Category Contrast for Unsupervised Domain Adaptation in Visual Tasks
Appendix

A. Theoretical insights of CaCo

A.1. Proof of Proposition 1

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

Proof:

Maximum likelihood (ML) was initially proposed to model clustering tasks, and can be optimized by expectation maximization (EM). For our proposed category contrastive learning, the objective is to find the encoder weights θ_{f_q} that maximizes the log-likelihood function of both labeled data X_s and unlabeled data X_t:

$$\theta^*_{f_q} = \arg\max_{\theta_{f_q}} \sum_{x_s \in X_s} \log p(x_s; \theta_{f_q}) + \sum_{x_t \in X_t} \log p(x_t; \theta_{f_q}). \tag{1}$$

As the *labeled data* are with annotations, the first term of the right-hand side (RHS) in Eq. 1 can be maximized by the supervised learning that minimizes a cross-entropy loss between the predictions of X_s and their annotations Y_s:

$$\arg\min_{\theta_G} \mathcal{L}_{sup} = \arg\min_{\theta_{f_q}, \theta_h} \sum_{x_s \in X_s, y_s \in Y_s} -y_s \log(h(f_q(x_s)))), \tag{2}$$

where h is the category classifier and the combination of h and f_q forms the visual task model $G = h(f_q(\cdot))$.

Please refer to [8,12,34] for the detailed proofs that minimizing the cross-entropy loss leads to the likelihood maximization.

As the *unlabeled data* are without annotations, CaCo maximizes the second term by the proposed category contrastive learning. Below please find the proof.

We assume that the unlabeled samples X_t are related to latent variable $\{k_c\}_{c=1}^{C}$ which denotes the categorical keys of the data. C stands for the number of categories. In this way, we can re-write the second term of the RHS in Eq. 1 as follows:

$$\theta^*_{f_q} = \arg\max_{\theta_{f_q}} \sum_{x_t \in X_t} \log \sum_{c=1}^{C} p(x_t, k_c; \theta_{f_q}) \tag{3}$$

As it is difficult to optimize Eq.3 directly, we utilize a surrogate function to lower-bound the log-likelihood function:

$$\begin{aligned} \sum_{x_t \in X_t} \log \sum_{c=1}^{C} p(x_t, k_c; \theta_{f_q}) &= \sum_{x_t \in X_t} \log \sum_{c=1}^{C} \mathcal{D}(k_c) \frac{p(x_t, k_c; \theta_{f_q})}{\mathcal{D}(k_c)} \\ &\geq \sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}(k_c) \log \frac{p(x_t, k_c; \theta_{f_q})}{\mathcal{D}(k_c)}, \end{aligned} \tag{4}$$

where $\mathcal{D}(k_c)$ denotes some distribution over k's ($\sum_{c=1}^{C} \mathcal{D}(k_c) = 1$), and the last step of derivation utilizes Jensen's inequality [10,19,26]. This equality holds if $\frac{p(x_t, k_c; \theta_{f_q})}{\mathcal{D}(k_c)} =$ Constant. Thus, we can get:

$$\mathcal{D}(k_c) = \frac{p(x_t, k_c; \theta_{f_q})}{\sum_{c=1}^{C} p(x_t, k_c; \theta_{f_q})} = \frac{p(x_t, k_c; \theta_{f_q})}{p(x_t; \theta_{f_q})} = p(k_c; x_t, \theta_{f_q}) \tag{5}$$

By ignoring the constant$-\sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}(k_c) \log \mathcal{D}(k_c)$ in Eq.4, we are supposed to maximize:

$$\sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}(k_c) \log p(x_t, k_c; \theta_{f_q}) \tag{6}$$

Expectation step. We estimate the posterior probability $p(k_c; x_t, \theta_{f_q})$. For this purpose, we conduct C-category pseudo labeling on the key embeddings $k = f_k(x_k)$ ($x_k \in X_s \cup X_t$) that are encoded by the momentum encoder to obtain category-level keys $\{k_c\}_{c=1}^{C}$. The categorical key k_c 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:

$$\arg\max_{\hat{y}_k} \sum_{c=1}^{C} \hat{y}_k^{(c)} \log p(c; k, \theta_h), \; s.t. \; \hat{y} \in \Delta^C, \forall k \tag{7}$$

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. To get the pseudo label $\hat{y}_q$ of the query embedding $q = f_q(x_t)$ ($x_t \in X_t$) encoded by current encoder, we simply repeat above steps by replacing all the notation "k" with "q".

Next, we calculate $p(k_c; x_t, \theta_{f_q}) = \hat{y}_q \times \hat{y}_{k_c}$, where $\hat{y}_q \times \hat{y}_{k_c} = 1$ if both refer to the same category; otherwise, $\hat{y}_q \times \hat{y}_{k_c} = 0$.

Maximization step. Now, we are ready to maximize the lower-bound in Eq.6.

$$\begin{aligned}\sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}(k_c) \log p(x_t, k_c; \theta_{f_q}) &= \sum_{x_t \in X_t} \sum_{c=1}^{C} p(k_c; x_t, \theta_{f_q}) \log p(x_t, k_c; \theta_{f_q}) \\ &= \sum_{x_t \in X_t} \sum_{c=1}^{C} (\hat{y}_q \times \hat{y}_{k_c}) \log p(x_t, k_c; \theta_{f_q})\end{aligned} \tag{8}$$

We assume a uniform prior over categorical keys. Then, we get:

$$p(x_t, k_c; \theta_{f_q}) = p(x_t; k_c, \theta_{f_q}) p(k_c; \theta_{f_q}) = \frac{1}{C} \cdot p(x_t; k_c, \theta_{f_q}), \tag{9}$$

where we let the prior probability $p(k_c; \theta_{f_q})$ for each k_c as $1/C$ as no data is provided.

Under the assumption that the embedding distribution around each categorical key k_c is an isotropic Gaussian [3], we get:

$$p(x_t; k_c, \theta_{f_q}) = \exp\left(\frac{-(q - k_+)^2}{2\sigma_+^2}\right) \Big/ \sum_{c=1}^{C} \exp\left(\frac{-(q - k_c)^2}{2\sigma_c^2}\right), \tag{10}$$

where $q = f_q(x_t)$, and k_+ is defined as the key k_c that belongs to the same category as q (*i.e.*, $\hat{y}_q \times \hat{y}_{k_+} = 1$). By applying ℓ_2-normalization to q and k, we get $(q - k)^2 = 2 - 2q \cdot k$. Combining this equation with Eqs.3, 4, 6, 8, 9, 10, we formulate the likelihood maximization as:

$$\theta_{f_q}^* = \arg\min_{\theta_{f_q}} \sum_{x_t \in X_t} -\log \frac{\exp(q \cdot k_+ / \tau_+)}{\sum_{c=1}^{C} \exp(q \cdot k_c / \tau_c)}, \tag{11}$$

where $\tau \propto \sigma^2$ represents the density level of the embedding distribution around a categorical key (*e.g.*, k_c).

In practice, Eq. 11 can be achieved by minimizing a category contrastive loss:

$$\arg\min_{\theta_G} \mathcal{L}_{\text{CatNCE}} = \arg\min_{\theta_{f_q}} \sum_{x_t \in X_t} -\left(\frac{1}{M} \sum_{m=1}^{M} \log \frac{\sum_{c=1}^{C} \exp(q \cdot k_m^c / \tau_m^c)(\hat{y}_q \times \hat{y}_{k_m^c})}{\sum_{c=1}^{C} \exp(q \cdot k_m^c / \tau_m^c)}\right). \tag{12}$$

Please note that Eq. 12 is an instance of Eq. 11. They look different becuase: 1) Eq. 11 uses k_+ to denote the positive key instead of using a complex expression to identify the positive key (*i.e.*, $\sum_{c=1}^{C} \exp(q \cdot k_m^c / \tau_m^c)(\hat{y}_q \times \hat{y}_{k_m^c})$), for the simplicity of theoretic proof; 2) Eq. 11 only shows one group of categorical keys instead of M-group categorical keys, for the simplicity of theoretic proof.

A.2. Proof of Proposition 2

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

Proof:

For the supervised learning on *labeled data*, please refer to [8, 12, 34] for the detailed proofs of the fact that the likelihood maximization by minimizing the cross-entropy loss is convergent under certain conditions.

For the unsupervised learning on *unlabeled data*, please find the convergence proof below.

We let

$$
\begin{aligned}
L(\theta_{f_q}) = \sum_{x_t \in X_t} \log p(x_t; \theta_{f_q}) &= \sum_{x_t \in X_t} \log \sum_{c=1}^{C} p(x_t, k_c; \theta_{f_q}) \\
&= \sum_{x_t \in X_t} \log \sum_{c=1}^{C} \mathcal{D}(k_c) \frac{p(x_t, k_c; \theta_{f_q})}{\mathcal{D}(k_c)} \\
&\geq \sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}(k_c) \log \frac{p(x_t, k_c; \theta_{f_q})}{\mathcal{D}(k_c)}.
\end{aligned} \tag{13}
$$

It has been illustrated in Section A.1 that the inequality in Eq.13 holds with equality if $\mathcal{D}(k_c) = p(k_c; x_t, \theta_{f_q})$.

In the n-th Expectation-step, we estimate $\mathcal{D}^n(k_c) = p(k_c; x_t, \theta_{f_q}^n)$. Thus, we get:

$$
L(\theta_{f_q}^n) = \sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}^n(k_c) \log \frac{p(x_t, k_c; \theta_{f_q}^n)}{\mathcal{D}^n(k_c)}. \tag{14}
$$

In the n-th Maximization-step, we fix $\mathcal{D}^n(k_c) = p(k_c; x_t, \theta_{f_q}^n)$ and optimize weights θ_{f_q} to maximize Equation 14. Thus, we always get:

$$
\begin{aligned}
L(\theta_{f_q}^{n+1}) &\geq \sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}^n(k_c) \log \frac{p(x_t, k_c; \theta_{f_q}^{n+1})}{\mathcal{D}^n(k_c)} \\
&\geq \sum_{x_t \in X_t} \sum_{c=1}^{C} \mathcal{D}^n(k_c) \log \frac{p(x_t, k_c; \theta_{f_q}^{n})}{\mathcal{D}^n(k_c)} \\
&= L(\theta_{f_q}^n).
\end{aligned} \tag{15}
$$

Eq. 15 indicates that $L(\theta_{f_q}^n)$ monotonously increases along with more training iterations.

As the log-likelihood is upper-bounded, our proposed category contrastive learning will thus converge.

We may utilize gradient descent to achieve Eq. 15 by minimizing the category contrastive loss in Eq. 12. With a proper learning rate, this loss is guaranteed to decrease monotonically. In practice, network training is normally conducted with mini-batch gradient descent instead of gradient descent. This may not strictly guarantee the monotonic decrease of the loss, but will almost certainly converge to a lower one.

B. Discussion

B.1. Conceptual comparisons

We provided conceptual comparisons of different UDA methods in Table 1.

B.2. 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 [1,6,7,13,14,16,17,28,41,45], patch ordering [9,27], rotation prediction [11], and denoising/context/colorization auto-encoders [31,38,47,48]. The experiments (shown in Table 2) over the UDA task GTA→Cityscapes show that existing unsupervised representation learning does not perform well in the UDA task. The major reason is that these methods were designed to learn 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.

B.3. Parameter studies

The parameter M (in the proposed CaCo) controls the length (or size) of the categorical dictionary. We studied M by changing it from 50 to 150 with a step of 25. The experiments (shown in Table 3) over the UDA segmentation task GTA → Cityscapes show that M does not affect UDA clearly while it changes from 50 to 150.

Methods	Mec.	Cross-domain adaptation	Intra-domain adaptation	Category aware	Task generalizable	Setup generalizable	Main assumption
AdaptSeg [36]	AT	✓	×	×	×	×	Domain-invariant representations can be learnt via
CLAN [24]	AT	✓	×	✓	×	×	adversarial training (AT) in output space (AdaptSeg),
AdvEnt [39]	AT	✓	×	×	✓	×	entropy space (ADVENT), patch space (PatAlign), and context
PatAlign [37]	AT	✓	×	×	×	×	space (CrCDA). They can also be learnt through sample
IDA [29]	AT	×	✓	×	×	×	or class joint AT (CLAN and SIM), multi-level AT (SWDA),
CrCDA [18]	AT	✓	×	×	×	×	regularized AT (CRDA), or intra-domain AT (IDA).
SIM [40]	AT	✓	×	✓	×	×	
SWDA [32]	AT	✓	×	×	×	×	
CRDA [42]	AT	✓	×	✓	×	×	
TIR [21]	IT	✓	×	×	×	×	Domain-invariant representations can be learnt via
FDA [44]	IT	✓	×	×	×	×	image translation (TIR), spectrum swapping (FDA).
CBST [51]	ST	×	✓	✓	✓	×	Category-discriminative representation can be learnt
CRST [50]	ST	×	✓	✓	✓	×	via self-training (CBST), regularized ST (CRST).
CaCo (ours)	IC	✓	✓	✓	✓	✓	Category-discriminative yet domain-invariant representation can be learnt via instance contrast (IC).

Table 1. Conceptual comparisons of different UDA methods. Mec. denotes Mechanisms. AT, IT, ST, and IC denote adversarial training, image translation, self-training, and instance contrast, respectively.

Method	mIoU	gain
Baseline [15]	36.6	N.A.
Jigsaw [27]	38.5	+1.9
Rotation [11]	37.0	+0.4
Colorization [47]	38.7	+2.1
SimCLR [6]	38.4	+1.8
InstDisc [41]	38.0	+1.4
MoCo [14]	38.9	+2.3
CaCo	**49.2**	**+12.6**

Table 2. Comparisons with existing unsupervised representation learning methods: For the semantic segmentation over GTA $\rightarrow$ Cityscapes adaptation, CaCo performs the best consistently by large margins.

	M (the length of categorical dictionary)				
Method	50	75	100	125	150
CaCo	48.9	49.1	49.2	49.1	49.1

Table 3. The length of categorical dictionary (parameter M) affects unsupervised domain adaptation (evaluated over semantic segmentation on GTA $\rightarrow$ Cityscapes adaptation).

B.4. 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*, *partial-set domain adaptation* and *open-set domain adaptation*. Experiments (in Tables 4-6) show that CaCo achieve competitive performance consistently across all the tasks.

Unsupervised model adaptation	mIoU	gain
Baseline [15]	36.6	N.A.
UR [35]	45.1	+8.5
SFDA [23]	45.8	+9.2
CaCo	47.6	+11.0

Table 4. Comparison on unsupervised model adaptation (UMA) over GTA5 → Cityscapes adaptation: For semantic segmentation, CaCo achieves competitive performance as compared with state-of-the-art UMA methods. (Compared with UDA, UMA dose not use labeled source data during adaptation.)

Partial-set DA	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Mean
ResNet-50 [15]	46.3	67.5	75.9	59.1	59.9	62.7	58.2	41.8	74.9	67.4	48.2	74.2	61.3
IWAN [46]	53.9	54.5	78.1	61.3	48.0	63.3	54.2	52.0	81.3	76.5	56.8	82.9	63.6
SAN [4]	44.4	68.7	74.6	67.5	65.0	77.8	59.8	44.7	80.1	72.2	50.2	78.7	65.3
ETN [5]	59.2	77.0	79.5	62.9	65.7	75.0	68.3	55.4	84.4	75.7	57.7	84.5	70.5
SAFN [43]	58.9	76.3	81.4	70.4	73.0	77.8	72.4	55.3	80.4	75.8	60.4	79.9	71.8
CaCo	61.2	83.7	90.5	73.9	75.4	81.5	76.7	61.3	89.4	80.5	66.1	86.9	77.3

Table 5. Comparison on partial-set UDA (PS-UDA) over Office-Home: For image classification, CaCo achieves competitive performance as compared with state-of-the-art PS-UDA methods. (In PS-UDA setting, source and target domains do not share a completely same label space.)

Open-set DA	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Mean
ResNet [15]	36.3	54.8	69.1	33.8	44.4	49.2	36.8	29.2	56.8	51.4	35.1	62.3	46.6
ATI-λ [30]	55.2	52.6	53.5	69.1	63.5	74.1	61.7	64.5	70.7	79.2	72.9	75.8	66.1
OSBP [33]	56.7	51.5	49.2	67.5	65.5	74.0	62.5	64.8	69.3	80.6	74.7	71.5	65.7
OpenMax [2]	56.5	52.9	53.7	69.1	64.8	74.5	64.1	64.0	71.2	80.3	73.0	76.9	66.7
STA [22]	58.1	53.1	54.4	71.6	69.3	81.9	63.4	65.2	74.9	85.0	75.8	80.8	69.5
CaCo	63.5	78.7	83.8	61.1	74.0	79.6	64.2	58.2	82.3	68.8	62.9	81.7	71.6

Table 6. Comparison on open-set UDA (OS-UDA) over Office-Home: For image classification, CaCo achieves competitive performance as compared with state-of-the-art OS-UDA methods. (In OS-UDA setting, source and target domains do not share a completely same label space.)

B.5. Category-aware dictionary

What about assigning all keys with the same temperature? In Eq.5 in the submitted manuscript, we assigned different temperatures to different keys as their predicted labels have different uncertainties (labeled source samples also have corresponding prediction uncertainties even if they are labeled). In this section, we conduct experiments to show how this adaptive temperature design affects the performance. Table 7 shows that CaCo (with adaptive temperature) outperforms its uncertainty-independent version (with fixed temperature). The reason is that CoCo (with adaptive temperature) alleviates the negative effects from the wrongly pseudo-labeled keys, $i.e.$, suppressing the effect of keys with high uncertainty (about the category pseudo labeling). In another word, CaCo encourages to employ well-learnt embeddings (instead of using under-learnt embeddings) as keys in representation learning.

What about using two individual dictionaries (for source and target data) instead of a single domain-mixed dictionary? In the description of categorical dictionary in the submitted manuscript, the dictionary keys are domain-mixed, $i.e.$, evenly sampled from source and target domains. In this section, we conduct experiments to show how domain-mixing design affects the performance. Table 8 shows that CaCo (with domain-mixed dictionary) outperforms its vanilla version (with two individual dictionaries) clearly. The reason is that CaCo (with domain-mixed dictionary) enables information communication across domains (like in Shufflenet [49]) that helps mitigate inter-domain discrepancy. For instance, with a group

Fixed or Adaptive temperature	mIoU	gain
Baseline [15]	36.6	N.A.
CaCo (with fixed temperature)	47.9	+11.3
CaCo (with adaptive temperature)	49.2	+12.6

Table 7. Fixed or adaptive temperature in the proposed Category Contrast: CaCo with adaptive temperature clearly outperforms its uncertainty-independent version with fixed temperature, evaluated over UDA-based semantic segmentation task GTA $\rightarrow$ Cityscapes.

of domain-mixed keys that contains a "car" key encoded from target domain and $C-1$ keys (the rest categories) encoded from source domain, a target "car" query could be pulled closer to its positive target-domain key and pushed away from its negative source-domain keys, which makes the learning process more efficient and effective.

Domain-mixed or individual dictionary	mIoU	gain
Baseline [15]	36.6	N.A.
CaCo-S	46.8	+10.2
CaCo-T	48.3	+11.7
CaCo (with two individual dictionaries)	48.5	+11.9
CaCo (with domain-mixed dictionary)	49.2	+12.6

Table 8. Domain-mixed or individual dictionary: CaCo with domain-mixed dictionary clearly outperforms its vanilla version with two individual dictionaries, evaluated over the UDA-based semantic segmentation task GTA $\rightarrow$ Cityscapes.

What about sampling queries from source-domain ($i.e.$, training networks with an extra supervised source-domain contrastive loss)? As described in Fig. 1 and Eq.5 in the submitted manuscript, CaCo samples queries from target domain only in contrastive learning of its unlabelled samples. This strategy is intuitive and reasonable as the annotated source-domain data can be well learnt with supervised losses without bothering with an extra contrastive loss. Nevertheless, [20] shows that incorporating an extra supervised contrastive loss could further improve the supervised learning of source-domain data. We thus conduct new experiments to explore whether training networks with an extra source-domain contrastive loss [20] could further improve domain adaptation. As Table 9 shows, the baseline with a supervised source-domain contrast marginally outperforms its vanilla version, which indicates that the supervised contrastive learning could improve the generalization of the baseline model. However, incorporating the supervised source-domain contrast into CaCo slightly degrades the domain adaptation. We conjecture that further including an extra supervised source-domain contrastive loss may distract the network from learning target-domain data which leads to the slight degradation.

Sampling queries from source-domain?	mIoU	gain
Baseline [15]	36.6	N.A.
+supervised source-domain contrast	38.2	+1.6
CaCo	49.2	+12.6
+supervised source-domain contrast	49.0	+12.4

Table 9. Sampling queries from source-domain ($i.e.$, training networks with an extra supervised source-domain contrastive loss [20]) or not: The baseline with supervised source-domain contrast marginally outperforms its vanilla version, which indicates that the supervised contrastive learning could improve the generalization of the baseline model. However, incorporating supervised source-domain contrast into CaCo slightly degrades the domain adaptation. We conjecture that further including an extra source-domain contrastive loss may distract the networks from learning from target-domain samples which leads to the slight degradation. All experiments are conducted over the UDA-based semantic segmentation task GTA $\rightarrow$ Cityscapes.

What about updating the dictionary by memory bank [41] or current mini-batch [6]? In this paper, we use a momentum encoder to encode the keys and update the dictionary. In this section, we conduct experiments to show how dictionary-update strategy affects the performance. Table 10 shows that the performance is not sensitive to dictionary-update strategies. The reason is that CaCo improves domain adaptation largely by reducing domain gaps and enhancing category discrimination whereas dictionary-update strategy has little effect on these two factors.

Dictionary updating strategies	mIoU	gain
Baseline [15]	36.6	N.A.
CaCo (end-to-end updating [6])	49.0	+12.4
CaCo (memory bank updating [41])	48.9	+12.3
CaCo (momentum updating [14])	49.2	+12.6

Table 10. Dictionary updating strategies: Different dictionary updating strateties (i.e. end-to-end updating [6], memory bank updating [41], and momentum updating [14]) have little effect on CaCo's performance, evaluated over the UDA-based semantic segmentation task GTA $\rightarrow$ Cityscapes.

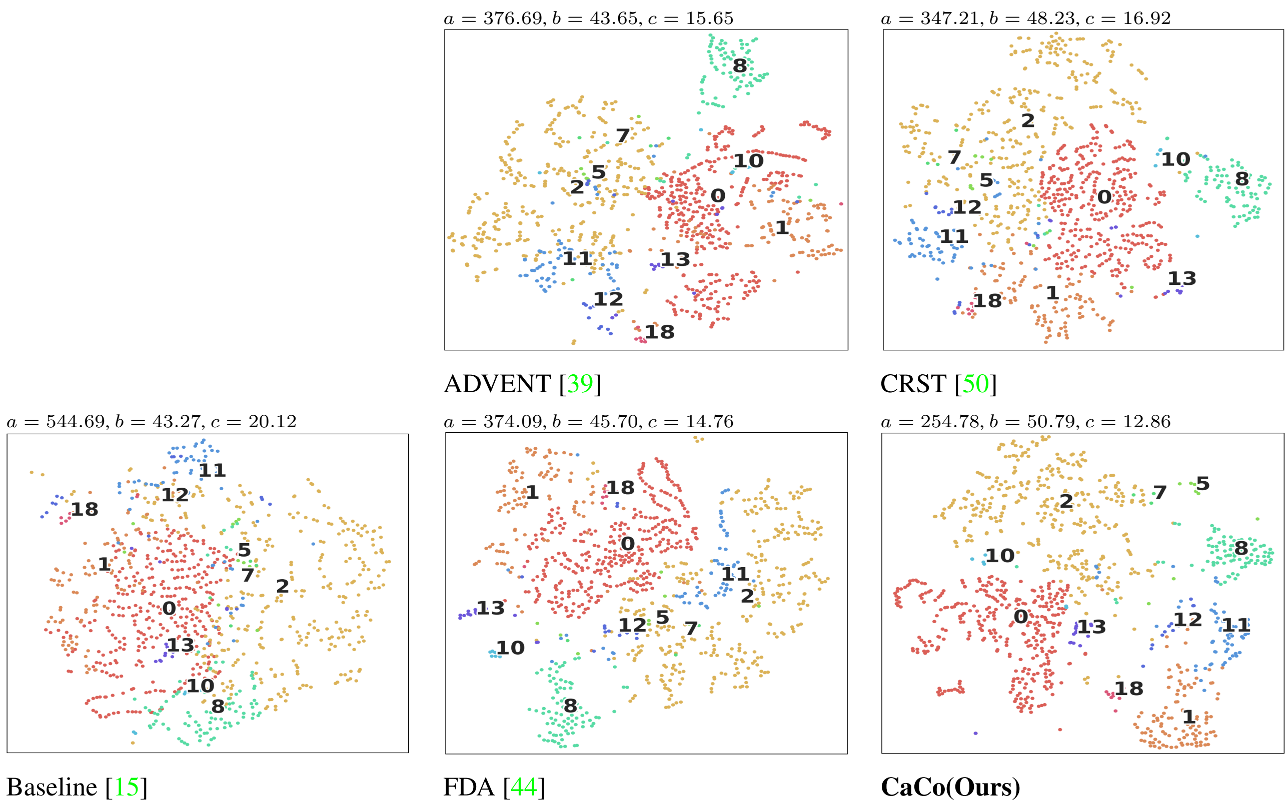

Figure 1. The t-SNE [25] visualization of feature distribution for target images on task GTA $\rightarrow$ Cityscapes: Each colour represents one semantic category of image pixels with a digit showing the category center. a, b and c on the top of each graph are intra-category variance, inter-category distance and cross-domain distance of the corresponding feature distribution. The proposed CaCo greatly outperforms "Baseline", "ADVENT" (adversarial training based), "CRST" (self-training based) and "FDA" (image translation based) qualitatively and quantitatively. Please note that we did not include the source feature distribution for simplicity and clarity.

B.6. Feature distribution analysis

Feature distribution visualization. We provide the t-SNE [25] visualization of feature distribution for target images on the GTA $\rightarrow$ Cityscapes task. To illustrate the unique features of the proposed "CaCo", we compare it with the "Baseline" and three typical UDA approaches, *i.e.*, "ADVENT" [39] (adversarial training based), "CRST" [50] (self-training based) and "FDA" [44] (image translation based) as shown in Fig.1. In addition, we evaluate the learnt features by using three metrics including intra-category variance, inter-category distance, and cross-domain distance which are labelled by a, b, and

c, respectively, as shown at the top of each graph in Fig.1.

It can be observed that the feature distribution of the Baseline model is messy and not category-discriminative ($i.e.$, $a = 544.69$ and $b = 43.27$) due to the large distribution gaps across domains ($i.e.$, $c = 20.12$). ADVENT generates features with less cross-domain discrepancy ($i.e.$, $c = 15.65$) as it employs adversarial training to reduce domain gaps. However, adversarial training is category-unaware, which leads to sub-optimal category discrimination ($i.e.$, $a = 376.69$ and $b = 43.65$) for visual recognition which requires category-discriminative features. FDA adopts image translation and generates features ($i.e.$, $a = 374.09$, $b = 45.70$ and $c = 14.76$) in a similar manner to ADVENT. In addition, CRST generates category-discriminative features ($i.e.$, $a = 347.21$ and $b = 48.23$) as it employs category-wise self-training. On the other hand, self-training is less effective on cross-domain gaps reduction, leading to sub-optimal cross-domain distance ($c = 16.92$) for domain adaptation. The proposed CaCo learns category-discriminative yet domain-invariant features ($i.e.$, $a = 254.78$, $b = 50.79$ and $c = 12.86$) as it employs a category-aware and domain-mixed dictionary for categorical contrastive learning. The visualization verifies the first and second claims as mentioned in the fifth paragraph of the Introduction.

Category imbalance mitigation. As shown in Fig.1, it can be also observed that the feature distribution generated by the proposed CaCo is more category-balanced. For example, the features of both dominant and less-dominant categories are well separated, $i.e.$, category $0 - 18$. On the contrary, the other four models ($i.e.$, "Baseline", "ADVENT", "FDA" and "CRST") generate feature distributions that are more category-imbalanced. For example, the features of dominant categories are well learnt and separated ($e.g.$, category 0-2), but the features of less-dominant categories are poorly learnt and separated ($e.g.$, category 5&7, 8&10 in "Baseline", category 2&5 in "ADVENT", category 5&7&12, 2&11 in "FDA" and category 5&11&12 in "CRST").

References

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

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 6

[3] Wlodzimierz Bryc. *The normal distribution: characterizations with applications*, volume 100. Springer Science & Business Media, 2012. 3

[4] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, 2018. 6

[5] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2985–2994, 2019. 6

[6] 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. 4, 5, 7, 8

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[8] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 2, 3

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 4

[10] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. 2

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 4, 5

[12] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 2, 3

[13] 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. 4

[14] 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. 4, 5, 8

[15] 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. 5, 6, 7, 8

[16] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 4

[17] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 4

[18] 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. 5

[19] Johan Ludwig William Valdemar Jensen et al. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30:175–193, 1906. 2

[20] 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. 7

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

[22] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2019. 6

[23] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. *arXiv preprint arXiv:2103.16372*, 2021. 6

[24] 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. 5

[25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[26] Tristan Needham. A visual explanation of jensen's inequality. *The American mathematical monthly*, 100(8):768–771, 1993. 2

[27] 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. 4, 5

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

[29] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[30] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. 6

[31] 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. 4

[32] 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. 5

[33] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018. 6

[34] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 2, 3

[35] Prabhu Teja Sivaprasad and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number CONF. IEEE, 2021. 6

[36] 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. 5

[37] 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. 5

[38] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 4

[39] 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. 5, 8

[40] 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

[41] 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. 4, 5, 7, 8

[42] 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

[43] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1426–1435, 2019. 6

[44] 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. 5, 8

[45] 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. 4

[46] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018. 6

[47] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 4, 5

[48] 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. 4

[49] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 6

[50] 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. 5, 8

[51] 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. 5