Category Contrast for Unsupervised Domain Adaptation in Visual Tasks

Jiaxing Huang [1], Dayan Guan [1], Aoran Xiao [1], Shijian Lu* [1], Ling Shao [2]
[1] Nanyang Technological University, Singapore [2] Terminus Group, China
{Jiaxing.Huang, Dayan.Guan, Aoran.Xiao, Shijian.Lu}@ntu.edu.sg, ling.shao@ieee.org

Abstract

Instance contrast for unsupervised representation learning has achieved great success in recent years. In this work, we explore the idea of instance contrastive learning in unsupervised domain adaptation (UDA) and propose a novel Category Contrast technique (CaCo) that introduces semantic priors on top of instance discrimination for visual UDA tasks. By considering instance contrastive learning as a dictionary look-up operation, we construct a semantics-aware dictionary with samples from both source and target domains where each target sample is assigned a (pseudo) category label based on the category priors of source samples. This allows category contrastive learning (between target queries and the category-level dictionary) for category-discriminative yet domain-invariant feature representations: samples of the same category (from either source or target domain) are pulled closer while those of different categories are pushed apart simultaneously. Extensive UDA experiments in multiple visual tasks (e.g., segmentation, classification and detection) show that CaCo achieves superior performance as compared with state-of-the-art methods. The experiments also demonstrate that CaCo is complementary to existing UDA methods and generalizable to other learning setups such as unsupervised model adaptation, open-/partial-set adaptation etc.

1. Introduction

Though deep neural networks (DNNs) [20, 57] have revolutionized various computer vision tasks [4, 20, 47, 57], they generally perform not well on new domains due to the *cross-domain mismatch*. Unsupervised domain adaptation (UDA) aims to mitigate the cross-domain mismatch via exploiting unlabelled target-domain samples. To achieve this purpose, researchers have designed different unsupervised training objectives on target-domain samples to train a well-performed model in target domain [7, 30, 40, 59, 62, 63, 69]. The existing unsupervised losses can be broadly classified into three categories: 1) *adversarial loss* that enforces source-like target representations [38, 40, 53, 59, 60, 62, 63]; 2) *image translation loss* that translates source images to have target-like styles and appearance [8, 27, 36, 72, 74]; and 3) *self-training loss* that re-trains networks iteratively with confidently pseudo-labelled target samples [15, 36, 80, 81].

Unsupervised representation learning [5, 19, 41, 44, 58, 68, 73, 77, 78] addresses a related problem, $i.e.$, unsupervised network pre-training which aims to learn discriminative embeddings from unlabelled data. In recent years, instance contrastive learning [5, 19, 42, 58, 68, 73] has led to major advances in unsupervised representation learning. Despite different motivations, instance contrast methods can be thought of as a dictionary look-up task [19] that trains a visual encoder by matching an encoded query q with a dictionary of encoded keys k: the encoded query should be similar to the encoded positive keys and dissimilar to encoded negative keys. With no labels available for unlabelled data, the positive keys are often randomly augmented versions of query samples, and all other samples are considered as negative keys.

In this work, we explore the idea of instance contrast in UDA. Considering contrastive learning as a dictionary look-up task, we hypothesize that a UDA dictionary should be category-aware and domain-mixed with keys from both source and target domains. Intuitively, a category-aware dictionary with category-balanced keys will encourage to learn category-discriminative yet category-unbiased representations, while the keys from both source and target domains will allow to learn invariant representations within and across the two domains, both being aligned with the objective of UDA.

With above motivation, this paper presents *Category Contrast* (CaCo) as a way of building category-aware and domain-mixed dictionaries with corresponding contrastive losses for UDA. As shown in Fig. 1, this dictionary includes keys that are evenly sampled in both categories and domains, where each target key comes with a predicted pseudo category. Take the illustrative dictionary $\mathbf{K} = \{k_m^c\}_{1 \leq c \leq C, 1 \leq m \leq M}$ as an example. Each category c will have M keys while each domain has $(C \times M)/2$ keys.

*Corresponding author.

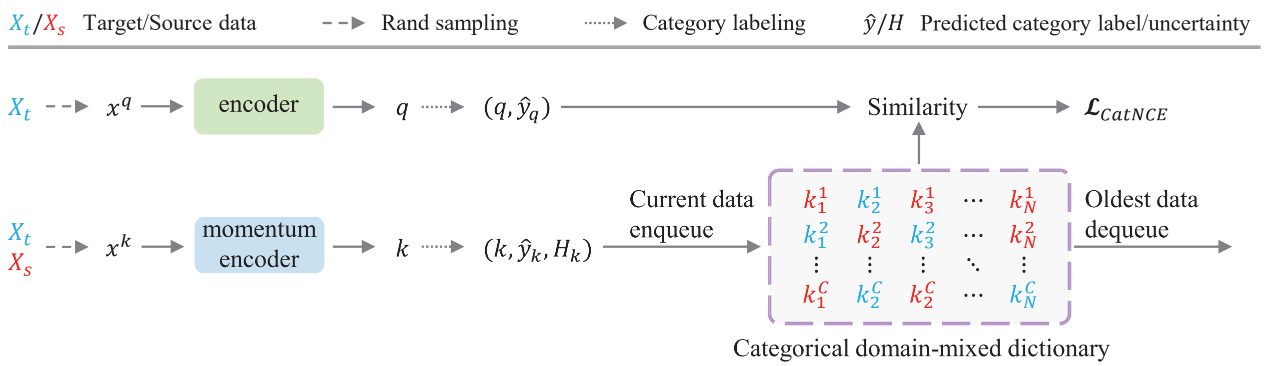

Figure 1. The proposed Category Contrast trains an unsupervised domain adaptive encoder by matching a query q (from an unlabelled target sample $x^q \in X_t$) to a dictionary of keys via a category contrastive loss $\mathcal{L}_{\text{CatNCE}}$. The dictionary keys are domain-mixed from both source domain X_s (in red with labels) and target domain X_t (in blue with pseudo labels), which allows to learn invariant representations within and across the two domains. They are also category-ware and category-balanced allowing to learn category-discriminative yet category-unbiased representations. Note the *category-balanced* means that each query q is compared with all the dictionary keys (in loss computation) that are evenly distributed over all data categories which mitigates data imbalance issue.

The network learning will thus strive to minimize a *category contrastive loss* $\mathcal{L}_{\text{CatNCE}}$ between target queries and dictionary keys: samples of the same category are pulled close while those of different categories are pushed away. This naturally leads to category-discriminative yet domain-invariant representations that perfectly match the objective of UDA.

With the category-aware and domain-mixed dictionary together with the category contrastive loss, the proposed Category Contrast tackles the UDA challenges with three desirable features: 1) It concurrently minimizes the intra-category variation and maximizes the inter-category distance with the *category-aware* dictionary design; 2) It achieves inter-domain and intra-domain alignment simultaneously thanks to the *domain-mixed* dictionary design by including both source and target samples; 3) It greatly mitigates the data balance issue due to the *category-balanced* dictionary design which allows to compute contrast losses evenly across all categories during learning.

We summary the contributions of this paper as follows: (1) we explore instance contrast for UDA, aiming to learn discriminative representation for unlabelled target-domain samples. (2) we propose Category Contrast that builds a category-aware and domain-mixed dictionary with a category contrastive loss. It encourages to learn category-discriminative yet domain-invariant representation that perfectly matches the objective of UDA. (3) extensive experiments demonstrate that our CaCo achieves superior UDA performance consistently as compared with state-of-the-art. Additionally, CaCo complements previous UDA approaches and generalizes to other learning setups that involves unlabeled data.

2. Related Works

This work relates to two main fields of research, namely, unsupervised learning in unsupervised domain adaptation and instance contrast in unsupervised representation learning.

Unsupervised domain adaptation aims to leverage unlabelled target data to improve network performance in target domain. To learn from unlabelled target data, most existing works propose various unsupervised losses. We roughly sort them into three subcategories. The first subcategory is *adversarial loss* that enforces source-like target representation in terms of encoded features [7, 16, 38, 52, 62, 75], generated predictions [28, 40, 51, 53, 59] or converted latent representations [29, 60, 63]. The second category is *image translation loss* that generates source data with target-like styles and appearance via GANs [8, 10, 36] and spectrum matching [25, 72]. The third category is *self-training loss* that re-trains the network iteratively with pseudo-labelled target samples [14, 24, 26, 36, 64, 72, 80, 81].

We tackle UDA from a new perspective of instance contrastive learning, and propose a novel Category Contrast (CaCo) that introduces a generic category contrastive loss that can work for various UDA tasks. To the best of our knowledge, CaCo is the first effort to investigate instance contrastive learning for UDA.

Instance Contrastive Learning [5, 19, 42, 58, 68, 73] aims to learn an embedding space where positive samples are pulled close to an anchor and negative samples are pushed away. Despite different motivations, instance contrastive learning can be viewed as a dictionary look-up task [19] that trains a visual encoder by matching an encoded query q with a dictionary of encoded keys k: q should be similar to positive k and dissimilar to negative k. Three typical dictionary creation strategies have been proposed. The first builds a *memory bank* [68] that stores the keys of all samples every training epoch. The second one builds a *momentum-encoded queue* [19] that collects encoded samples online as keys. The third one creates an *end-to-end* dictionary [5, 58, 73] that takes encoded samples of the current training batch as keys. Instance contrast with various dictionaries helps to learn better unsupervised representations clearly.

On the other hand, existing instance contrastive learning methods [5, 19, 42, 58, 68, 73] were designed for unsupervised representation, which has two main limitations in UDA: 1). With little category priors, existing instance contrast techniques learn rich low-level features without capturing much high-level semantic information. This is sub-optimal to many visual recognition tasks ($e.g.$, segmentation, detection and classification) that require discriminative semantic features. Recent studies [56, 61] verify this issue; 2). Most existing instance contrastive learning methods [5, 19, 42, 58, 68, 73] employ a super-large/category-agnostic dictionary that could introduce category collision [56], where negative pairs share the same semantic category but are undesirably pushed away in the feature space. This impairs most learning setups that require semantic-level discrimination including various visual UDA tasks. The proposed CaCo introduces a categorical domain-mixed dictionary which introduces category priors and addresses the two problems effectively.

Other recent related contrastive learning works. [35] explores contrastive learning with semantic distributions and proposes semantic distribution-aware contrastive adaptation that contrasts each sample with estimated category centroids. [1, 65] explore pixel-level contrast with a memory bank for supervised and semi-supervised semantic segmentation.

3. Method

3.1. Task Formulation

This work tackles the task of unsupervised domain adaptation, where labelled source-domain samples $\{X_s, Y_s\}$ are accessible while only unlabelled data X_t are available in the target domain. The learning objective is to train a well-performing network G for X_t. The *baseline* performance is acquired by training network G with annotated source-domain sample only:

$$\mathcal{L}_{sup} = l(G(X_s), Y_s), \tag{1}$$

where $l(\cdot)$ denotes an accuracy-related loss.

3.2. Preliminaries of Instance Contrastive Learning

The idea of instance contrastive learning [18] can be considered as training an encoder (feature extractor) for a *dictionary look-up* task. Given a query q and a dictionary that consists of a number of keys $\{k_0, k_1, ..., k_N\}$, instance discriminative representations are learnt with an instance contrastive loss [18] ($e.g.$, InfoNCE [42]), minimization of which will pull q close to its positive key and push it away from all other keys (considered negative for q):

$$\mathcal{L}_{\text{InfoNCE}} = \sum_{x_q \in X} -\log \frac{\sum_{i=0}^{N} \mathbb{1}(k_i \in q) \exp(q \cdot k_i / \tau)}{\sum_{i=0}^{N} \exp(q \cdot k_i / \tau)} \tag{2}$$

where $\mathbb{1}(k_i \in q) = 1$ if k_i is the positive key of q and $\mathbb{1}(k_i \in q) = 0$ otherwise. Parameter τ is a temperature parameter [68]. In general, the query representation is $q = f_{\text{q}}(x^q)$ where f_{q} is an encoder network and x^q is a query sample (likewise in $k = f_{\text{k}}(x^k)$).

3.3. Category Contrast for Unsupervised Domain Adaptation

We tackle UDA from a perspective of instance contrastive learning. Specifically, we design Category Contrast that builds a category-aware and domain-mixed dictionary to learn category-discriminative yet domain-invariant representations under the guidance of a category contrastive loss.

Overview. For *supervised training* over a labelled source domain, we feed source samples $\{X_s, Y_s\}$ to a model G and optimize G with Eq. 1. In this work, G consists of an encoder f_q and a classifier h that classifies the encoded embeddings into pre-defined categories, $i.e.$, $G(\cdot) = h(f_q(\cdot))$. For *unsupervised training* over an unlabelled target domain, the training involves a query *encoder* f_q and a key *momentum encoder* f_k (the momentum update of f_q, $i.e.$, $\theta_{f_k} = b\theta_{f_k} + (1-b)\theta_{f_q}$, and b is a momentum coefficient) as illustrated in Fig. 1. During the training, we evenly sample the key x_k from both source and target domains ($i.e.$, X_s and X_t) and feed them to the key encoder f_k to build a category-aware dictionary $\mathbf{K}$. We sample query x_q from the target domain (i.e. X_t) only and feed them to the query encoder f_q for category contrastive learning with the category-aware dictionary $\mathbf{K}$.

Categorical domain-mixed dictionary. One key component in the proposed CaCo is a category-aware and domain-mixed dictionary with keys from both source and target domains. The dictionary allows to perform category contrastive learning: the embeddings of the *same category* are pulled close together while those of *different categories* are pushed apart. The category awareness encourages the network to learn category-discriminative embeddings. This feature is critical to various visual tasks ($e.g.$, segmentation, classification and detection) that require to learn discriminative features and classify them to pre-defined categories. In addition, the dictionary is domain-mixed which encourages to learn invariant representations within and across domains as category contrast is computed between target queries and keys from both source and target domains.

As stated in the Overview, given an encoded key $k = f_k(x_k)$ ($x_k \in X_s \cup X_t$), the classifier h predicts a category label $\hat{y}_k$ and converts k into a categorical key k^c which is further queued into the categorical dictionary $\mathbf{K}$. These processes are carried out in parallel for a mini-batch of inputs, and the formal definition of the categorical dictionary $\mathbf{K}$ is presented in Definition. 1.

Definition 1 *A Categorical Dictionary $\mathbf{K}$ with C-category is defined by:*

$$\mathbf{K} = \{k^1, k^2, ..., k^C\}, \tag{3}$$

where the categorical key $k^c \in \mathbf{K}$ is defined as the key k that belongs to the c-th semantic category ($c = \arg\max_i \hat{y}_k^{(i)}$) and the predicted category label $\hat{y}_k$ of $k = f_k(x_k)$ is derived by:

$$\arg\max_{\hat{y}_k} \sum_{c=1}^{C} \hat{y}_k^{(c)} \log p(c; k, \theta_h), \ s.t. \ \hat{y}_k \in \Delta^C, \forall k, \tag{4}$$

where h is the category classifier that predicts C-category probabilities for each embedding (e.g., k), and $\hat{y} = (\hat{y}^{(1)}, \hat{y}^{(2)}, ..., \hat{y}^{(C)})$ is the predicted category label. The key x_k is sampled from a training dataset X and encoded by the momentum encoder f_k to get the encoded key $k = f_k(x_k)$. Δ^C denotes a probability simplex, with which a point can be represented by C non-negative numbers that add up to 1.

Remark 1 *It is worth highlighting that Eq. 3 only shows one group of categorical keys for the simplicity of illustration and theoretic proof. In practice, we take the same strategy as [19] and maintain a dynamic categorical dictionary with M-size queue (i.e., $\{k_m^c\}_{1\leq c\leq C, 1\leq m\leq M}$), where the categorical keys are progressively updated in a category-wise manner. Specifically, for the queue of each category, we have $\{k_1^c, k_2^c, ..., k_M^c\}$, in which the oldest key is dequeued and the currently sampled key (belongs to c-th semantic category) is enqueued.*

Category contrastive loss. Given the categorical dictionary $\mathbf{K} = \{k_m^c\}_{1\leq c\leq C, 1\leq m\leq M}$ defined in Definition. 1, the proposed CaCo performs contrastive learning on unlabeled target data X_t via a category contrastive loss CatNCE that is defined by:

$$\mathcal{L}_{\text{CatNCE}} = \sum_{x_q \in X_t} -\Bigg(\frac{1}{M}\sum_{m=1}^{M} \log \frac{\sum_{c=1}^{C} \exp(q \cdot k_m^c / \tau_m^c)(\hat{y}_q \times \hat{y}_{k_m^c})}{\sum_{c=1}^{C} \exp(q \cdot k_m^c / \tau_m^c)}\Bigg), \tag{5}$$

where $q = f_q(x_q)$, $(\hat{y}_q \times \hat{y}_{k_m^c})$ is equal to 1 if both refer to the same category and 0 otherwise, τ_m^c is a temperature hyper-parameter and the $\cdot$ denotes the inner (dot) product. For each group of categorical keys $\{k_m^1, k_m^2, ..., k_m^C\}$, only one key is positive for the current query q ($i.e.$, $(\hat{y}_q \times \hat{y}_{k_m^c}) = 1$) as every sample belongs to a single category. This loss is thus the log loss of a C-way softmax-based classifier that strives to classify q as the positive key (of same category).

Remark 2 *Note that the CatNCE loss in Eq.5 has a similar form as the InfoNCE loss in Eq.2. Therefore, InfoNCE can be interpreted as a special case of CatNCE, where each instance (with its augmentations) itself is a category and the temperature is fixed ($i.e.$, $\tau_m^c = \tau, \forall c, m$). For CaCo, we assign different temperatures to different keys as their predicted labels have different uncertainties, $i.e.$, scaled by the prediction entropy $\mathcal{H}(\cdot)$. The adjustable temperature parameter has also been explored in [5, 17, 31].*

Remark 3 *Note that our category contrastive loss serves as an unsupervised objective function for training the encoder networks that represent the queries and keys [18]. In general, the query representation is $q = f_q(x^q)$ where f_q is an encoder network and x^q is a query sample (likewise, $k = f_k(x^k)$). Their instantiations depend on the specific pretext task. The input x^q and x^k can be images [18, 68, 73], patches [42] or context consisting of a set of patches [42], etc. The networks f_q and f_k can be identical [18, 66, 73], partially shared [2, 42], or different [19, 58].*

Relations to existing instance contrast methods. Beyond instance-discriminative representations as learnt by instance contrast [5, 19, 42, 58, 68, 73], CaCo learns category-discriminative yet domain-invariant representation.

3.4. Theoretical Insights

The category contrast (CaCo) is inherently connected with some probabilistic models. Specifically, CaCo can be modeled as an example of Expectation Maximization (EM):

Proposition 1 *The category contrastive learning can be modeled as a maximum likelihood (ML) problem optimized via Expectation Maximization (EM).*

Proposition 2 *The categorical contrastive learning is convergent under certain conditions.*

The proofs of **Propositions 1** and **2** are provided in the Appendix.

4. Experiments

This section presents experimental results. Sections 4.1 and 4.2 describe the dataset and implementation details. Sections 4.3, 4.4 and 4.5 present the UDA experiments in segmentation, detection and classification, respectively. Section 4.6 discusses different features of the proposed method.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg.	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Baseline [4]	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CaCo-S	91.1	54.4	79.6	27.0	22.9	36.9	40.2	33.4	83.7	36.3	65.2	59.7	22.4	83.5	37.5	49.3	10.1	23.3	31.8	46.8
CaCo-T	92.0	53.5	81.6	28.9	26.3	36.5	42.7	36.3	81.8	37.2	75.5	59.8	26.5	84.9	40.0	44.9	11.6	27.0	29.9	48.3
CaCo	91.9	54.3	82.7	31.7	25.0	38.1	46.7	39.2	82.6	39.7	76.2	63.5	23.6	85.1	38.6	47.8	10.3	23.4	35.1	49.2
AdaptSeg [59]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CBST [81]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
CLAN [40]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
AdvEnt [63]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
IDA [43]	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
BDL [36]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
CrCDA [29]	92.4	55.3	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
SIM [67]	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
TIR [32]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
CRST [80]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
+CaCo	93.0	58.4	83.1	34.0	29.3	37.0	47.1	42.9	84.6	41.5	82.8	61.8	32.2	86.9	39.2	48.0	22.4	31.1	45.7	52.7
FDA [72]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
+CaCo	93.2	54.5	84.6	32.9	29.3	39.7	46.9	42.7	84.4	40.1	83.7	61.1	32.2	85.6	41.7	51.2	19.2	35.6	45.9	52.9
ProDA [76]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
+CaCo	93.8	64.1	85.7	43.7	42.2	46.1	50.1	54.0	88.7	47.0	86.5	68.1	2.9	88.0	43.4	60.1	31.5	46.1	60.9	58.0

Table 1. Results over unsupervised domain adaptive semantic segmentation task GTA5-to-Cityscapes: CaCo-S, CaCo-T and CaCo construct the category-aware dictionary by sampling key samples x_k from the source dataset X_s only, the target dataset X_t only, and both datasets, respectively.

Method	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg.	Sky	PR	Rider	Car	Bus	Motor	Bike	mIoU	mIoU*
Baseline [4]	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
PatAlign [60]	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
AdaptSeg [59]	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
CLAN [40]	81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
AdvEnt [63]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
IDA [43]	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
CrCDA [29]	86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	42.9	50.0
TIR [32]	92.6	53.2	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3
SIM [67]	83.0	44.0	80.3	-	-	-	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	-	52.1
BDL [36]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
CRST [80]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
+CaCo	88.8	48.0	79.5	6.9	0.3	36.9	28.0	22.1	83.5	84.1	63.9	31.0	85.8	38.1	29.4	49.1	48.5	56.2
FDA [72]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.W1	83.9	40.8	38.4	51.1	-	52.5
+CaCo	86.4	43.3	78.7	9.0	0.1	28.5	26.7	29.7	81.7	82.9	59.3	28.1	82.9	38.6	35.7	50.0	47.6	55.7
CaCo	87.4	48.9	79.6	8.8	0.2	30.1	17.4	28.3	79.9	81.2	56.3	24.2	78.6	39.2	28.1	48.3	46.0	53.6

Table 2. Results of unsupervised domain adaptive semantic segmentation task SYNTHIA-to-Cityscapes.

4.1. Datasets

Adaptation for semantic segmentation: It involves three public datasets over two challenging UDA tasks, $i.e.$, GTA5 [48]-to-Cityscapes [9] and SYNTHIA [49]-to-Cityscapes. Specifically, GTA5 is a synthesized dataset with $24,966$ images and 19 common categories with Cityscapes. SYNTHIA is a synthesized dataset with $9,400$ images and 16 common categories with Cityscapes. Cityscapes is a real-image dataset with 2975 training samples and 500 validation samples.

Adaptation for object detection: It involves three public datasets over two adaptation tasks, $i.e.$, Cityscapes-to-Foggy Cityscapes [54] and Cityscapes-to-BDD [?]. Specifically, Foggy Cityscapes is a synthesized dataset that applies simulated fog on Cityscapes images. BDD is a real dataset with $70k$ samples in training set, $10k$ samples for validating and 7 common classes with Cityscapes dataset. As in [7, 52, 71], only a subset of BDD "*daytime* set" is used for experiments.

Adaptation for classification tasks: It involves two domain adaptive classification datasets VisDA17 [45] and Office-31 [50]. The former consists of a source domain with $152,409$ synthesized samples with twelve classes and a target domain with $55,400$ real samples. The latter consists of images of 31 categories which were collected from Amazon (2817 images), Webcam (795 images) and DSLR (498 images), respectively. The evaluation is on every pair of them as in [50, 55, 80].

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Baseline [47]	24.4	30.5	32.6	10.8	25.4	9.1	15.2	28.3	22.0
MAF [21]	28.4	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SCDA [79]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
DA [7]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
MLDA [70]	33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
DMA [33]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
CAFA [23]	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
SWDA [52]	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
+CaCo	39.3	46.1	48.0	32.4	45.7	38.7	31.3	35.3	39.6
CRDA [71]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
+CaCo	39.4	47.4	47.9	32.5	46.4	39.9	32.7	35.4	40.2
CaCo	38.3	46.7	48.1	33.2	45.9	37.6	31.0	33.0	39.2

Table 3. Results over unsupervised domain adaptive object detection task Cityscapes-to-Foggy-Cityscapes.

Method	person	rider	car	truck	bus	mcycle	bicycle	mAP
Baseline [47]	26.9	22.1	44.7	17.4	16.7	17.1	18.8	23.4
DA [7]	29.4	26.5	44.6	14.3	16.8	15.8	20.6	24.0
SWDA [52]	30.2	29.5	45.7	15.2	18.4	17.1	21.2	25.3
+CaCo	32.1	32.9	51.6	20.5	23.7	20.1	25.6	29.5
CRDA [71]	31.4	31.3	46.3	19.5	18.9	17.3	23.8	26.9
+CaCo	32.5	34.1	51.1	21.6	25.1	20.5	26.5	30.2
CaCo	32.7	32.2	50.6	20.2	23.5	19.4	25.0	29.1

Table 4. Results over unsupervised domain adaptive object detection task Cityscapes-to-BDD.

4.2. Experiment Detail

Segmentation Task: As in [59, 81], DeepLabV2 [4] is employed as segmentation architecture and ResNet-101 [20] is adopted as the backbone. We employ SGD [3] as the optimizer with momentum 0.9, weight decay 0.0001 and learning rate 0.00025. We follow previous works [59,81] to schedule the learning rate [4].

Detection Task: We follow previous works [7,34,52,71] to conduct experiments, where VGG16-based [57] Faster R-CNN [47] is employed base detecting backbone. For network optimization, Stochastic gradient descent optimizer [3] is adopted with a momentum of 0.9 and a weight decay of 0.0005. The shorter side of input image is set to 600 and RoIAlign is employed for feature extraction. The learning rate is set as 0.001 for $50,000$ training iterations and adjusted as 0.0001 in following $20,000$ training iterations [7,52,71].

Classification Task: Following [50,55,80], we employ ResNet101 (for VisDA17 dataset) and ResNet50 [20] (for Office-31 dataset) as the base backbones. For optimization, Stochastic gradient descent optimizer [3] is employed with momentum 0.9, weight decay 0.0005, learning rate 0.001 and batch size 32.

We set the length of dictionary queue M at 100 in all experiments except in parameter analysis. In addition, we set the momentum update coefficient b at 0.999 and the basic temperature τ at 0.07 as in [19].

4.3. UDA for Semantic Segmentation

Table 1 reports semantic segmentation results on the task GTA5-to-Cityscapes. It can be seen that the proposed CaCo achieves comparable performance with state-of-the-art methods. In addition, CaCo is complementary to existing UDA approaches that exploit adversarial loss, image translation loss and self-training loss. As shown in Table 1, incorporating CaCo as denoted by "+CaCo" boosts the performance of state-of-the-art methods clearly and consistently. Fig. 2 presents the qualitative comparisons.

Ablation studies. We perform ablation studies over a widely adopted *Baseline* [20] as shown on the top of Table 1, where *CaCo-S*, *CaCo-T* and *CaCo* mean that the category-aware dictionary is built with keys from source domain, target domain and both domains, respectively. We can observe that *CaCo-S* and *CaCo-T* outperform the *Baseline* clearly. *CaCo-S* and *CaCo-T* provide orthogonal self-supervision signals, where *CaCo-S* focuses on inter-domain category contrastive learning between target samples and source keys and *CaCo-T* focuses on intra-domain category contrastive learning between target samples and target keys. In addition, *CaCo* performs clearly the best, showing that the keys from the source and target domains are complementary.

Table 2 reports semantic segmentation results on the

Method	Aero	Bike	Bus	Car	Horse	Knife	Motor	Person	Plant	Skateboard	Train	Truck	Mean
Baseline [20]	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
MMD [37]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN [11]	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
ENT [13]	80.3	75.5	75.8	48.3	77.9	27.3	69.7	40.2	46.5	46.6	79.3	16.0	57.0
MCD [53]	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
ADR [51]	87.8	79.5	83.7	65.3	92.3	61.8	88.9	73.2	87.8	60.0	85.5	32.3	74.8
SimNet-Res152 [46]	94.3	82.3	73.5	47.2	87.9	49.2	75.1	79.7	85.3	68.5	81.1	50.3	72.9
GTA-Res152 [55]	-	-	-	-	-	-	-	-	-	-	-	-	77.1
CBST [81]	87.2	78.8	56.5	55.4	85.1	79.2	83.8	77.7	82.8	88.8	69.0	72.0	76.4
+CaCo	90.7	80.8	79.4	57.0	89.2	88.6	82.4	79.0	87.9	87.9	87.0	65.9	81.3
CRST [80]	88.0	79.2	61.0	60.0	87.5	81.4	86.3	78.8	85.6	86.6	73.9	68.8	78.1
+CaCo	91.4	80.6	80.0	56.5	89.5	89.4	82.8	79.9	88.8	86.8	87.3	66.0	81.6
CaCo	90.4	80.7	78.8	57.0	88.9	87.0	81.3	79.4	88.7	88.1	86.8	63.9	80.9

Table 5. Results over UDA-based classification benchmark VisDA17.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Mean
Baseline [20]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DAN [37]	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RTN [38]	84.5	96.8	99.4	77.5	66.2	64.8	81.6
DANN [11]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
ADDA [62]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
JAN [39]	85.4	97.4	99.8	84.7	68.6	70.0	84.3
GTA [55]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
CBST [81]	87.8	98.5	100.0	86.5	71.2	70.9	85.8
+CaCo	90.3	98.6	100.0	92.4	73.2	72.8	87.9
CRST [80]	89.4	98.9	100.0	88.7	72.6	70.9	86.8
+CaCo	90.4	98.9	100.0	92.8	73.7	72.5	88.1
CaCo	89.7	98.4	100.0	91.7	73.1	72.8	87.6

Table 6. Results over domain adaptive image classification task Office-31.

task SYNTHIA-to-Cityscapes. It can be observed that CaCo achieves comparable performance with the state-of-the-art methods, and it boosts their performance (denoted by "+CaCo") as well.

4.4. UDA for Object Detection

Tables 3 and 4 report object detection experiments on Cityscapes-to-Foggy Cityscapes and Cityscapes-to-BDD. They show that CaCo outperforms the state-of-the-art methods [52, 71] by clear margins. Besides, combining CaCo with state-of-the-art could boost the detection performance.

4.5. UDA for Image Classification

Tables 5 and 6 report image classification experiments on VisDA17 and Office-31, respectively. It can be observed that CaCo outperforms state-of-the-art by clear margins. Additionally, combining CaCo with state-of-the-art approaches could boost image classification performance.

4.6. Discussion and Analysis

Generalization ability: We investigate the generalization of the proposed CaCo via assessing it on several cornerstone visual UDA applications, $i.e.$, *segmentation*, *detection* and *classification*. We present the experimental results in Tables 1- 6, which demonstrate CaCo generates comparable performance consistently.

Complementariness ability: We investigate the synergetic benefits of our CaCo by combining it with existing UDA approaches. We present experimental results in Tables 1- 6 (the rows with '+CaCo'), which show that CaCo when incorporated improves all existing methods consistently across different visual tasks.

Comparisons with existing unsupervised representation learning methods: We compared CaCo with unsupervised representation learning methods over the UDA task. Most existing methods achieve unsupervised representation learning through certain pretext tasks, such as instance contrastive learning [2, 5, 6, 18, 19, 22, 42, 68, 73], patch ordering [41], rotation prediction [12], and denoising/context/colorization auto-encoders [44, 77, 78]. The experiments (shown in Appendix) over the UDA task GTA→Cityscapes show that existing unsupervised representation learning works not well on UDA tasks. The main reason lies in that these approaches are designed for learning instance-discriminative representations without considering semantic priors and domain gaps. CaCo also performs unsupervised learning but works for UDA effectively, largely because it learns category-discriminative yet domain-invariant representations which is essential to various visual UDA tasks.

Parameter studies: The parameter M (in the proposed CaCo) controls the length (or size) of the categorical dictionary. We investigate M via varying it from 50 to 150 progressively. The experiments (shown in Appendix) over the

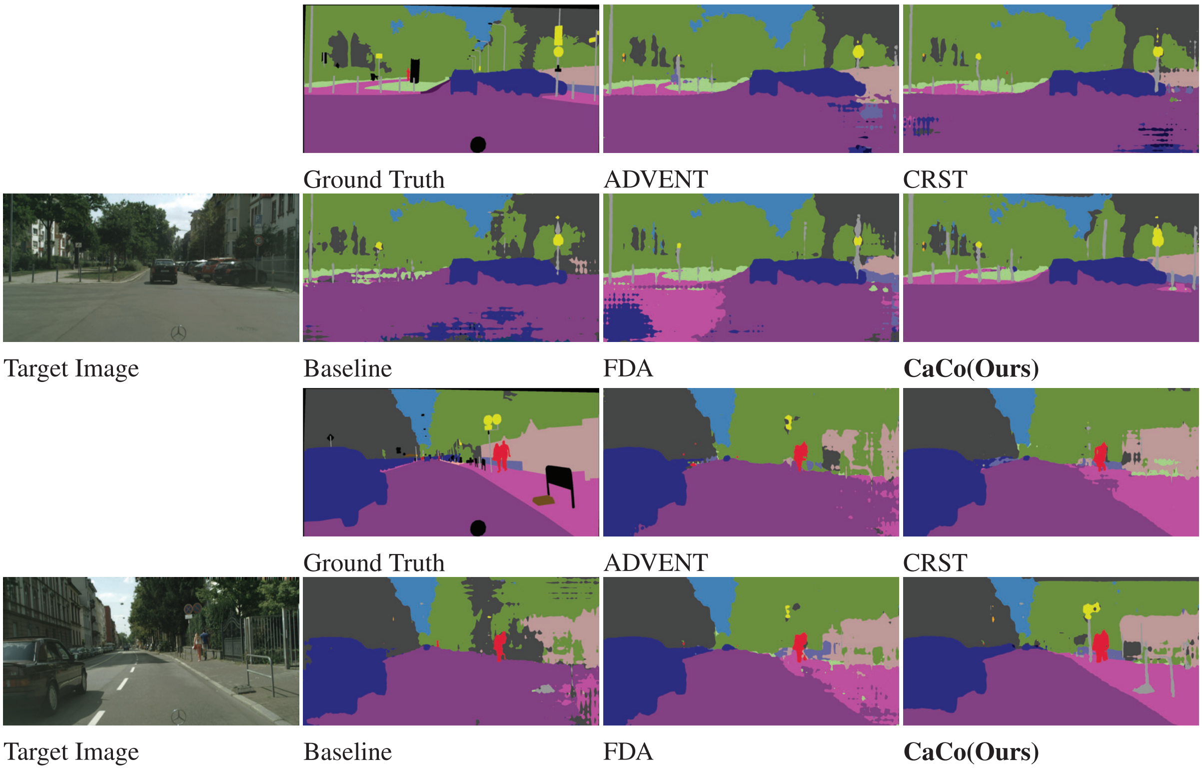

Figure 2. Qualitative comparisons over UDA-based semantic segmentation GTA5-to-Cityscapes.

UDA segmentation task GTA-to-Cityscapes demonstrate that M does not affect UDA clearly while it changes from 50 to 150.

Generalization across different learning setups: We studied the scalability of the proposed CaCo from the view of learning setups. Specifically, we evaluated CaCo over a variety of tasks that involve unlabeled data learning and certain semantic priors such as *unsupervised model adaptation*, and *partial-set/open-set UDA*. We present the experimental results in Appendix, which illustrates that CaCo generates comparable performance robustly.

Category-aware dictionary: We studied three variant designs of the proposed category-aware dictionary: 1) Assign all keys with the same temperature; 2) Using two individual dictionaries (for source and target data) instead of a single domain-mixed dictionary; 3) Update the dictionary by memory bank [68] or current mini-batch [5]. Experiments (in Appendix) verify the superiority of the design as described in this paper.

5. Conclusion

This paper presents CaCo, a category contrast technique that introduces a generic category contrastive loss that can work for various visual UDA tasks effectively. We construct a semantics-aware dictionary with samples from both source and target domains where each target sample is assigned a (pseudo) category label based on the category priors of source samples. This allows category contrastive learning (between target queries and the category-level dictionary) for category-discriminative yet domain-invariant feature representations: samples of the same category (from either source or target domain) are pulled close together while those of different categories are pushed away simultaneously. Extensive experiments over multiple visual tasks ($e.g.$, segmentation, classification and detection) show that the simple implementation of CaCo achieves superior performance as compared with state-of-the-art methods. In addition, we demonstrate that CaCo is also complementary to existing UDA methods and generalizable to other learning setups such as unsupervised model adaptation, open-/partial-set adaptation etc.

Acknowledgement

This research was conducted at Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is supported by A*STAR under its Industry Alignment Fund (LOA Award number: I1701E0013).

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 3

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. 4, 7

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 6

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 5, 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 4, 7, 8

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 7

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1, 2, 5, 6

[8] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019. 1, 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[10] Kaiwen Cui, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Genco: Generative co-training for generative adversarial networks with limited data. *arXiv preprint arXiv:2110.01254*, 2021. 2

[11] 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. 7

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 7

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 7

[14] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021. 2

[15] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8053–8064, 2021. 1

[16] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 2021. 2

[17] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020. 4

[18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3, 4, 7

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 3, 4, 6, 7

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6, 7

[21] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 6

[22] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 7

[23] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 6

[24] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10133–10144, 2021. 2

[25] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[26] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[27] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8988–8999, 2021. 1

[28] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Multi-level adversarial network for domain adaptive semantic segmentation. *Pattern Recognition*, 123:108384, 2022. 2

[29] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European Conference on Computer Vision*, pages 705–722. Springer, 2020. 2, 5

[30] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 1

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 4

[32] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. *arXiv preprint arXiv:2003.00867*, 2020. 5

[33] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 6

[34] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 6

[35] Shuang Li, Binhui Xie, Bin Zang, Chi Harold Liu, Xinjing Cheng, Ruigang Yang, and Guoren Wang. Semantic distribution-aware contrastive adaptation for semantic segmentation. *arXiv preprint arXiv:2105.05013*, 2021. 3

[36] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 1, 2, 5

[37] 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, 2015. 7

[38] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 1, 2, 7

[39] 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. 7

[40] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 1, 2, 5

[41] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 1, 7

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 3, 4, 7

[43] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. *arXiv preprint arXiv:2004.07703*, 2020. 5

[44] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 7

[45] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018. 5

[46] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018. 7

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 6

[48] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 5

[49] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 5

[50] 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. 5, 6

[51] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *International Conference on Learning Representations*, 2017. 2, 7

[52] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 2, 5, 6, 7

[53] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 1, 2, 7

[54] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 5

[55] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 5, 6, 7

[56] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019. 3

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 6

[58] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 2, 3, 4

[59] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1, 2, 5, 6

[60] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1456–1465, 2019. 1, 2, 5

[61] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019. 3

[62] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 1, 2, 7

[63] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 1, 2, 5

[64] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 2

[65] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 3

[66] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 4

[67] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerior Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. *arXiv preprint arXiv:2003.08040*, 2020. 5

[68] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 1, 2, 3, 4, 7, 8

[69] Aoran Xiao, Jiaxing Huang, Dayan Guan, and Shijian Lu. Unsupervised representation learning for point clouds: A survey. *arXiv preprint arXiv:2202.13589*, 2022. 1

[70] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 6

[71] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 5, 6, 7

[72] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 1, 2, 5

[73] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019. 1, 2, 3, 4, 7

[74] Jingyi Zhang, Jiaxing Huang, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 0–0, 2022. 1

[75] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, and Shijian Lu. Da-detr: Domain adaptive detection transformer by hybrid attention. *arXiv preprint arXiv:2103.17084*, 2021. 2

[76] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021. 5

[77] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1, 7

[78] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, pages 1058–1067, 2017. 1, 7

[79] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 6

[80] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 1, 2, 5, 6, 7

[81] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 1, 2, 5, 6, 7