TRUST: Leveraging Text Robustness for Unsupervised Domain Adaptation

Anonymous submission

Abstract

Recent unsupervised domain adaptation (UDA) methods have shown great success in addressing classical domain shifts (e.g., synthetic-to-real), but they still suffer under complex shifts (e.g. geographical shift), where both the background and object appearances differ significantly across domains. Prior works showed that the language modality can help in the adaptation process, exhibiting more robustness to such complex shifts. In this paper, we introduce TRUST, a novel UDA approach that exploits the robustness of the language modality to guide the adaptation of a vision model. TRUST generates pseudo-labels for target samples from their captions and introduces a novel uncertainty estimation strategy that uses normalised CLIP similarity scores to estimate the uncertainty of the generated pseudo-labels. Such estimated uncertainty is then used to reweight the classification loss, mitigating the adverse effects of wrong pseudo-labels obtained from low-quality captions. To further increase the robustness of the vision model, we propose a multimodal soft-contrastive learning loss that aligns the vision and language feature spaces, by leveraging captions to guide the contrastive training of the vision model on target images. In our contrastive loss, each pair of images acts as both a positive and a negative pair and their feature representations are attracted and repulsed with a strength proportional to the similarity of their captions. This solution avoids the need for hardly determining positive and negative pairs, which is critical in the UDA setting. Our approach outperforms previous methods, setting the new state-of-the-art on classical (DomainNet) and complex (GeoNet) domain shifts. The code will be available upon acceptance.

1 Introduction

Although deep learning models have achieved remarkable performance in lots of computer vision tasks, they still fall short in their ability to generalise to different domains. Retraining deep learning models on new data requires a big effort to acquire and manually label images and should be avoided. To address these limitations, Unsupervised Domain Adaptation (UDA) approaches for image classification have been proposed. UDA aims at transferring knowledge acquired on a labelled source domain to an unlabelled target domain, bridging the gap between them (Xu et al. 2019; Wei et al. 2024; Sharma, Kalluri, and Chandraker 2021; Chen et al. 2022b). Recent UDA methodologies have been successful on classical domain shifts (*e.g.* synthetic-to-real), but they suffer a lot on more complex shifts (*e.g.* geographical), where both the background and objects' appearance change between domains (Kalluri, Xu, and Chandraker 2023).

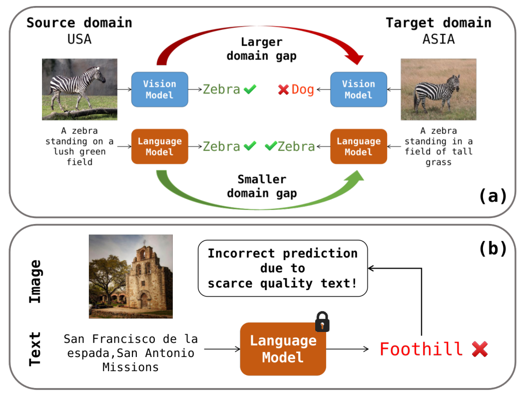

Figure 1: (a) geographical shifts strongly impact the appearance of both foreground objects and background. Contrarily, captions contain valuable information about the semantic class and they are minimally affected by the geographic shift, as stated in (Kalluri, Majumder, and Chandraker 2024). (b) Despite such apparent robustness, low quality captions may still lead to low classification accuracy.

Kalluri et al. (Kalluri, Majumder, and Chandraker 2024) posited that, under complex shift, solely relying on images is challenging and they proposed LaGTran that integrates textual data for guiding the adaptation. Indeed, authors showed that textual data are more robust to complex shifts, as they semantically describe the image content instead of focusing on appearance details, which is essential for bridging the domain gap (*c.f.* Figure 1(a)). However, LaGTran (Kalluri, Majumder, and Chandraker 2024) used textual data only for generating pseudo-labels for the target domain, highly underusing the potential of the language modality to reduce the domain shift. Moreover, they blindly rely on the pseudo-labels generated from the text, which may be incorrect due to two main factors: (a) the scarce quality of text descriptions, especially for crowd-sourced texts (*e.g.* Figure 1 (b)); (b) the impact of domain shifts on language models, which is lower than image-based models but still present. For those reasons, solely training the vision model with pseudo-labels generated from the text is suboptimal for transferring the shift robustness from the language to the vision model.

To overcome these limitations, we propose **TRUST** (**T**ext **R**ob**UST**ness for unsupervised domain adaptation), a novel approach for UDA in image classification that exploits the potential of language guidance for the adaptation of a vision model. Similarly to LaGTran (Kalluri, Majumder, and Chandraker 2024), we use a language model to generate pseudo-labels on target samples from captions, but we also propose two novel components to cope with the shortcomings discussed above.

First, to reduce the impact of incorrect pseudo-labels generated from low-quality captions, we propose a novel multimodal uncertainty estimation strategy that reweights the classification loss for target samples, by evaluating how well the captions semantically describe the target images. To this aim, we use a pretrained vision-language model (*i.e.* CLIP (Radford et al. 2021)) to evaluate the semantic correlation between target images and their captions, which serves as a measure of the uncertainty/reliability of generated pseudo-labels. The estimated uncertainty is then used to reweight the contribution of the pseudo-labels in the classification loss, reducing the negative impact of wrong pseudo-labels obtained from low-quality captions.

Second, to enhance the generalisation of the vision model, we aim to transfer the robustness to complex shifts from the language to the vision model, by promoting an effective alignment between their feature spaces. We propose to integrate a novel multimodal soft-contrastive learning loss, which uses the language modality to guide the contrastive training of the vision model. Unlike previous works, our soft-contrastive framework does not require to identify positive and negative pairs, as it assigns to each pair a score of "*positiveness*" and "*negativeness*" based on how likely they share the same semantic content. Then, a pair of images acts simultaneously as a positive and negative pair and attracts/repulses samples with a strength proportional to the similarity of their captions. The benefits of this strategy are multiple: (a) we encourage the vision model to match the language model's feature space by attracting representations of images with similar captions and repulsing those with dissimilar ones; (b) differently than Chen et al. (2022a); Litrico, Del Bue, and Morerio (2023), we avoid contrasting images of the same class without relying on the output of the vision model, therefore limiting the confirmation bias; (c) we achieve a smoother contrastive training, where representations of each pair of samples are both attracted and repulsed based on how likely they share the same semantic content.

We benchmark TRUST on three datasets representing classical (DomainNet, VisDA) and complex (GeoNet) shifts. On GeoNet, we obtain the best performance with a gain of $+2.96\%$ to the best competitor; on DomainNet and VisDA, we improve performance of $+2.53\%$ and $+2.10\%$, compared to state-of-the-art.

To summarise, our main contributions are:

- We introduce a novel uncertainty estimation strategy that leverages CLIP to evaluate the uncertainty of pseudo-labels, by measuring the semantic correlation between images and their captions.
- We propose a novel multimodal soft-contrastive learning loss that uses the language modality to guide the contrastive training of the vision model. Our solution removes the problem of identifying positive and negative samples in UDA, preventing the confirmation bias.
- We validate our method on classical and complex domain shifts outperforming the state-of-the-art on both settings.

2 Related Works

Unsupervised Domain Adaptation. Unsupervised Domain Adaptation aims to adapt a model trained on a labelled source domain to generalise well on an unlabelled target domain, in presence of domain shift (Ganin and Lempitsky 2015). Earlier UDA approaches focused on aligning statistical distributions (Peng et al. 2019). More recent methods proposed domain alignment strategies, including Maximum Mean Discrepancy (Tan, Peng, and Saenko 2020; Long et al. 2017), adversarial learning (Wei et al. 2024; Chen et al. 2022b), clustering (Kalluri and Chandraker 2022; Li et al. 2021) and self-training (Chen et al. 2022a; Litrico, Del Bue, and Morerio 2023). Other works leveraged generative models (Hoffman et al. 2018) and transformers (Zhu, Bai, and Wang 2023; Xu et al. 2021) to learn the underlying distributions of data using the generation process or the attention mechanism. However, all of these approaches solely operate on the image space, which has been demonstrated to be suboptimal for complex shifts (Kalluri, Xu, and Chandraker 2023). Differently, Chen et al. (2024) proposed to integrate a large language model (LLM) for UDA, while other works (Dunlap et al. 2023; Wang et al. 2024; Liu and Wang 2023; Huang et al. 2023) leveraged the language modality for domain generalisation, but they rely on short class descriptors, which are not semantically rich as crowd-sourced text. For example, Goyal et al. (2023) used class names as text descriptors, which are less semantically informative than free-form texts. To overcome these limitations, LaGTran (Kalluri, Majumder, and Chandraker 2024) used captions to generate pseudo-labels as a source of supervision for target samples. However, given the uncurated nature of captions, the generated supervision may be incorrect, often leading to overfitting the introduced noise.

VLMs in Knowledge Transfer. Vision-Language Models, such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), demonstrated success in capturing modality-invariant features, opening new avenues for knowledge transfer. Some works (Lai et al. 2023; Zhekai Du 2024; Li et al. 2024) leveraged the zero-shot ability of VLMs, applying or finetuning them to the target domain. In (Cho et al. 2023), the authors used prompt or adaptor learning to fine-tune the VLM to the target domain in a semi-supervised fashion. DAPL (Ge et al. 2022) learned domain-specific prompts to separate domain and class information in the CLIP visual feature space. DIFO (Tang et al. 2023) used prompt learning to adapt the VLM to the target domain and distilled it to a target model. In contrast, we propose a different paradigm using CLIP to estimate the uncertainty of pseudo-labels generated by a language model, by leveraging the multimodal semantic alignment of CLIP's feature spaces.

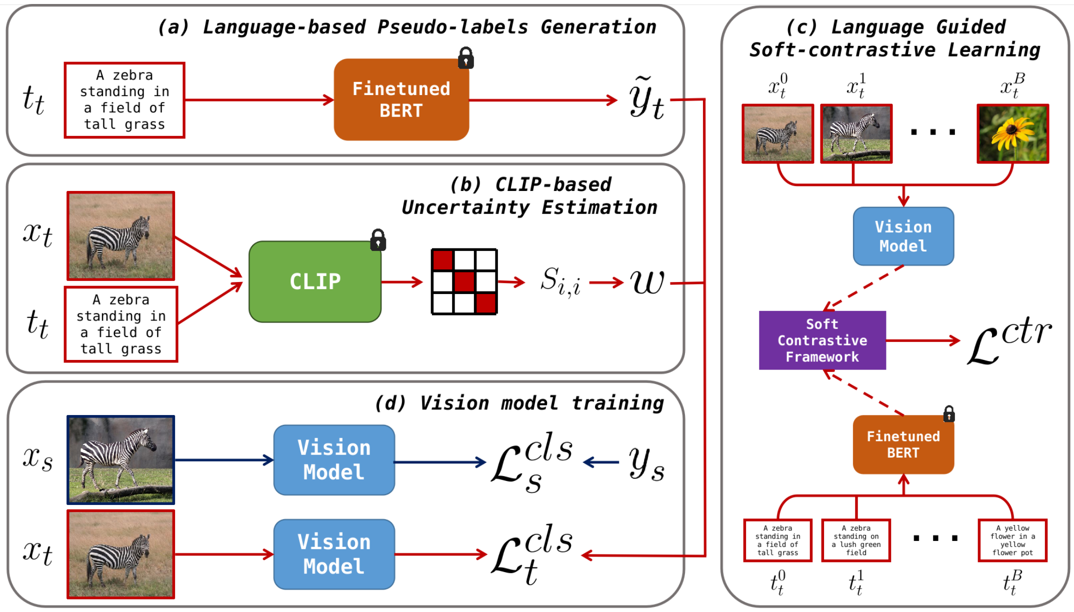

Figure 2: (a) We use the finetuned BERT model on target captions to generate pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ for target samples (Sec. 3.1). (b) From target images and texts, we compute normalised CLIP similarity scores $s^{i,i}$ used to estimate $\mathbf{w}$ as a measure of the reliability of the generated pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ (Sec. 3.2). (c) Feature representations of images go through a soft-contrastive framework, where they are attracted and repulsed to each other based on the similarity of their captions (dashed lines indicates features extraction). (d) A vision model is trained on both source and target images. On source images, we compute a classification loss $\mathcal{L}_s^{cls}$ using the ground truth labels $\mathbf{y}_s$. On target images we use the pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ as supervision and the classification loss $\mathcal{L}_t^{cls}$ is reweighted based on the estimated reliability score $\mathbf{w}$.

Contrastive Learning in Domain Adaptation. Self-supervised methods proved their effectiveness in learning generalised representations of visual data (Chen et al. 2020a; Caron et al. 2018; Chen et al. 2020b). Specifically, contrastive-based approaches (Chen et al. 2020a; Chen and He 2020) have demonstrated to improve the generalisation of deep learning models. In (van den Oord, Li, and Vinyals 2018; Chen et al. 2020a), the authors proposed a self-supervised contrastive framework that used augmentations to generate positive pairs and all the other training samples for the negative pairs. Other approaches (Wu, Wu, and Huang 2021; Chuang et al. 2020; Kalantidis et al. 2020) proposed strategies to optimise the selection of negative samples to improve the contrastive training. However, these methods include in the negative pairs also samples sharing the same class, disrupting the adaptation. Recent approaches (Chen et al. 2022a; Litrico, Del Bue, and Morerio 2023) proposed to exclude samples of the same class from the negative pairs, by looking at pseudo-labels (Chen et al. 2022a), historical predictions (Litrico, Del Bue, and Morerio 2023), or clustering assignments (Kang et al. 2022). Since this exclusion is based on the model predictions, it easily leads to confirmation bias. Differently, our contrastive framework does not require the selection of positive and negative pairs, overcoming the aforementioned issues. Specifically, we propose a soft-contrastive loss, where each pair simultaneously acts as both a positive and negative pair. The representations of each pair are then attracted and repulsed based on an estimated score of "*positiveness*" and "*negativeness*".

3 Proposed Method

An overview of TRUST is shown in Figure 2. Our aim is unsupervised domain adaptation (UDA), *i.e.* to learn a model for the target domain without having access to any ground-truth labels from it. Here, we have a labelled source domain $\mathcal{D}_s : \{x_s^i, t_s^i, y_s^i\}_{i=1}^{N_s}$, where x_s and y_s are source images and ground-truth labels, and t_s are source captions. Similarly, we have an unlabelled target domain $\mathcal{D}_t : \{x_t^i, t_t^i\}_{i=1}^{N_t}$. Captions are obtained from either associated metadata in web-collected images (Mahajan et al. 2018), or generated with image-to-text models (Li et al. 2023) and they are used *only* at training time. TRUST trains a vision model composed of a feature extractor $f : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^P$ and a classifier $h : \mathbb{R}^P \to \mathbb{R}^C$, where P is the length of the feature vector, and C is the number of classes.

3.1 Language Guided Domain Adaptation

We address the problem of UDA leveraging pseudo-labels generated from captions. Similarly to (Kalluri, Majumder, and Chandraker 2024), we fine-tune a BERT sentence classifier (Devlin et al. 2019) in a supervised fashion using captions and labels from the source domain. The trained BERT model is then frozen and inputted with target captions to generate pseudo-labels for target samples. Then, source labels and the obtained target pseudo-labels are used as supervision to train a vision model on both source and target domains simultaneously, as detailed in Sec. 3.4.

More formally, we fine-tune a pretrained DistilBERT (Sanh et al. 2019) model $\mathcal{B}$ on the source domain $\{t_s^i, y_s^i\}_{i=1}^{N_s}$

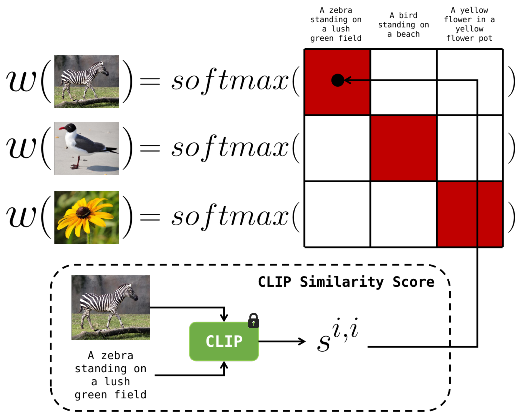

Figure 3: We compute CLIP similarity scores on each pair of target images and texts in a batch, obtaining the similarity matrix S. To calculate the reliability weight $\mathbf{w}$, we normalise the CLIP scores with a softmax on each row of S and select the values in the main diagonal, which measure the semantic correlation between much each image and its caption.

using source captions and labels, optimising this objective:

$$\underset{\phi}{\operatorname{argmin}}\ \mathbb{E}_{\{t_s^i, y_s^i\}\sim\mathcal{D}_s}\ \mathcal{L}_{\text{CE}}(\mathcal{B}_\phi(t_s^i), y_s^i), \tag{1}$$

where ϕ denotes the parameters of the BERT model and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss. Then, we use the fine-tuned model $\mathcal{B}$ to generate pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ on target captions $\mathbf{t_t}$ by running inference passes as follows:

$$\tilde{y}_t^i = \underset{C}{\operatorname{argmax}}\ \mathcal{B}_\phi(t_t^i). \tag{2}$$

3.2 CLIP-based Pseudo-labels Uncertainty Estimation

Although previous works (Kalluri, Majumder, and Chandraker 2024) showed that the language modality is more robust to complex shifts, blindly relying on the knowledge acquired from captions may lead to the generation of incorrect pseudo-labels $\tilde{\mathbf{y}}_\mathbf{t}$ due to: (a) the scarce quality of captions (*c.f.* Figure 1(b)), especially when they are crowd-sourced from the web; and (b) the domain shift that still exists in the language modality. Therefore, training an image classifier on such pseudo-labels (as in (Kalluri, Majumder, and Chandraker 2024)) risks to disrupt the adaptation process.

To mitigate these issues, we propose a novel strategy to estimate the uncertainty/reliability of the pseudo-labels generated from target captions, based on the capacity of captions to semantically describe the corresponding images. Such uncertainty scores are then used to reweight the classification loss, accordingly. With this aim, we use a pretrained CLIP model (Radford et al. 2021) to evaluate the correlation between images and corresponding captions. The underlying idea, shown in Figure 3, is that when the CLIP image/text similarity is high, the text accurately describes the image content. Consequently, we will assume that the pseudo-label generated with BERT (*c.f.* Sec. 3.1) is reliable. On the contrary, when the CLIP similarity is low, the text does not describe well the content of the image (e.g. Figure 1 (b)). This may occur when the text is of limited quality or because it does not capture the semantics of the image. In this case, the pseudo-label is considered unreliable. Differently from other uncertainty estimation strategies (Litrico, Del Bue, and Morerio 2023), which perform such estimation based on the output of the training model, our solution based on CLIP prevents the confirmation bias (Li, Socher, and Hoi 2020), since the CLIP estimation is not affected by the training of the TRUST's vision model, and CLIP's parameters are frozen during the adaptation process.

More formally, given a batch of image/caption pairs from the target domain, we compute the similarity matrix $S(E_{\text{im}}(\mathbf{x_t}), E_{\text{text}}(\mathbf{t_t}))$ as the cosine similarity between the embeddings of target images $\mathbf{x_t}$ and texts $\mathbf{t_t}$ obtained from the CLIP's image and text encoders E_{im} and E_{text}. The element $s^{i,j}$ of the similarity matrix indicates the semantic correlation between the i-th image and the j-th caption in the batch. Consequently, the diagonal elements $s^{i,i}$ indicates how much each target image is semantically related to its caption. However, this score is unbounded and needs to be normalised to measure the reliability of the pseudo-labels obtained from the captions. Therefore, we compute the softmax function for each row of the similarity matrix S to obtain a normalised score w_i, as follows:

$$w_i = \frac{\exp(s^{i,i})}{\sum_{j=1}^{B}\exp(s^{i,j})}. \tag{3}$$

The resulting score w_i is an estimation of the quality of the caption t_t^i to describe the image x_t^i. We use the softmax function for normalising $s^{i,i}$, to produce a score proportional to how much each image is semantically related to its caption (with respect to the other texts in the batch). The larger w_i (high reliability), the better the i-th image is described by its corresponding text, which leads to a higher confidence in the pseudo-label $\tilde{\mathbf{y}}_\mathbf{i}$. Conversely, a lower value of w_i (low reliability) means that the caption semantically describes the image as less as the other texts in the batch. Note that CLIP is pretrained and frozen during this process, to avoid adding an additional overhead for finetuning CLIP and preventing the confirmation bias.

3.3 Language Guided Soft-contrastive Learning

The language modality intrinsically benefits of a larger robustness to complex domain shifts compared to visual data, as demonstrated in (Kalluri, Xu, and Chandraker 2023). We posit that combining the benefits of language and vision modalities improves the performance on the target domain. LaGTran (Kalluri, Majumder, and Chandraker 2024) used a language model to generate pseudo-labels on target images for training a vision model. Despite its simplicity, this strategy alone does not encourage the vision model to inherit the robustness of the language model.

Therefore, we propose a novel language guided soft-contrastive learning framework to transfer the intrinsic robustness from the language to the vision model, by aligning their feature spaces. Differently than previous works (He et al. 2019; Chen et al. 2020a, 2022a), our contrastive framework treats each pair of images as both a positive and a negative pair, with a score of "*positiveness*" and "*negativeness*" based on the semantic similarity between image captions. In this way, the feature representation of each pair will be both attracted and repulsed with a strength proportional to how likely they share the same semantic content. This strategy leads to multiple benefits. Firstly, it aligns the language and vision feature spaces, transferring the intrinsic robustness to complex domain shifts from the language to the vision model. Secondly, it does not require to determine positive and negative pairs, since each pair plays simultaneously as a positive and a negative pair. This avoids using the vision model itself to discriminate between positive and negative pairs (Litrico, Del Bue, and Morerio 2023), reducing the effects of confirmation bias. Finally, our solution reduces the adverse effects of mistakenly assigning a pair as positive or negative. If two samples have incorrect pseudo-labels, the pair will not be strictly assigned to either the positive or negative class. Instead, their feature representations are both attracted and repulsed, limiting the impact of the incorrect assignment.

More formally, let $\mathcal{S} = \{a_w(x_t^i), a_s(x_t^i)\}_{i=1}^{B}$ be a batch composed of target samples, where we apply a weak $a_w \in \mathcal{A}_w$ and a strong augmentation $a_s \in \mathcal{A}_s$ drawn from distributions $\mathcal{A}_w$ and $\mathcal{A}_s$. Standard self-supervised contrastive methods (Chen et al. 2020a) optimise the following:

$$\mathcal{L}_{self}^{ctr} = -\sum_{i}^{B} \log \frac{exp(z_i \cdot \bar{z}_i/\tau)}{\sum_{j\in\mathcal{N}} exp(z_i \cdot z_j/\tau)}, \tag{4}$$

where $z_i = f(a_w(x_t^i))$ and $\bar{z}_i = f(a_s(x_t^i))$ are feature representations extracted from the weakly and strongly augmented i-th target sample, τ is a temperature parameter, and $\mathcal{N} : \{j | 1 \leq j \leq B, j \neq i\}$ is the set of indices of all the negative samples. In this formulation, only the augmented version of the same sample is treated as positive, while all the other samples are treated as negatives.

Inspired by (Khosla et al. 2020), we generalise this formulation to account for multiple positive samples. Each pair of images in the batch is treated as positive, with a score of positiveness depending on how similar their captions are, leading to the following objective:

$$\mathcal{L}^{ctr} = -\sum_{i}^{B} \log\left\{ \frac{1}{|\mathcal{S}|} \sum_{p=1}^{|\mathcal{S}|} \frac{\mathrm{sim}_{(i,p)} \cdot \exp(z_i \cdot \bar{z}_p/\tau)}{\sum_{j\in\mathcal{N}} (1-\mathrm{sim}_{(i,j)}) \cdot \exp(z_i \cdot z_j/\tau)} \right\}, \tag{5}$$

where $\mathrm{sim}_{(a,b)} = \frac{f^{\mathcal{B}}(t_t^a) \cdot f^{\mathcal{B}}(t_t^b)}{||f^{\mathcal{B}}(t_t^a)|| \cdot ||f^{\mathcal{B}}(t_t^b)||}$ is the cosine similarity between the DistilBERT features vectors for the text descriptions of target samples x_t^a and x_t^b. Differently than eq. (4), we treat each sample in the batch (e.g. $\bar{z}_p$) as positive of the i-th sample, requiring to include another summation over the cardinality of the multiviewed batch. This results in attracting the feature representations of a pair of images (x_t^a, x_t^b) with a strength equal to $\mathrm{sim}_{(a,b)}$, while repulsing them with a strength equal to $1 - \mathrm{sim}_{(a,b)}$.

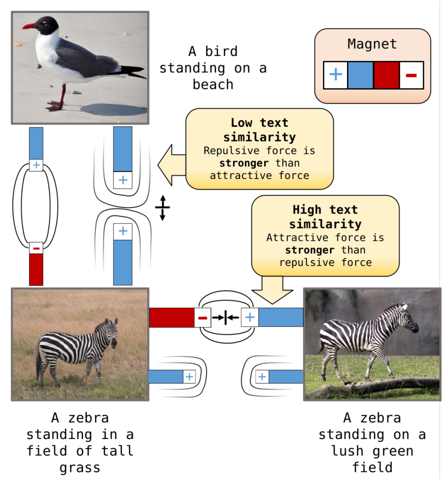

Figure 4: Representation of our soft-contrastive framework, where each pair of images is both attracted and repulsed based on their caption similarity: if two images have semantically similar captions, they are more attracted than repulsed and vice versa. Attraction/Repulsion is shown by magnets with opposite/same polarity. The strength of each force is indicated by magnet size and field lines.

3.4 Overall Framework

To adapt the source to the target domain, we train a vision model on both the source and target domains, combining the source labels and the target pseudo-labels obtained by the finetuned BERT model in Sec. 3.1. Differently than (Kalluri, Majumder, and Chandraker 2024), we reweight the classification loss for target samples based on the reliability of the pseudo-labels estimated through CLIP scores in Sec. 3.2. The higher the estimated reliability, the more it will contribute to the classification loss. Hence, we train TRUST with the following classification losses:

$$\mathcal{L}_s^{cls} = \frac{1}{|\mathcal{D}_s|} \sum_{i=1}^{|\mathcal{D}_s|} \mathcal{L}_{\mathrm{CE}}(h(f(a_w(x_s^i))), y_s^i), \tag{6}$$

$$\begin{aligned}\mathcal{L}_t^{cls} = \frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{|\mathcal{D}_t|} w_i \cdot \mathcal{L}_{\mathrm{CE}}(h(f(a_w(x_t^i))), \tilde{y}_t^i) + \\ (1-w_i) \cdot \mathcal{L}_{\mathrm{CE}}(h(f(a_w(x_t^i))), p_t^i),\end{aligned} \tag{7}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss, w_i is the reliability weight calculated from CLIP (*c.f.* Section 3.2), $p_t^i = h(f(a_s(x_t^i)))$ are predictions obtained by the vision model on strongly augmented samples, and a_w and a_s are weak and strong augmentations, respectively. Finally, on target images and captions, we employ our proposed soft-contrastive framework optimising $\mathcal{L}^{ctr}$.

The overall loss function is the following:

$$\mathcal{L} = \mathcal{L}_s^{cls} + \mathcal{L}_t^{cls} + \mathcal{L}^{ctr}. \tag{8}$$

4 Experimental Results

Datasets, implementation details and additional analyses are reported in the supplementary material.

	Real →			Clipart →			Sketch →			Painting →			Avg.
	C	S	P	R	S	P	R	C	P	R	C	S	
Zero-shot Classification													
CLIP (Radford et al. 2021)	72.39	60.90	66.81	81.37	60.90	66.81	81.37	72.39	66.81	81.37	72.39	60.90	70.38
CLIP* (Radford et al. 2021)	73.45	62.81	67.07	**81.71**	62.81	<u>67.07</u>	<u>81.71</u>	73.45	67.07	<u>81.71</u>	73.45	62.81	71.26
UDA													
Source only	63.02	49.47	60.48	70.52	56.09	52.53	70.42	65.91	54.47	73.34	60.09	48.25	60.38
MCD (Saito et al. 2017)	39.40	25.20	41.20	44.60	31.20	25.50	34.50	37.30	27.20	48.10	31.10	22.80	34.01
MDD (Zhang et al. 2019)	52.80	41.20	47.80	52.50	42.10	40.70	54.20	54.30	43.10	51.20	43.70	41.70	47.11
CGDM (Du et al. 2021)	49.40	38.20	47.20	53.50	36.90	35.30	55.60	50.10	43.70	59.40	37.70	33.50	45.04
SCDA (Li et al. 2021)	54.00	42.50	51.90	55.00	44.10	39.30	53.20	55.60	44.70	56.20	44.10	42.00	48.55
SSRT-B (Sun et al. 2022)	69.90	58.90	66.00	75.80	59.80	60.20	73.20	70.60	62.20	71.40	61.70	55.20	65.41
MemSAC (Kalluri, Sharma, and Chandraker 2022)	63.49	42.14	60.32	72.33	54.92	46.14	73.46	68.04	52.75	74.42	57.79	43.57	59.11
CDTrans (Xu et al. 2021)	66.20	52.90	61.50	72.60	58.10	57.20	72.50	69.00	59.00	72.10	62.90	53.90	63.16
PMTrans (Zhu, Bai, and Wang 2023)	74.10	61.10	**70.00**	79.30	63.70	62.70	77.50	73.80	62.60	79.80	69.70	61.20	69.63
Language-guided UDA													
TextMatch (Kalluri, Majumder, and Chandraker 2024)	71.36	64.30	65.32	81.25	65.65	64.85	81.09	72.65	63.94	81.08	70.84	**64.17**	70.14
nGramMatch (Kalluri, Majumder, and Chandraker 2024)	68.92	59.82	63.15	76.35	61.72	62.87	76.35	69.28	62.51	76.04	68.52	60.52	67.17
CLIP+TextMatch (Kalluri, Majumder, and Chandraker 2024)	72.89	63.56	66.97	<u>81.53</u>	62.98	66.74	**81.89**	72.81	65.97	**81.87**	72.07	63.34	71.05
LaGTran (Kalluri, Majumder, and Chandraker 2024)	<u>77.30</u>	<u>68.25</u>	67.35	81.31	<u>67.03</u>	66.81	80.78	<u>75.62</u>	<u>68.08</u>	79.23	<u>73.80</u>	63.44	<u>72.41</u>
TRUST	**81.04**	**71.12**	<u>69.95</u>	81.42	**69.15**	**69.03**	81.44	**79.37**	**72.77**	81.61	**78.33**	<u>64.11</u>	**74.95**

Table 1: Classification accuracy (%) under the classical shifts of DomainNet-345. All methods use the Swin-base backbone. The symbol * indicates models finetuned on the source data.

	GeoImnet		GeoPlaces		Total Avg.
	U → A	A → U	U → A	A → U	
Zero-shot classification					
CLIP (Radford et al. 2021)	49.84	53.83	43.41	54.34	50.36
CLIP* (Radford et al. 2021)	57.79	59.12	48.91	55.89	55.42
UDA					
Source only	52.46	51.91	44.90	36.85	46.53
CDAN (Long et al. 2018)	54.48	53.87	42.88	36.21	46.86
MemSAC (Kalluri, Sharma, and Chandraker 2022)	53.02	54.37	42.05	38.33	46.94
ToAlign (Wei et al. 2024)	55.67	55.92	42.32	38.40	48.08
MDD (Zhang et al. 2019)	51.57	50.73	42.54	39.23	46.02
DALN (Chen et al. 2022b)	55.36	55.77	41.06	40.41	48.15
PMTrans (Zhu, Bai, and Wang 2023)	56.76	57.60	46.18	40.33	50.22
Language-guided UDA					
TextMatch (Kalluri, Majumder, and Chandraker 2024)	49.68	54.82	53.06	50.11	51.92
nGramMatch (Kalluri, Majumder, and Chandraker 2024)	49.53	51.02	51.70	49.87	50.93
CLIP+TextMatch	50.11	54.92	50.36	52.14	51.88
LaGTran (Kalluri, Majumder, and Chandraker 2024)	<u>63.67</u>	<u>64.16</u>	<u>56.14</u>	<u>57.02</u>	<u>60.24</u>
TRUST	**65.77**	**67.02**	**59.89**	**60.14**	**63.20**

Table 2: Classification accuracy (%) under the geographical shifts of GeoNet. All methods use a ViT-base backbone.

Baselines. We employ two text baselines (TextMatch and nGramMatch) as in (Kalluri, Majumder, and Chandraker 2024) to evaluate the effect of using the source data for finetuning the language model, instead of using directly the target captions. We also compare with CLIP employing two baselines: CLIP zero-shot inference (Dunlap et al. 2023), performed by incorporating the domain information into the text prompt of CLIP (*e.g.*, *A clipart of a* `<class>`); CLIP model finetuned on the source to adapt it to the data distribution. Finally, we employ a training-free baseline by ensembling CLIP and TextMatch (CLIP+TextMatch) by averaging the CLIP's image-text similarity and the TextMatch's caption-label similarity, and then computing the argmax to obtain the ensemble predictions.

4.1 Results

DomainNet-345. Table 1 presents results on DomainNet-345 that, exhibiting classical shifts, leads to generally higher scores than in GeoNet. CLIP and CLIP* achieve a notable accuracy of 70.38% and 71.26%. TRUST outperforms standard UDA approaches, CLIP and CLIP* by $+5.32\%$, 4.57% and $+3.69\%$, respectively. On non-real to real transfers, TRUST performs worse than CLIP and CLIP+TextMatch. A possible explanation is that CLIP is trained on a massive amount of real-world data, potentially eliminating any domain shifts between the train and test settings. Differently, TRUST is trained on non-real images (the source domain), which instead exhibit domain shifts with testing data. Nonetheless, TRUST achieves overall superior performance across diverse target domains, demonstrating its adaptability to different domain shifts. Compared to LaGTran, which also uses text during training, TRUST improves accuracy on average by $+2.54\%$, demonstrating the effectiveness of the introduced components. Moreover, TRUST performs best in almost all the transfer scenarios, achieving the best or second best results on 9 out of 12 scenarios.

GeoNet. Table 2 presents results of TRUST on geographical shifts of GeoNet. Our method outperforms the source-only baseline by $+16.67\%$. Compared to LaGTran (Kalluri, Majumder, and Chandraker 2024), we improve performance by $+2.96\%$, achieving better performance on all the shift settings of both GeoImnet and GeoPlaces. We also compare our method with the zero-shot performance of CLIP and CLIP+TextMatch. Despite CLIP being trained on a significantly larger amount of data, our approach obtains a gain in performance of $+12.84\%$. While in most of the settings the CLIP+TextMatch performs better than CLIP and TextMatch used alone, results are still lower than TRUST . We hypothesise that CLIP and TextMatch make coherent mistakes, leading to a low improvement when ensembling the two models. Finally, compared to the CLIP model finetuned on source data (CLIP*), TRUST improves accuracy by $+7.78\%$.

VisDA. Although in TRUST, CLIP is not finetuned and used only for uncertainty estimation, we compare with other approaches that use CLIP for UDA and we present results on VisDA in Tab. 3. TRUST outperforms PADCLIP by $+2.1\%$ on average, achieving the best performance on 4 out of 12 classes while maintaining competitive accuracies on the other classes (see Tab. 1 in the supplementary material). These results also demonstrate the effectiveness of TRUST on large-scale datasets.

4.2 Analysis

Ablation Studies. In Table 4 (Left), we report ablation studies for the TRUST components on GeoNet. When using only the pseudo-labels, TRUST achieves the lowest accuracy of

Method	Avg. Acc.
DAPL (Ge et al. 2022)	86.9
DAMP (Zhekai Du 2024)	88.4
UniMoS (Li et al. 2024)	88.1
PADCLIP (Lai et al. 2023)	88.5
TRUST	**90.6**

Table 3: UDA performance (%) on VisDA dataset with ResNet101 compared with CLIP-based approaches. (Full table in Supplementary Material.)

Hard Contrastive	Soft Contrastive	CLIP Uncertainty	Avg. Acc.
✗	✗	✗	62.31
✓	✗	✗	62.86
✗	✓	✗	65.38
✗	✗	✓	63.41
✗	✓	✓	**66.40**

Method	Avg. Acc.
TRUST w/ CLIP	61.36
TRUST w/ CLIP*	63.32
TRUST w/ CLIP-Text	64.11
TRUST w/ BERT*	**66.40**

Table 4: (Left) Ablation studies of components of the proposed method. (Right) Classification accuracy (%) on GeoImnet comparing CLIP and BERT for generating pseudo-labels (Sec. 3.1). The symbol * indicates models finetuned on the source data.

62.31%. When adding the standard contrastive loss (Chen et al. 2020a) (second row), the improvement in performance remains negligible. But when using our proposed language guided soft-contrastive loss (Sec. 3.3), we boost the performance by $+2.52\%$ (third row). The fourth row presents the results enabling only our CLIP-based uncertainty estimation (Sec. 3.2), which brings a gain in performance of $+1.10\%$. Finally, in the last row, we show the gain obtained by the full TRUST model which, enabling both the two introduced components, further improves performance by $+4.09\%$.

Generating pseudo-labels with CLIP. Table 4 (Right) presents results using CLIP (Dunlap et al. 2023) for generating pseudo-labels in Sec. 3.1, as an alternative to finetuning BERT. We compare three different CLIP-based approaches: the CLIP zero-shot inference (TRUST w/ CLIP), the CLIP model finetuned on source data (TRUST w/ CLIP*) and the generation of pseudo-labels by evaluating the similarity between captions and class names in the CLIP's text feature space (TRUST w/ CLIP-Text). Results show that using CLIP (instead of BERT) for obtaining the pseudo-labels leads to lower performance. Moreover, text-only based approaches, like CLIP-Text and BERT, achieve better results, justifying the assumption that language models benefit of a larger robustness to domain shift than vision models.

Effectiveness of CLIP-based uncertainty estimation. Figure 5 illustrates the effectiveness of the proposed CLIP-based uncertainty estimation strategy, showing the distribution of the reliability weights $\mathbf{w}$ (Sec. 3.2) for target samples with correct and wrong pseudo-labels $_{\mathbf{t}}$. Figure 5 shows that the two distributions are clearly separable and samples with correct pseudo-labels have high values of $\mathbf{w}$ (high reliability), while samples with wrong pseudo-labels have lower values of $\mathbf{w}$ (low reliability). These results validate our strategy to use CLIP to identify low-quality text descriptions, which is able to down-weight wrong pseudo-labels and to reduce their contribution in the classification loss.

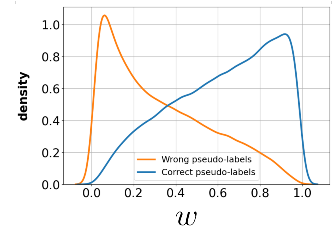

Figure 5: Probability density function of reliability weights estimated in Sec. 3.2 for target samples having correct and wrong pseudo-labels.

Figure 6: Visualisation of nearest neighbors for 2 target images (grey borders) comparing the standard contrastive loss (Chen et al. 2020a) (top 2 rows), and our soft-contrastive loss (*c.f.* Sec. 3.3) (bottom 2 rows). Images with green/red borders represent correctly/incorrectly retrieved images.

Effects of the soft-contrastive loss. Figure 6 shows a qualitatively analysis of the effectiveness of our soft-contrastive loss in aggregating the representations of semantically similar samples that belong to the same class. We show the top-3 nearest neighbour retrievals using features computed by the vision model, when trained with the standard hard contrastive loss (He et al. 2019; Chen et al. 2020a), or with the proposed soft-contrastive loss. Our loss produces a more fine-grained retrieval even in presence of semantically similar classes (e.g. *streetcar*, *cable_car*, *shuttle_bus*). On the contrary, the hard contrastive loss leads to less accurate retrievals, likely due to a coarser-grained feature space, where visually similar classes are aggregated even if they represent different semantic concepts (e.g. *farmers_market* and *plaza*).

5 Conclusions

In this work, we introduced TRUST, a novel approach to UDA for image classification that leverages textual data to guide the adaptation to the target domain. TRUST generates pseudo-labels for target samples from captions and estimates their uncertainty using CLIP to mitigate the impact of wrong pseudo-labels in the classification loss. We also proposed a novel soft-contrastive learning framework that aligns vision and language feature space, to transfer the shift robustness from the language to the vision model. Our extensive evaluations on DomainNet and GeoNet benchmarks demonstrated that TRUST outperforms the current state-of-the-art in both classical and complex domain shifts.

References

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *European Conference on Computer Vision*.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022a. Contrastive Test-time Adaptation. In *CVPR*.

Chen, L.; Chen, H.; Wei, Z.; Jin, X.; Tan, X.; Jin, Y.; and Chen, E. 2022b. Reusing the Task-specific Classifier as a Discriminator: Discriminator-free Adversarial Domain Adaptation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7171–7180.

Chen, S.; Zhang, Y.; Jiang, W.; Lu, J.; and Zhang, Y. 2024. VLLaVO: Mitigating Visual Gap through LLMs. arXiv:2401.03253.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. 2020b. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Chen, X.; and He, K. 2020. Exploring Simple Siamese Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15745–15753.

Cho, J.; Nam, G.; Kim, S.; Yang, H.; and Kwak, S. 2023. PromptStyler: Prompt-driven Style Generation for Source-free Domain Generalization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15656–15666. Los Alamitos, CA, USA: IEEE Computer Society.

Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; and Jegelka, S. 2020. Debiased Contrastive Learning. *ArXiv*, abs/2007.00224.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Du, Z.; Li, J.; Su, H.; Zhu, L.; and Lu, K. 2021. Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3936–3945.

Dunlap, L.; Mohri, C.; Guillory, D.; Zhang, H.; Darrell, T.; Gonzalez, J. E.; Raghunathan, A.; and Rohrbach, A. 2023. Using Language to Extend to Unseen Domains. arXiv:2210.09520.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1180–1189. JMLR.org.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2022. Domain Adaptation via Prompt Learning. *IEEE transactions on neural networks and learning systems*, PP.

Goyal, S.; Kumar, A.; Garg, S.; Kolter, Z.; and Raghunathan, A. 2023. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19338–19347. Los Alamitos, CA, USA: IEEE Computer Society.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1989–1998. PMLR.

Huang, Z.; Zhou, A.; Lin, Z.; Cai, M.; Wang, H.; and Lee, Y. J. 2023. A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11651–11661.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard Negative Mixing for Contrastive Learning. *ArXiv*, abs/2010.01028.

Kalluri, T.; and Chandraker, M. 2022. Cluster-to-adapt: Few Shot Domain Adaptation for Semantic Segmentation across Disjoint Labels. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 4120–4130. Los Alamitos, CA, USA: IEEE Computer Society.

Kalluri, T.; Majumder, B.; and Chandraker, M. 2024. Tell, Don't Show! Language Guidance Eases Transfer Across Domains in Images and Videos. *ICML*.

Kalluri, T.; Sharma, A.; and Chandraker, M. 2022. MemSAC: Memory Augmented Sample Consistency for Large Scale Domain Adaptation. *ArXiv*, abs/2207.12389.

Kalluri, T.; Xu, W.; and Chandraker, M. 2023. GeoNet: Benchmarking Unsupervised Adaptation across Geographies. *CVPR*.

Kang, G.; Jiang, L.; Wei, Y.; Yang, Y.; and Hauptmann, A. 2022. Contrastive Adaptation Network for Single- and Multi-Source Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1793–1804.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Lai, Z.; Vesdapunt, N.; Zhou, N.; Wu, J.; Huynh, C. P.; Li, X.; Fu, K. K.; and Chuah, C.-N. 2023. PADCLIP: Pseudo-labeling with Adaptive Debiasing in CLIP for Unsupervised Domain Adaptation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16109–16119.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. *ArXiv*, abs/2002.07394.

Li, S.; Xie, M.; Lv, F.; Liu, C. H.; Liang, J.; Qin, C.; and Li, W. 2021. Semantic Concentration for Domain Adaptation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9082–9091.

Li, X.; Li, Y.; Du, Z.; Li, F.; Lu, K.; and Li, J. 2024. Split to Merge: Unifying Separated Modalities for Unsupervised Domain Adaptation. In *2024 IEEE Conference on Computer Vision and Pattern Recognition*.

Litrico, M.; Del Bue, A.; and Morerio, P. 2023. Guiding Pseudo-labels with Uncertainty Estimation for Source-free Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, G.; and Wang, Y. 2023. TDG: Text-guided Domain Generalization. arXiv:2308.09931.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 1647–1657. Red Hook, NY, USA: Curran Associates Inc.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2208–2217. JMLR.org.

Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and van der Maaten, L. 2018. Exploring the Limits of Weakly Supervised Pretraining. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, 185–201. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-01215-1.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1406–1415.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2017. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3723–3732.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sharma, A.; Kalluri, T.; and Chandraker, M. 2021. Instance Level Affinity-Based Transfer for Unsupervised Domain Adaptation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5357–5367.

Sun, T.; Lu, C.; Zhang, T.; and Ling, H. 2022. Safe Self-Refinement for Transformer-based Domain Adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7181–7190. Los Alamitos, CA, USA: IEEE Computer Society.

Tan, S.; Peng, X.; and Saenko, K. 2020. Class-Imbalanced Domain Adaptation: An Empirical Odyssey. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 585–602. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-66414-5.

Tang, S.; Su, W.; Ye, M.; and Zhu, X. 2023. Source-Free Domain Adaptation with Frozen Multimodal Foundation Model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23711–23720.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv*, abs/1807.03748.

Wang, Z.; Zhang, L.; Wang, L.; and Zhu, M. 2024. LanDA: Language-Guided Multi-Source Domain Adaptation. *ArXiv*, abs/2401.14148.

Wei, G.; Lan, C.; Zeng, W.; Zhang, Z.; and Chen, Z. 2024. ToAlign: task-oriented alignment for unsupervised domain adaptation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713845393.

Wu, C.; Wu, F.; and Huang, Y. 2021. Rethinking InfoNCE: How Many Negative Samples Do You Need? *ArXiv*, abs/2105.13003.

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1426–1435. Los Alamitos, CA, USA: IEEE Computer Society.

Xu, T.; Chen, W.; Wang, P.; Wang, F.; Li, H.; and Jin, R. 2021. CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation. *ArXiv*, abs/2109.06165.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. I. 2019. Bridging Theory and Algorithm for Domain Adaptation. In *International Conference on Machine Learning*.

Zhekai Du, F. L. K. L. L. Z. J. L., Xinyao Li. 2024. Domain-Agnostic Mutual Prompting for Unsupervised Domain Adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, J.; Bai, H.; and Wang, L. 2023. Patch-Mix Transformer for Unsupervised Domain Adaptation: A Game Perspective. arXiv:2303.13434.

Reproducibility Checklist

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) yes

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) yes

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) no

2.4. Proofs of all novel claims are included (yes/partial/no) no

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) yes

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) yes

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) NA

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) partial

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) NA

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) no

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) no

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) no