TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation

Jinyu Yang[1], Jingjing Liu[2], Ning Xu[2], and Junzhou Huang[1]
[1]University Of Texas at Arlington, [2]Kuaishou Technology
jinyu.yang@mavs.uta.edu, jjliu08cas@gmail.com, ningxu01@gmail.com, jzhuang@uta.edu

Abstract

Unsupervised domain adaptation (UDA) aims to transfer the knowledge learnt from a labeled source domain to an unlabeled target domain. Previous work is mainly built upon convolutional neural networks (CNNs) to learn domain-invariant representations. With the recent exponential increase in applying Vision Transformer (ViT) to vision tasks, the capability of ViT in adapting cross-domain knowledge, however, remains unexplored in the literature. To fill this gap, this paper first comprehensively investigates the performance of ViT on a variety of domain adaptation tasks. Surprisingly, ViT demonstrates superior generalization ability, while the performance can be further improved by incorporating adversarial adaptation. Notwithstanding, directly using CNNs-based adaptation strategies fails to take the advantage of ViT's intrinsic merits (e.g., attention mechanism and sequential image representation) which play an important role in knowledge transfer. To remedy this, we propose an unified framework, namely Transferable Vision Transformer (TVT), to fully exploit the transferability of ViT for domain adaptation. Specifically, we delicately devise a novel and effective unit, which we term Transferability Adaption Module (TAM). By injecting learned transferabilities into attention blocks, TAM compels ViT focus on both transferable and discriminative features. Besides, we leverage discriminative clustering to enhance feature diversity and separation which are undermined during adversarial domain alignment. To verify its versatility, we perform extensive studies of TVT on four benchmarks and the experimental results demonstrate that TVT attains significant improvements compared to existing state-of-the-art UDA methods.

1. Introduction

Deep neural networks (DNNs) demonstrate unprecedented achievements on various machine learning problems and applications. However, such impressive performance heavily relies on massive amounts of labeled data which requires considerable time and labor efforts to collect. Therefore, it is desirable to train models that can leverage rich labeled data from a different but related domain and generalize well on target domains with no or limited labeled examples. Unfortunately, the canonical supervised-learning paradigm suffers from the domain shift issue that poses a major challenge in adapting models across domains. This motivates the research on unsupervised domain adaptation (UDA) [52] which is a special scenario of transfer learning [36]. The key idea of UDA is to project data points of the labeled source domain and the unlabeled target domain into a common feature space, such that the projected features are both discriminative (semantic meaningful) and domain-invariant, in turn, generalize well to bridge the domain gap. To achieve this goal, various methods have been proposed in the past decades, among which adversarial adaptation has become the dominant technique in this field, which attempts to align cross-domain representations by minimizing an adversarial loss through a domain discriminator [13, 47, 30, 59].

Recently, Vision Transformer (ViT) [11] has received increasing attention in the vision community. Different from CNNs that act on local receptive fields of the given image, ViT models long-range dependencies among visual features across the entire image, through the global self-attention mechanism. Specifically in ViT, each image is split into a sequence of fixed-size non-overlapping patches, which are then linearly embedded and concatenated with position embeddings. To be consistent with NLP paradigm, a class token is prepended to the patch tokens, serving as the representation of the whole image. Then, those sequential embeddings are fed into a stack of transformers to learn desired visual representations. Due to its advantages in global context modeling, ViT has obtained excellent results on various vision tasks, such as image classification [11], object detection [5, 53], segmentation [64, 28], and video understanding [14, 34].

Despite that ViT is becoming increasingly popular, two important questions related to domain adaption remain unanswered. First, *how does the generalization ability of*

[1]This work was done while Jinyu Yang was interning at Kuaishou Technology; code: https://github.com/uta-smile/TVT

ViT across different domains? There are several contemporary work [58, 55, 32] that apply DeiT [46] and Swin [28] to UDA, yet the ViT has not been investigated. The second question is, *how can we properly improve ViT in adapting different domains?* One intuitive approach is to directly apply adversarial discriminator onto the class tokens to perform adversarial alignment, where the state of a class token represents the entire image. However, cross-domain alignment of such global features assumes all regions or aspects of the image have the equal transferability and discriminative potential, which is not always tenable. For instance, background regions can be easier aligned across domains, while foreground regions are more discriminative. In other words, some discriminative features may lack transferability, and some transferable features may not contribute much to the downstream task (e.g., classification). Therefore, in order to properly enhance the transferability of ViT, it is essential to identify fine-grained features that are both transferable and discriminative.

In this paper we aim to present our answers to the two aforementioned questions. Firstly, to fill the blank of understanding ViT's generalization ability, we first conduct a comprehensive study of vanilla ViT [11] on public UDA benchmarks. As expected, our experimental results demonstrate that ViT even in the source-only setting outperforms its strong CNNs-based counterparts. There could be multiple deep reasons behind the strong performance of ViT [40, 66], which are not in the scope of this paper. Besides, we observe further improvements by applying an adversarial discriminator to the class tokens of ViT, which only aligns global representations. However, such strategy suffers from the oversimplified assumption and ignores the inherent properties of ViT that are beneficial for domain adaptation: i) sequential patch tokens actually give us the free access to fine-grained features; ii) the self-attention mechanism in transformer naturally works as a discriminative probe. In the light of this, we propose an unified UDA framework that makes full use of ViT's inherent merits. We name it Transferable Vision Transformer (TVT).

The key idea of our method is to retain both transferable and discriminative features which are essential in knowledge adaptation. To achieve this goal, we first introduce the novel Transferability Adaption Module (TAM) built upon a conventional transformer. TAM uses a patch-level domain discriminator to measure the transferabilities of patch tokens, and injects learned transferabilities into the multi-head self-attention block of a transformer. On one hand, the attention weights of patch tokens in the self-attention block are used to determine their semantic importance, i.e., the features with larger attention are more discriminative yet without transferability guarantees. On the other hand, as patch tokens can be regarded as fine-grained representations of an image, the higher transferability of a token means the local features are more transferable across domains though not necessarily discriminative. By simply replacing the last transformer of ViT with a plug-and-play TAM, we could drive ViT to focus on both transferable and discriminative features.

Since our method performs adversarial adaptation that forces the learned features of two domains to be similar, one underlying side-effect is that the discriminative information of target domain might be destroyed during feature alignment. To address this problem, we design a Discriminative Clustering Module (DCM) inspired by the clustering assumption. The motivation is to enforce the individual target prediction close to one-hot encoding (well separated) and the global target prediction to be uniformly distributed (global diverse), such that the learnt target-domain representation could retain maximum discriminative information about the input values.

Contributions of this paper are summarized as follows:

- As far as we know, this is the first comprehensive investigation of ViT's capability in transferring knowledge on the domain adaptation task. We believe this work gives good insights to understand and explore ViT's generalization ability while applied to various vision tasks.
- We propose TAM that delicately leverages the intrinsic characteristics of ViT, such that our method can capture both transferable and discriminative features for domain adaptation. Moreover, we adopt discriminative clustering assumption to alleviate the discrimination destruction during adversarial alignment.
- Without any bells and whistles, our method set up a new competitive baseline cross several public UDA benchmarks.

2. Related Work

Unsupervised Domain Adaptation Transfer learning aims to learn transferable knowledge that are generalizable across different domains with different distributions [36, 62]. This is built upon the evidence that feature representations in machine learning models, especially in deep neural networks, are transferable [63]. The main challenge of transfer learning is to reduce the domain shift or the discrepancy of the marginal probability distributions across domains [52]. In the past decades, various methods have been proposed to address one canonical transfer learning problem, i.e., unsupervised domain adaptation (UDA), where no labels are available for the target domain. For instance, DDC [48] attempted to learn domain-invariant features by minimizing Maximum Mean Discrepancy (MMD) [3] between two domains. Long et al. further improved DDC by embedding hidden representations of all

task-specific layers in a reproducing Hilbert space and used a multiple kernel variant of MMD to measure the domain distance [29]. Long et al. proposed to align joint distributions of multiple domain-specific layers across domains through a joint maximum mean discrepancy metric [31]. Another line of effort was inspired by the success of adversarial learning [16, 61]. By introducing a domain discriminator and modeling the domain adaption as a minimax problem [13, 47, 30, 60], an encoder is trained to generate domain-invariant features, through deceiving a discriminator which tries to distinguish features of source domain from that of target domain.

It is noteworthy that all of these methods completely or partially used CNNs as the fundamental block [22, 21, 17]. By contrast, our method explores ViT [11] to tackle the UDA problem, as we believe ViT has better potential and capability in domain adaptation owning to some of its properties. Although previous UDA methods (e.g., adversarial learning) are able to improve vanilla ViT to some extent, they were not well designed for transformer-based models, and thereby cannot leverage ViT's inherent characteristic of providing attention information and fine-grained representations. However, Our method is delicately designed with the nature of ViT and could effectively leverages the transferability and discrimination of each feature for knowledge transfer, thus having better chance in fully exploiting the adaptation power of ViT.

Vision Transformer Transformers [49] was firstly proposed in the NLP field and demonstrate record-breaking performance on various language tasks, e.g., text classification and machine translation [10, 2, 65]. Much of such impressive achievement is attributed to the power of capturing long-range dependencies through attention mechanism. Spurred by this, some recent studies attempted to integrate attention into CNNs to augment feature maps, aiming to provide the capability in modeling heterogeneous interactions [54, 1, 19]. Another pioneering work of completely convolution-free architecture is Vision Transformer (ViT), which applied transformers on a sequence of fixed-size non-overlapping image patches. Different from CNNs that rely on image-specific inductive biases (e.g., locality and translation equivariance), ViT takes the benefits from large-scale pre-training data and global context modeling. One such method [11], known for its simplicity and accuracy/compute trade-off, competes favorably against CNNs on the classification task and lays the foundation for applying transformer to different vision tasks. ViT and its variants have proved their wide applicability in object detection [5, 67, 53], segmentation [64, 57], and video understanding [14, 34], etc.

Despite the success of ViT on different vision tasks, to the best of our knowledge, neither their transferability nor the design of UDA methods with ViT have been previously discussed in the literature. To this end, we focus in this paper on the investigation of ViT's capability in knowledge transferring across different domains. Furthermore, we propose a novel UDA framework tailored for ViT by exploring its intrinsic merits and prove its superiority over existing methods. It is noteworthy that there are several contemporary work [58, 55, 32] that apply DeiT [46] and Swin [28] to UDA. Specifically, [58, 55] uses cross-attention to obtain the mixup representations of source and target images, [32] uses two class tokens to learn domain-specific information. Different from these works, our paper focuses on the empirical investigation of ViT's generalization ability and proposes a plug-and-play module to boost ViT's performance in knowledge transfer.

3. Preliminaries

3.1. Adversarial Learning UDA

We consider the image classification task in UDA, where a labeled source domain $\mathcal{D}_s\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s examples and an unlabeled target domain $\mathcal{D}_t\{x_j^t\}_{j=1}^{n_t}$ with n_t examples are given. The goal of UDA is to learn features that are both discriminative and invariant to the domain discrepancy, and in turn guarantee accurate prediction on the unlabeled target data. Here, a common practice is to jointly performs feature learning, domain adaptation, and classifier learning by optimizing the following loss function:

$$\mathcal{L}_{clc}(x^s, y^s) + \alpha\mathcal{L}_{dis}(x^s, x^t) \tag{1}$$

where $\mathcal{L}_{clc}$ is supervised classification loss, $\mathcal{L}_{dis}$ is a transfer loss with various possible implementations, and α is used to control the importance of $\mathcal{L}_{dis}$. One of the most commonly used $\mathcal{L}_{dis}$ is the adversarial loss which encourages a domain-invariant feature space through a domain discriminator [13].

3.2. Self-attention Mechanism

The main building block of ViT is Multi-head Self-Attention (MSA), which is used in the transformer to capture long-range dependencies [49]. Specifically, MSA concatenates multiple scaled dot-product attention (short for SA) modules, where each SA module takes a set of queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) as inputs. In order to learn dependencies between distinct positions, SA computes the dot products of the query with all keys, and applies a softmax function to obtain the weights on the values.

$$\text{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \tag{2}$$

where d is the dimension of $\mathbf{Q}$ and $\mathbf{K}$. With $\text{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, MSA is defined as:

$$\begin{aligned} \text{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Concat}(\text{head}_1, ..., \text{head}_k)\mathbf{W}^O \\ \text{where head}_i &= \text{SA}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \end{aligned} \tag{3}$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ are projections of different heads, $\mathbf{W}^O$ is another mapping function. Intuitively, using multiple heads allows MSA to jointly attend to information from different representation subspaces at different positions.

4. Methodology

In this section, we first investigate ViT's ability in knowledge transfer on various adaptation tasks. After that, we conduct the early attempts to improve ViT's transferability by incorporating adversarial learning. Finally, we introduce our method named Transferable Vision Transformer (TVT), which consists two new adaptation modules to further improve ViT's capability for cross-domain adaptation..

4.1. ViT's Generalization Ability

To the best of our knowledge, the generalization ability of ViT has not been studied in the literature before, although ViT and its variants have shown great success in various vision task. To probe into ViT's capability of domain adaptation, we choose the vanilla ViT [11] as the backbone in all of our studies, owing to its simplicity and popularity. We train vanilla ViT by labeled source data only and assess its generalization ability by the classification accuracy on target data. As mentioned above, CNNs-based approaches dominate UDA research in the past decades and demonstrate great successes. Therefore, we compare vanilla ViT with CNNs-based architectures, including LeNet [22], AlexNet [21], and ResNet [17]. All experiments are performed on well-established benchmarks with standard evaluation protocols.

Take the results on Office-31 dataset for example. As shown in Table 2, Source Only ViT obtains impressing classification accuracy 89.5%, which is much better than its strong CNN opponents AlexNet (70.1%) and ResNet (76.1%). Similar phenomenon can be observed in other benchmark results, where ViT competes favorably against, if not better than, the other state-of-the-arts CNNs backbones, as shown in Table 1,3,4. Surprisingly, Source Only ViT even outperforms strong CNNs-based UDA approaches without any bells and whistles. For instance, it achieves an average accuracy 78.7% on Office-Home dataset (Table 3), beating all CNN-based UDA methods. Compared to SHOT [26] recognized as the best UDA model nowadays, Source Only ViT obtains 7% absolute accuracy boost, a big step in pushing the frontier of UDA research. There could be multiple reasons behind the strong performance of ViT [40, 66], for example, the striking differences between the features learned by ViTs and CNNs [40]. We leave this as future work. Despite this, a large gap still exists between the Source Only and Target Only models (88.3% vs 99.2%) as shown in Table 1, which indicates potential improvement space of ViT's generalization ability.

4.2. ViT w/ Adversarial Adaptation: Baseline

We first investigate how ViT benefits from adversarial adaptation [13], which is widely used in CNNs-based UDA methods. We follow the typical adversarial adaptation fashion that employs an encoder G_f for feature learning, a classifier G_c for classification, and a domain discriminator D_g for global feature alignment. Here, G_f is implemented as ViT and D_g is applied to output state of the class tokens of the source and target images. To accomplish domain knowledge adaptation, G_f and D_g play a minimax game: G_f learns domain-invariant features to deceive D_g, while D_g distinguishes source-domain features from that of target-domain. The objective can be formulated as:

$$\begin{aligned}\mathcal{L}_{clc}(x^s, y^s) &= \frac{1}{n_s}\sum_{x_i \in \mathcal{D}_s} \mathcal{L}_{ce}(G_c(G_f(x_i^s)), y_i^s) \\ \mathcal{L}_{dis}(x^s, x^t) &= -\frac{1}{n}\sum_{x_i \in \mathcal{D}} \mathcal{L}_{ce}(D_g(G_f(x_i^*)), y_i^d),\end{aligned} \tag{4}$$

where $n = n_s + n_t$, $\mathcal{D} = \mathcal{D}_s \bigcup \mathcal{D}_t$, $\mathcal{L}_{ce}$ is cross-entropy loss, the superscript $*$ can be either s or t to denote a source or a target domain, and y^d denotes the domain label (i.e., $y^d = 1$ is source, $y^d = 0$ is target).

We denote ViT with adversarial adaptation as our Baseline. As shown in Table 1,2,3,4, Baseline shows 7.8%, 0.8%, 1.6%, and 3.2% absolute accuracy improvements over vanilla ViT, respectively on the four benchmarks. Those results reveal that global feature alignment with a domain discriminator helps ViT's generalization ability. However, compared with the digit recognition task, Baseline achieves limited improvements on object detection which is more complicated and challenging. We boils down such observation to a conclusion that simply applying global adversarial alignment cannot exploit ViT's full transferable power, since it fails to consider two key factors: (i) not all regions/features are equally transferable or discriminative. For effective knowledge transfer, it is essential to focus on both transferable and discriminative features; (ii) ViT naturally provides fine-grained features given its forward passing sequential tokens, and attention weights in transformer actually convey discriminative potentials of patch tokens. To address these challenges and fully leverage the merits of ViT, a new UDA framework named Transferable Vision Transformer (TVT) is further proposed.

4.3. Transferable Vision Transformer (TVT)

An overview of TVT is shown in Figure 1, which contains two main modules: (i) a Transferability Adaptation Module (TAM) and (ii) a Discriminative Clustering Module (DCM). These two modules are highly interrelated and play a complementary role in transferring knowledge for ViT-based architectures. TAM encourages the output state

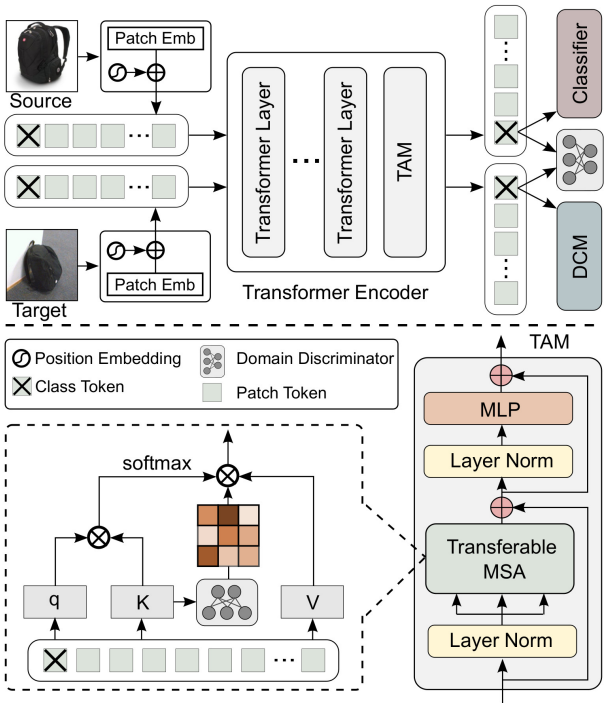

Figure 1. An overview of the proposed TVT framework. As in ViT, both source and target images are split into fixed-size patches which are then linearly mapped and embedded with positional information. The generated patches are fed into a transformer encoder whose last layer is replaced by Transferability Adaptation Module (TAM). Feature learning, adversarial domain adaptation and classification are accomplished by ViT-akin backbone, two domain discriminators (on patch-level and global-level), Discriminative Clustering Module (DCM) and the MLP-based classifier

of class token to focus on both transferable and semantic meaningful features, and DCM enforces the aligned features of target-domain samples to be clustered with large margins. As a consequence, the features learnt by TVT are discriminative in classification and transferable across domains as well. We detail each module in what follows.

4.3.1 Transferability Adaptation Module

As shown in Figure 1, we introduce the Transferability Adaptation Module (TAM) that explicitly considers the intrinsic merits of ViT, i.e., attention mechanisms and sequential patch tokens.

As the patch tokens are regarded as local features of an image, they are corresponded to different image regions or captures different visual aspects as fine-grained representations of an image. Assuming patch tokens of different semantic importance and transferabilities, TAM aims at assigning different weights to those tokens, to encourage the learned image representations, i.e., the output state of class token, to attend to patch tokens that are both transferable and discriminative. While the self-attention weights in ViT could be employed as discriminative weights, one major hurdle here is, the transferability of each patch token is not available. To bypass this difficulty, we adopt a patch-level domain discriminator D_l that matches cross-domain local features [37, 56] by optimizing:

$$\mathcal{L}_{pat}(x^s, x^t) = -\frac{1}{nR}\sum_{x_i\in\mathcal{D}}\sum_{r=1}^{R}\mathcal{L}_{ce}(D_l(G_f(x^*_{ir})), y^d_{ir}), \tag{5}$$

where R is number of patches, and $D_l(f_{ir})$ is the probability of this region belonging to the source domain. During adversarial learning, D_l tries to assign 1 for a source-domain patch and 0 for the target-domain ones, while G_f combats such circumstances. Conceptually, a patch that can easily deceive D_l (i.e., D_l is around 0.5) is more transferable across domains and should be given a higher transferability. We therefore use $t_{ir} = T(f_{ir}) = H(D_l(f_{ir})) \in [0, 1]$ to measure the transferability of r^{th} token of i^{th} image, where $H(\cdot)$ is the standard entropy function. An other explanation of the transferability is: by assigning weights to different patches, it disentangles an image into common space representations and domain-specific representations, while the passing paths of domain-specific features are softly suppressed.

We then convert the conventional MSA into the transferable MSA (T-MSA) by transferability adaptation, i.e., injecting the learned transferabilities into attention weights of the class token. Our T-MSA is built upon the transferable self-attention (TSA) block that is formally defined as:

$$\text{TSA}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d}}) \odot [1; T(\mathbf{K}_{patch})]\mathbf{V} \tag{6}$$

where $\mathbf{q}$ is the query of the class token, $\mathbf{K}_{patch}$ is the key of the patch tokens, $\odot$ is Hadamard product, and $[;]$ is concatenation operation. Obviously, softmax($\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d}}$) and $[1; T(\mathbf{K}_{patch})]$ indicate the discrimination (semantic importance) and the transferability of each patch token, respectively. To jointly attend to the transferabilities of different representation subspaces and of different locations, we thus define T-MSA as:

$$\begin{aligned}\text{T-MSA}(\mathbf{q}, \mathbf{K}, \mathbf{V}) &= \text{Concat}(\text{head}_1, ..., \text{head}_k)\mathbf{W}^O \\ \text{where head}_i &= \text{TSA}(\mathbf{q}\mathbf{W}^q_i, \mathbf{K}\mathbf{W}^K_i, \mathbf{V}\mathbf{W}^V_i)\end{aligned} \tag{7}$$

Taken them together, we get the TAM as follows:

$$\begin{aligned}\hat{\mathbf{z}}^l &= \text{T-MSA}(\text{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1} \\ \mathbf{z}^l &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^l)) + \hat{\mathbf{z}}^l,\end{aligned} \tag{8}$$

where LN is LayerNorm layer, MLP denotes Multi-Layer Perception, $\mathbf{z}^l$ is hidden representation at layer l. We only apply TAM to the last transformer layer where patch features are spatially non-local and of higher semantic meanings. By this means, TAM focuses on fine-grained features

that are transferable across domains and are discriminative for classification. So we have $l = L$, where L is the total number of transformer layers in ViT.

4.3.2 Discriminative Clustering Module

Towards the challenging problem of learning a probabilistic discriminative classifier with unlabeled target data, it is desirable to minimize the expected classification error on the target domain. However, cross-domain feature alignment through TAM by forcing the two domains to be similar may destroy the discriminative information of the learned representation, if no semantic constrains of the target domain is introduced. As shown in Figure 2, although the target feature is indistinguishable from the source feature, it is distributed in a mess which limits its discriminative power. To address this limitation, we are inspired by the assumptions that: (i) $p^t = \text{softmax}(G_c(G_f(x^t)))$ are expected to retain as much information about x^t as possible [4, 33, 45, 42, 51]; and (ii) decision boundary should not cross high density regions, but instead lie in low density regions, which is also known as cluster assumption [6]. Fortunately, these two assumptions can be met by maximizing mutual information between the empirical distribution on the target inputs and the induced target label distribution [15, 44, 20, 25, 39], which can be formally defined as:

$$\begin{aligned}\mathcal{I}(p^t; x^t) &= H(\bar{p^t}) - \frac{1}{n_t}\sum_{j=1}^{n_t} H(p_j^t) \\ &= -\sum_{k=1}^{K} \bar{p_k^t} log(\bar{p_k^t}) + \frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{K} p_{jk}^t log(p_{jk}^t)\end{aligned} \tag{9}$$

where $p_j^t = \text{softmax}(G_c(G_f(x_j^t)))$, $\bar{p^t} = \mathbb{E}_{x_t}[p^t]$, and K is the number of classes. Note that maximizing $-\frac{1}{n_t}\sum_{j=1}^{n_t} H(p_j^t)$ enforces the target predictions close to one-hot encoding, therefore the cluster assumption is guaranteed. To ensure the global diversity, we also maximize $H(\bar{p^t})$ to avoid that every target data is assigned to the same class. With $\mathcal{I}(p^t; x^t)$, our model is encouraged to learn tightly clustered target features with uniform distribution, such that the discriminative information in the target domain are retained.

To summarize, the objective function of TVT is:

$$\mathcal{L}_{clc}(x^s, y^s) + \alpha\mathcal{L}_{dis}(x^s, x^t) + \beta\mathcal{L}_{pat}(x^s, x^t) - \gamma\mathcal{I}(p^t; x^t) \tag{10}$$

where α, β, and γ are hyper-parameters.

5. Experiments

To verify the effectiveness of our model, we conduct comprehensive studies on commonly used benchmarks and present experimental comparisons against state-of-the-art UDA methods as shown below.

Algorithm		S→M	U→M	M→U	Avg
Source Only	LeNet	67.1	69.6	82.2	73.0
RevGrad [12]	LeNet	73.9	73.0	77.1	74.7
ADDA [47]	LeNet	76.0	90.1	89.4	85.2
SHOT-IM [26]	LeNet	89.6	96.8	91.9	92.8
CyCADA [18]	LeNet	90.4	96.5	95.6	94.2
MCD [43]	LeNet	96.2	94.1	94.2	94.8
Target Only		99.4	99.4	98.0	98.9
Source Only	ViT	88.6	88.2	73.1	88.3
Baseline	ViT	92.7	98.6	97.0	96.1
TVT*	ViT	98.0	98.9	97.7	98.2
TVT	ViT	99.0	99.4	98.2	98.9
Target Only		99.7	99.7	98.3	99.2

Table 1. Performance comparison on the Digits dataset. TVT* indicates that the backbone is pre-trained on ImageNet

Digits is an UDA benchmark on digit classification. We follow the same setting in previous work to perform adaptations on MNIST [22], USPS, and Street View House Numbers (SVHN) [35]. For each source-target domain pair, we train our model using the training sets of each domain, and perform evaluations on the standard test set of the target domain.

Office-31 [41] contains 4,652 images of 31 categories, which were collected from three domains: Amazon (A), DSLR (D), and Webcam (W). The Amazon (A) image were downloaded from amazon.zom, while the DSLR (D), and Webcam (W) were photoed under the office environment by web and digital SLR camera, respectively.

Office-Home [50] consists of images from four different domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-World images (Rw). A total of 65 categories are covered within each domain.

VisDA-2017 [38] is a synthesis-to-real object recognition task used for the 2018 VisDA challenge. It covers 12 categories. The source domain contains 152,397 synthetic 2D renderings generated from different angles and under different lighting conditions, while the target domain contains 55,388 real-world images.

5.1. Existing Methods

We use the results in their original papers for fair comparison. For each type of backbone, we report its lower bound performance, denoted as Source Only, meaning the models are trained with source data only. For digit recognition, we also show the Target Only results as the high-end performance, which is obtained by both training and testing on the labeled target data. Baseline denotes vanilla ViT with adversarial adaptation [13].

5.2. Implementation Details

The ViT-Base with 16×16 input patch size (or ViT-B/16) [11] pre-trained on ImageNet-21K [9] is used as our back-

Algorithm		A→ W	D→ W	W→ D	A→ D	D→ A	W→ A	Avg
Source Only	AlexNet	61.6	95.4	99.0	63.8	51.1	49.8	70.1
DDC [48]	AlexNet	61.8	95.0	98.5	64.4	52.1	52.2	70.6
DAN [29]	AlexNet	68.5	96.0	99.0	67.0	54.0	53.1	72.9
RevGrad [12]	AlexNet	73.0	96.4	99.2	72.3	53.4	51.2	74.3
PFAN [7]	AlexNet	83.0	99.0	99.9	76.3	63.3	60.8	80.4
Source Only	ResNet	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DDC [48]	ResNet	75.6	96.0	98.2	76.5	62.2	61.5	78.3
DAN [29]	ResNet	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RevGrad [12]	ResNet	82.0	96.9	99.1	79.7	68.2	67.4	82.2
TAT [27]	ResNet	92.5	99.3	100.0	93.2	73.1	72.1	88.4
SHOT [26]	ResNet	90.1	98.4	99.9	94.0	74.7	74.3	88.6
ALDA [8]	ResNet	95.6	97.7	100.0	94.0	72.2	72.5	88.7
Source Only-S	DeiT	86.9	97.7	99.6	87.6	74.9	73.5	86.7
CDTrans-S [58]	DeiT	93.5	98.2	99.6	94.6	78.4	78.0	90.4
Source Only-B	DeiT	90.4	98.2	100.0	90.8	76.8	76.4	88.8
CDTrans-B [58]	DeiT	96.7	99.0	100.0	97.0	81.1	81.9	92.6
Source Only	Swin	89.2	94.1	100.0	93.1	80.9	81.3	89.8
BCAT [55]	Swin	99.2	99.5	100.0	99.6	85.7	86.1	95.0
Source Only	ViT	89.2	98.9	100.0	88.8	80.1	79.8	89.5
Baseline	ViT	91.6	99.0	100.0	90.6	80.2	80.1	90.2
TVT*	ViT	95.7	98.7	100.0	95.4	80.6	80.3	91.8
TVT	ViT	96.4	99.4	100.0	96.4	84.9	86.1	93.9

Table 2. Performance comparison on the Office-31 dataset. TVT* indicates that the backbone is pre-trained on ImageNet. "-S" and "-B" indicate that the backbone is DeiT-Small and DeiT-Base, respectively

bone. The transformer encoder of ViT-B/16 contains 12 transformer layers in total. We train all ViT-based models using mini-batch Stochastic Gradient Descent (SGD) optimizer with the momentum of 0.9. We initialized the learning rate as 0 and linearly increase it to $lr = 0.03$ after 500 training steps. We then decrease it by the cosine decay strategy. The only exception is that we set $lr = 0.003$ for D→ A and W→ A in Office-31 dataset.

5.3. Results of Digit Recognition

For the digit recognition task, we perform evaluations on SVHN→MNISt, USPS→MNIST, and MNIST→USPS, following the standard evaluation protocol of UDA. Shown in Table 1, TVT obtains the best mean accuracy for each task and outperforms prior work in terms of the average classification accuracy. TVT also performs better than Baseline (+2.7%) due to the contribution of the proposed TAM and DCM. In particular, TVT achieves comparable results to Target Only model, indicating that the domain shift problem is well alleviated.

5.4. Results of Object Recognition

For object recognition task, Office-31, Office-Home, and VisDA-2017 are used in evaluation. As shown in Table 2 3, 4, TVT sets up new benchmark results for all the three datasets. On the medium-sized Office-Home dataset (Table 3), we achieve the significant improvement over the best prior UDA method (83.6% vs 71.8%).

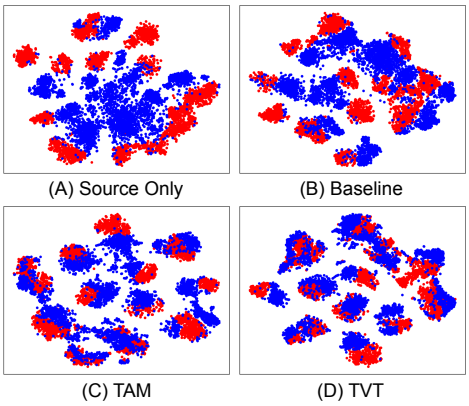

Figure 2. t-SNE visualization of VisDA-2017 dataset, where red and blue points indicate the source (synthetic rendering) and the target (real images) domain, respectively

Results on the large-scale VisDA-2017 dataset (Table 4) show that we not only achieve a higher average accuracy, but also compete favorably against ALDA and SHOT. Specifically, we use the most naive pseudo-labeling strategy (pseudo labels with high confidence) [23] in this experiment. Note that DTA also enforces the cluster assumption to learn discriminative features, but it fails to encourage the global diversity which may leads to a degenerate solution where every point is assigned to the same class. Besides, TVT surpasses both Source Only and Baseline, revealing its effectiveness in transferring domain knowledge by (i) capturing both transferable and discriminative fine-grained features and (ii) retaining discriminative information while searching for the domain-invariant representations. This is also evidenced by the t-SNE visualization of learned features as showcased in Figure 2. Obviously, TAM can effectively align source and target domain features by exploiting the local feature transferability. However, the target feature is not well-separated due to that target labels in training are absent and the discriminative information are destroyed by adversarial alignment. Fortunately, this problem is alleviated by DCM by assuming that datapoints should be classified with large margin, as illustrated in Figure 2 (D). It is noteworthy that several contemporary work [58, 55, 32] use DeiT [46] or Swin [28] as the backbone and outperforms our method. We argue that this can be mainly explained by the data-efficient merits of DeiT and Swin. Detailed discussion are referred to the supplementary.

5.5. Ablation Study

To learn the individual contribution of TAM and DCM in improving the knowledge transferability of ViT, we conduct the ablation study in Table 5. Compared to Source Only, TAM consistently improves the classification accuracy with average 4.9% boost, indicating the significance of capturing both transferable and discriminative features. The

Algorithm		A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Source Only	AlexNet	26.4	32.6	41.3	22.1	41.7	42.1	20.5	20.3	51.1	31.0	27.9	54.9	34.3
DAN [29]	AlexNet	31.7	43.2	55.1	33.8	48.6	50.8	30.1	35.1	57.7	44.6	39.3	63.7	44.5
RevGrad [12]	AlexNet	36.4	45.2	54.7	35.2	51.8	55.1	31.6	39.7	59.3	45.7	46.4	65.9	47.3
Source Only	ResNet	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [29]	ResNet	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
RevGrad [12]	ResNet	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
SHOT [26]	ResNet	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
Source Only-S	DeiT	55.6	73.0	79.4	70.6	72.9	76.3	67.5	51.0	81.0	74.5	53.2	82.7	69.8
CDTrans-S [58]	DeiT	60.6	79.5	82.4	75.6	81.0	82.3	72.5	56.7	84.4	77.0	59.1	85.5	74.7
WinTR-S [32]	DeiT	65.3	84.1	85.0	76.8	84.5	84.4	73.4	60.0	85.7	77.2	63.1	86.8	77.2
Source Only-B	DeiT	61.8	79.5	84.3	75.4	78.8	81.2	72.8	55.7	84.4	78.3	59.3	86.0	74.8
CDTrans-B [58]	DeiT	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
Source Only	Swin	64.5	84.8	87.6	82.2	84.6	86.7	78.8	60.3	88.9	82.8	65.3	89.6	79.7
BCAT [55]	Swin	75.3	90.0	92.9	88.6	90.3	92.7	87.4	73.7	92.5	86.7	75.4	93.5	86.6
Source Only	ViT	66.2	84.3	86.6	77.9	83.3	84.3	76.0	62.7	88.7	80.1	66.2	88.7	78.7
Baseline	ViT	71.9	80.7	86.7	79.9	80.4	83.5	76.9	70.9	88.3	83.0	72.9	88.4	80.3
TVT*	ViT	67.1	83.5	87.3	77.4	85.0	85.6	75.6	64.9	86.6	79.1	67.2	88.0	78.9
TVT	ViT	74.9	86.8	89.5	82.8	88.0	88.3	79.8	71.9	90.1	85.5	74.6	90.6	83.6

Table 3. Performance comparison on the Office-Home dataset. TVT* indicates that the backbone is pre-trained on ImageNet. "-S" and "-B" indicate that the backbone is DeiT-Small and DeiT-Base, respectively

Algorithm		plane	bcycl	bus	car	house	knife	mcycl	person	plant	sktbrd	train	truck	Avg
Source Only	ResNet	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
RevGrad [12]	ResNet	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD [43]	ResNet	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
ALDA [8]	ResNet	93.8	74.1	82.4	69.4	90.6	87.2	89.0	67.6	93.4	76.1	87.7	22.2	77.8
DTA [24]	ResNet	93.7	82.2	85.6	83.8	93.0	81.0	90.7	82.1	95.1	78.1	86.4	32.1	81.5
SHOT [26]	ResNet	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
Source Only-B	DeiT	97.7	48.1	86.6	61.6	78.1	63.4	94.7	10.3	87.7	47.7	94.4	35.5	67.1
CDTrans-B [58]	DeiT	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
WinTR-B [32]	DeiT	98.7	91.2	93.0	91.9	98.1	96.1	94.0	72.7	97.0	95.5	95.3	57.9	90.1
Source Only	Swin	98.7	63.0	86.7	68.5	94.6	59.4	98.0	22.0	81.9	91.4	96.7	25.7	73.9
BCAT [55]	Swin	99.1	91.6	86.6	72.3	98.7	97.9	96.5	82.3	94.2	96.0	93.9	61.3	89.2
Source Only	ViT	98.2	73.0	82.5	62.0	97.3	63.5	96.5	29.8	68.7	86.7	96.7	23.7	73.2
Baseline	ViT	94.6	81.6	81.8	69.9	93.5	69.9	88.6	50.5	86.8	88.5	91.5	20.1	76.4
TVT*	ViT	97.1	88.8	86.4	64.4	96.4	97.4	90.6	64.1	92.0	90.3	93.7	59.6	85.1
TVT	ViT	97.1	92.9	85.3	66.4	97.1	97.1	89.3	75.5	95.0	94.7	94.5	55.1	86.7

Table 4. Performance comparison on the VisDA-2017 dataset. TVT* indicates that the backbone is pre-trained on ImageNet. "-B" indicates that the backbone is DeiT-base

Methods	Digits	Office-31	Office-Home	VisDA-2017	Avg
Source Only	88.3	89.5	78.7	73.2	82.4
+TAM	97.2	91.2	81.3	79.3	87.3
+DCM	98.9	93.9	83.6	86.7	90.8

Table 5. Ablation study of each module

performance is further improved by incorporating DCM, justifying the necessary of retaining the discriminative information of the learned representation. It is noteworthy that DCM brings the largest improvement on the large-scale synthetic-to-real VisDA-2017 dataset. We suspect that the large domain gap in VisDA-2017 (synthetic 2D rendering to natural image) is the leading reason, since simply aligning two domains with large domain shift results in a mess distributed feature space. This challenge, however, can be largely addressed by DCM that enables retaining discriminative information based on a cluster assumption.

6. Conclusion

In this paper, we perform a comprehensive investigation of ViT's generalization ability in UDA task. To further improve the power of ViT in transferring domain knowledge, we propose TVT by explicitly considering the intrinsic merits of transformer architecture. Specifically, TVT captures both transferable and discriminative features in the given image, and retains discriminative information of the learnt domain-invariant representations. Experimental results on widely used benchmarks show that TVT outperforms prior UDA methods by a large margin.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[4] John S Bridle, Anthony JR Heading, and David JC MacKay. Unsupervised classifiers, mutual information and'phantom targets'. 1992.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *International workshop on artificial intelligence and statistics*, pages 57–64. PMLR, 2005.

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019.

[8] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. Adversarial-learned loss for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020.

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

[13] 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.

[14] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[15] Ryan Gomes, Andreas Krause, and Pietro Perona. Discriminative clustering by regularized information maximization. 2010.

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[17] 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.

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[20] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International conference on machine learning*, pages 1558–1567. PMLR, 2017.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[24] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 91–100, 2019.

[25] Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020.

[26] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

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

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[29] 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.

[30] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.

[31] 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.

[32] Wenxuan Ma, Jinming Zhang, Shuang Li, Chi Harold Liu, Yulin Wang, and Wei Li. Exploiting both domain-specific and invariant knowledge via a win-win transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2111.12941*, 2021.

[33] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE international conference on computer vision*, pages 5067–5075, 2017.

[34] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

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

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

[37] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[38] 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.

[39] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8558–8567, 2021.

[40] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

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

[42] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

[43] 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.

[44] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. *arXiv preprint arXiv:1206.6438*, 2012.

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

[46] 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.

[47] 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.

[48] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: maximizing for domain invariance (2014). *Preprint. arXiv*, 1412, 2014.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[50] 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.

[51] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[52] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[53] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[55] Xiyu Wang, Pengxin Guo, and Yu Zhang. Domain adaptation via bidirectional cross-attention transformer. *arXiv preprint arXiv:2201.05887*, 2022.

[56] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[57] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *arXiv preprint arXiv:2011.14503*, 2020.

[58] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

[59] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European Conference on Computer Vision*, pages 480–498. Springer, 2020.

[60] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 514–524, 2021.

[61] Jinyu Yang, Chunyuan Li, Weizhi An, Hehuan Ma, Yuzhi Guo, Yu Rong, Peilin Zhao, and Junzhou Huang. Exploring robustness of unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9194–9203, 2021.

[62] Wei Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. Transfer learning via learning to transfer. In *International Conference on Machine Learning*, pages 5085–5094. PMLR, 2018.

[63] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

[64] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.

[65] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv preprint arXiv:2012.07436*, 2020.

[66] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, and Yizhou Yu. Convnets vs. transformers: Whose visual representations are more transferable? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2230–2238, 2021.

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.