



Energy Transfer Contrast Network for Unsupervised Domain Adaption

Jiajun Ouyang, Qingxuan Lv, Shu Zhang^(✉), and Junyu Dong

Ocean University of China, Qingdao 266100, China
{ouyangjiajun,lvqingxuan}@stu.ouc.edu.cn,
{zhangshu,dongjunyu}@ouc.edu.cn

Abstract. The main goal of unsupervised domain adaptation is to improve the classification performance on unlabeled data in target domains. Many methods try to reduce the domain gap by treating multiple domains as one to enhance the generalization of a model. However, aligning domains as a whole does not account for instance-level alignment, which might lead to sub-optimal results. Currently, many researchers utilize meta-learning and instance segmentation approaches to tackle this problem. But it can only obtain a further optimized the domain-invariant feature learned by the model, rather than achieve instance-level alignment. In this paper, we interpret unsupervised domain adaptation from a new perspective, which exploits the energy difference between the source and target domains to reduce the performance drops caused by the domain gap. At the same time, we improve and exploit the contrastive learning loss, which can push the target domain away from the decision boundary. The experimental results on different benchmarks against a range of the state-of-the-art approaches justify the performance and the effectiveness of the proposed method.

1 Introduction

Convolutional Neural Network (CNN) is one of the representative algorithms of deep learning. However, CNNs often rely on a large amount of labeled training data in practical applications. Although we can provide rich labels for some fields with many categories, this leads to high time costs. To address this problem, Unsupervised Domain Adaptation (UDA) can transfer the knowledge learned from the labeled source domain to the unlabeled target domain, which has attracted a lot of attention from academia [5, 18] and industry [22].

Unsupervised domain adaptation has made impressive progress so far, and the vast majority of methods adjust the distribution of source and target domains by reducing the domain discrepancy, such as Maximum Mean Discrepancy (MMD) [1], joint maximum mean Discrepancy (JMMD) [2], etc. Another predominant streams in UDA based on the Generative Adversarial Networks [25] to maximize the error of the domain discriminator to confuse the source and target domains (Fig. 1).

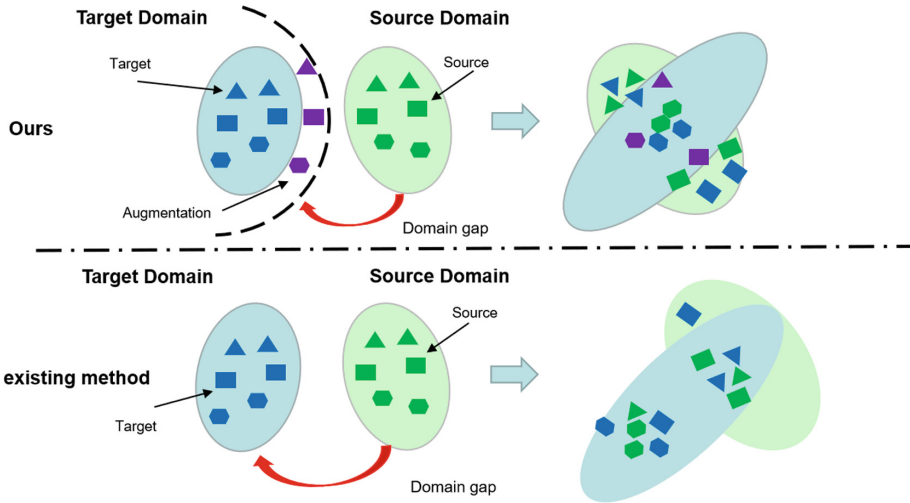


Fig. 1. We achieve instance-level alignment by contrasting the ability of learning to pull away positives samples and push away negatives.

However, directing alignment on the feature space may lead to the following problems: First, due to sampling variability, the label space of source and target domain samples on each mini-batch is different, which undoubtedly leads to outlier generation and negative optimization of generalization performance. Second, this direct approach to reducing the domain gap does not take into account instance-level alignment. Therefore, domain adaptation urgently need a solution that considers both distribution and categories discrepancy.

To tackle aforementioned problem, this paper propose an energy representation-based contrastive learning algorithm to avoid the first two problems: first, we improve contrastive learning and apply it to the UDA task for instance-level alignment. Secondly, we look at the UDA problem from another perspective, treating the target domain data as out-of-distribution data with the same labels as the source domain data. Due to the different data distributions in the target domain and the source domain, their energy values will be different to some extent [28, 32, 33]. So we use this difference to encourage the classifier to fit the energy value of the target domain to the vicinity of the source domain to mitigate the effects of domain shift. Since the energy is a non-probabilistic scalar value, it can be regarded as a certain norm of the output vector, which is less negatively affected by the label space in the mini-batch and reduce the domain gap can better avoid the negative optimization caused by Randomness of sampling.

We conduct experiments on several datasets to compare state-of-the-art methods, and the experimental results demonstrate the effectiveness of our method. Furthermore, we comprehensively investigate the impact of different components

of our approach, aiming to provide insights for the following research. The contribution of this article is summarized as follows:

1. We provide a new perspective that treats the target domain as out-of-distribution data with the same label space in the source domain, and achieves unsupervised domain adaptation by narrowing the difference between the OOD and ID data of the source and target domains
2. We improve the paradigm of contrastive learning, using contrastive learning to pull positive pairs closer and push negative pairs farther, enabling instance-level alignment
3. To verify the effectiveness of our method, we conduct extensive experiments on two datasets in UDA and select multiple state-of-the-art methods as our adversaries. Experiments show that our method has good consistency in UDA. We further conduct comprehensive ablation experiments to verify the effectiveness of our method in different settings.

2 Related Work

2.1 Contrastive Learning

Self-supervised learning aims to improve the feature extraction ability of models by designing auxiliary tasks to mine the representational features of data as supervised information for unlabeled data. [6, 16, 17]. At the same time, thanks to the emergence of contrastive learning, many methods have been proposed to further improve the performance of unsupervised learning by reducing the distance between positive samples. SimCLR [8] is mainly used to generate comparison pairs for the data in the current mini-batch through data augmentation and cosine similarity, which improves the generalization ability of the model; MoCov1 [9] updates the historical features of the stored samples through momentum, so that the contrastive learning samples can contain historical information to obtain better feature representation. Recent research shows that comparative learning is further extended as a paradigm. There are also many methods attempt to contrastive learning from the perspectives of clustering [23, 27, 31]. Inspired by this, we want to achieve instance-level alignment of UDA by contrasting the ability of learning to narrow the distance between positive samples.

2.2 Energy Based Model

The main purpose of an energy model is to construct a function that maps every point in space to a non-probabilistic scalar called energy based model (EBM) was first proposed by LeCun et al. in [29]. Through this non-probabilistic scalar, the problem that the model caused by probability density is difficult to optimize and unstable can be well solved. liu et al. [28] used energy to detect out-of-distribution (OOD) data; and [32] employs a formal connection of machine learning with thermodynamics to characterize the quality of learnt representations for transfer learning, the energy based model has also been explored

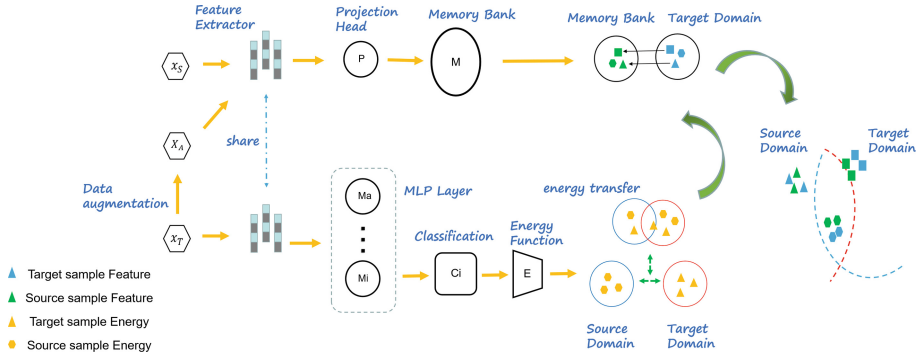


Fig. 2. We use the sample features of different dimensions to generate energy, and complete the knowledge transfer between the source domain and the target domain through energy transfer. Meanwhile, to achieve cross-domain instance-level alignment, we pull the positive samples of the source and target domains closer by contrastive learning.

in domain adaptation. Similarly, we approach the UDA problem from another perspective: taking energy as a domain-specific representation, and completing knowledge transfer in unsupervised domain adaptation through energy transfer.

2.3 Unsupervised Domain Adaption

The main purpose of unsupervised domain adaptation (UDA) is to transfer knowledge in the labeled source domain to the unlabeled target domain. Ben et al. [15] theoretically verifies that reduce the domain gap in the process of training data is more conducive to making the classifier suitable for the target domain. Based on this, reducing the domain gap [1, 3, 35] is a classic method to solve the UDA problem. Without dealing with instance information in each domain data, knowledge transfer can be accomplished by map the data distribution to the Reproducing Kernel Hilbert Space (RKHS) and by convolutional neural network to reduce the domain discrepancy [2, 13].

3 Proposed Method

3.1 Basic Definition

Given the well-annotated source domain $\{(x_i^s, y_i^s)\}_{i=1}^{n_s} = D_s$, and unlabeled target instances $\{(x_j^t)\}_{j=1}^{n_t} = D_t$, and an augmented to the target domain data $\{(x_j^a)\}_{j=1}^{n_a} = D_a$, where N denotes the number of classes. Our aim is to transfer the knowledge learned from the labeled source domain to the unlabeled target domain.

The overall structure of our network is shown in Fig. 2, we extract features with a feature extractor F and define it as f_i . To obtain better feature embeddings, we utilize the projection head P to map the features to the latent contrast

space, defined as p_i . Finally we will go through the classifier C to generate a probabilistic model for each sample.

3.2 Contrastive Learning at the Instance Level

Contrastive learning is a framework that usually uses the context of the same instance to learn representations by discriminating between positive queries and a collection of negative examples in an embedding feature space. We hope to accomplish cross-domain instance-level alignment through its ability to learn representations. However, contrastive learning methods that suitable for unsupervised learning do not involve knowledge transfer between domains, and They tend to fail if there is not enough contrast, e.g. samples in a mini-batch is insufficient.

After exploring a lot of recent work on contrastive learning, we found that memory-bank and data augmentation techniques can be used to reduce the risk of contrastive learning failure. memory-bank [9] can make up for the shortage of samples in mini-batch, while data augmentation can widen the gap between sample representation and facilitate the model to learn instance-level invariant features.

$$P_i^t = momentum \times p_i^{hs} + (1 - momentum) \times p_i^{ns} \quad (1)$$

p_i^{ns} is the contrast feature obtained through the projection head, and p_i^{hs} is the historical feature that already exists in the memory-bank. After obtaining the corresponding features, we compute the feature similarity in the contrast space.

$$sim = \frac{\sum_{i=1}^N (P_i^s \times \sum_{i=1}^n proj(f_i^t))}{\sqrt{\sum_{i=1}^N (P_i^s)^2} \times \sqrt{\sum_{i=1}^n (proj(f_i^t))^2}} \quad (2)$$

Note that in Eq. 2 we involve samples from both source and target domains and complete instance-level alignment. We use contrastive learning to relate samples of the same class whether they are in the same domain or not, which enables knowledge transfer across domains.

We take the sample with the largest feature similarity as the positive sample. We do the same operation on the augmented target domain samples to meet the requirement of discriminating positive and negative samples.

$$Sim_{p1} = \arg \max_D sim(P^s, f_{Au}) \quad (3)$$

$$Sim_{p2} = \arg \max_D sim(P^s, f_t) \quad (4)$$

Based on the above, we can get the final cross-domain contrastive learning loss.

$$L_{CD} = -\log \frac{\exp(S_{p1}/\tau) + \exp(S_{p2}/\tau)}{\exp(S_{p1} + S_{p2}/\tau) + \sum_{j=1}^{2N-1} \exp(N_{2n}/\tau)} \quad (5)$$

where S_{p1} and S_{p2} are both positive sample pairs and N_{2n} is a negative sample pair. Compared with N-pair loss, we learned UDA task by comparison considering knowledge transfer. At the same time, the source domain sample in memory-Bank is used as anchor to realize the instance-level alignment between source domain and target domain. Through formula (5), we implemented compact representation.

3.3 Energy Transfer

We explicitly express our desire to address the problem of domain distribution alignment in UDA problems from another perspective. In some researches of out-of-distribution detection, it has been clearly indicated that out-of-distribution samples have higher energy [28], and the purpose of our energy transfer is to encourage the classifier to obtain the target domain data of the distribution closer to the source domain.

We first introduce the definition of the energy model. The essence of energy-based models (EBM) is to construct a function $E(X)$ that maps each point in the input space to a non-probabilistic scalar called energy [15].

With the Gibbs distribution we can convert the energy into a probability density and get the Gibbs free energy $E(x)$ for any point as:

$$E(x) = -T \log \int_{y'} e^{-E(x,y)}/T \quad (6)$$

where T is the temperature parameter. We can easily associate the classification model with the energy model, and get the free energy for x as:

$$E(x, f) = -T \log \sum_{i=1}^N e^{f_i(x)}/T \quad (7)$$

Note that the energy here has nothing to do with the label of the data, it can be regarded as a kind of norm of the output vector $f_i(x)$. We use the definition of thermodynamic internal energy to express information entropy and energy at the same time [32], the internal energy of a system can be expressed as:

$$U = E + T'G \quad (8)$$

where T' is the temperature parameter, E is the free energy, G is the entropy, and U is the internal energy of the system. The temperature parameter T' is a hyperparameter. Since the classifier and feature extractor share parameters and weights, we believe that in both systems, the source and target domains, the internal energy is not affected by external factors.

However, if the energy transfer between the source and target domains is done directly through the internal energy of the system, two problems arise. First, we cannot guarantee that the energy transfer directions of the source and target domains are consistent. Second, the energy discrepancy of the mini-batch energy in the source and target domains may be too large, making it difficult

for the loss function to converge. Based on this, we restrict the energy transfer loss function as follows:

$$L1 = -\mathbb{E}_{D_s} X_s + \mathbb{E}_{D_t} X_t \quad (9)$$

$$L2 = \mathbb{1}(|U_s - U_t| < \beta) \quad (10)$$

To address the above issues, we use the L_1 proposed in [7] to reduce the range represented by the probability distribution in each mini-batch. The purpose of L_2 is to alleviate the negative optimization caused by the large energy discrepancy.

In Simclr [8], Hinton et al. effectively improves the performance of unsupervised learning through a simple projection head, and in [30], Wang et al. demonstrate in detail that MLP can effectively improve the representation ability of samples. In order to make the internal energy function U better represent the two different systems of the source domain and the target domain, we combine the features of different dimensions through the MLP layer to get a better expression. It can be expressed as:

$$x_i = \sum_{i=1}^K MLP_i(f_i) \quad (11)$$

where the f_i are features of different dimensions. We can integrate features of different dimensions together for better distributional representation, which can more effectively focus on domain-invariant features [7]. So energy expression and the energy transfer loss function can be summarized as follows:

$$E(x, x_i) = -T \log \sum_{i=1}^N e^{x_i} / T \quad (12)$$

$$L_{Trans} = \mathbb{1}(|U_s - U_t| < \beta) + L1 \quad (13)$$

Finally, we utilize a simple cross-entropy loss L_{cls} to guarantee classification accuracy in the source domain, while utilizing the domain adversarial loss L_{Da} [4] as a preliminary transfer target based on classification loss.

Our energy transfer contrast network can effectively learn a special feature representation, and use this to achieve knowledge transfer between source and target domains. Based on this, our overall loss function is as follows:

$$L = L_{cls} + L_{Da} + a \times L_{CD} + b \times L_{Trans}$$

4 Experiment

4.1 Datasets and Criteria

OfficeHome. [10] contains 4 domains, and each domain contains 65 categories, which is the most commonly used dataset for UDA tasks.

Office31. [11] Contains 3 domains, each domain has a total of 31 categories, and there are about 4100 images in total.

Criteria. Following [1, 20], we select two domain pairs (e.g. A2P, P2C) from the dataset for each training, and use the classification accuracy to judge the pros and cons of the model. Finally, we use the average accuracy of all domain pairs as the criterion for evaluating the algorithm

4.2 Implementation Details

For fair comparison, following [7, 13, 19], we use Resnet-50 [12] trained on ImageNet [34] as the backbone for UDA. In this paper, an SGD optimizer with momentum 0.9 is used to train all UDA tasks. The learning rate is adjusted by $l = l_1(1 + \alpha\beta)^\gamma$, where $l_1 = 0.01$, $\alpha = 10$, $\gamma = 0.75$, and β varies from 0 to 1 linearly with the training epochs.

4.3 Comparison with State of the Art

Table 2 shows the performance of different methods on the Office-Home dataset under the UDA task scenario. The experiments are conducted on 12 domain pairs, and we list the average scores under this dataset in the rightmost column. From the table we can observe As a result, our accuracy is at least 0.7% higher than other baselines, and even compared with MetaAlign, which has done further work on GVB, our average accuracy is still improved by 0.3%. Similarly, as shown in Table 1, for the Office31 dataset, our average accuracy is consistent with the state of the art.

Table 1. Accuracy(%) on Office-31 for unsupervised domain adaptation (ResNet-50).

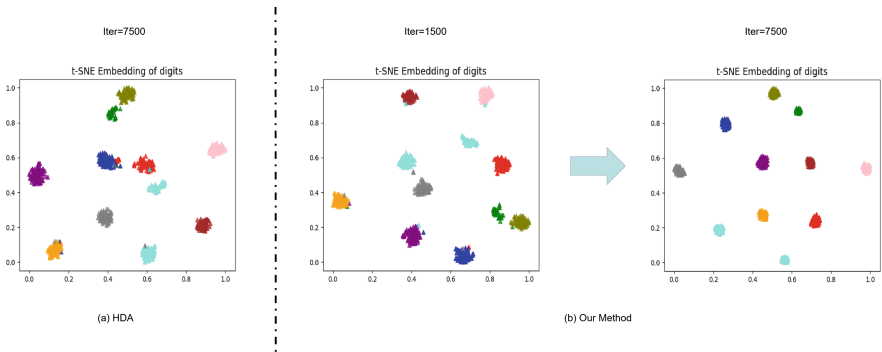
Dataset	Office31						
Task	W2D	D2W	A2W	A2D	D2A	W2A	Avg
Source-Only [12]	99.3	96.7	68.4	68.9	62.5	60.7	76.1
DAN [1]	99.6	97.1	89.9	78.6	63.6	62.8	80.4
JAN [2]	99.8	97.4	89.9	84.7	68.6	70.0	84.3
MDD [14]	100	98.4	94.5	93.5	74.6	72.2	88.9
GSDA [20]	100	99.1	95.7	94.8	73.5	74.9	89.7
CAN [26]	99.8	99.1	94.5	95.0	78.0	77.0	90.5
Ours	100	99.1	95.6	95.2	77.1	75.7	90.5

4.4 Ablation Studies

Feature Visualization. To demonstrate our approach’s achievement of intra-class compactness across samples across domains, we use T-SNE [24] to reduce sample dimensionality and visualize our method and HDA. We randomly selected

Table 2. Accuracy (%) of different UDAs on Office-Home with ResNet-50 as backbone. Best in bold.

Dataset	OfficeHome												
Task	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
Source-Only [12]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
MCD [13]	48.9	68.3	74.6	61.3	67.6	68.8	57.0	47.1	75.1	69.1	52.2	79.6	64.1
GSDA [20]	61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB [19]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
HDA [7]	56.8	75.2	79.8	65.1	73.9	75.2	66.3	56.7	81.8	75.4	59.7	84.7	70.9
MetaAlign [21]	59.3	76.0	80.2	65.7	74.7	75.1	65.7	56.5	81.6	74.1	61.1	85.2	71.3
Ours	57.6	75.3	80.0	65.5	75.4	74.9	66.5	58.0	82.1	76.5	62.1	85.1	71.6

**Fig. 3.** T-SNE visualization results of HDA and our method, which demonstrate that we achieve intra-class compactness.

11 categories from the P-R domain pair of Office-Home, where the same color represents the same label. As shown in Fig. 3, we can achieve the effect of HDA 7500 iterations with only 1500 iterations, and the intra-class compactness gets better with increasing iterations.

The Effect of the Number MLP Module. The energy transfer network is the most important part of our model. By setting up multiple MLP modules for the energy transfer network, we can effectively represent features of different dimensions as energy, providing better generalization ability for the whole model. In this ablation experiment, we selected a total of four sets of experiments with different source domains from the UDA task to explore the effect of different numbers of energy transfer networks on the model. The experimental results are shown in Fig. 4. Experiments show that we choose $N = 3$ as the default number setting for our MLP module.

Selecting Positive Samples by Thresholding in Contrastive Learning. In the experimental process of contrastive learning, we inevitably think of determining positive and negative sample pairs through a threshold τ . When it is

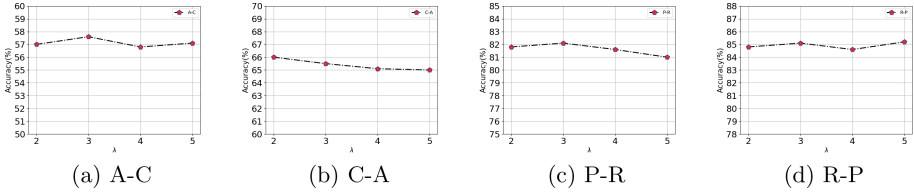


Fig. 4. Influence of the number of networks with different energy transfer on knowledge transfer from source domain to target domain.

greater than the threshold τ , it is a positive sample pair, otherwise it is a negative sample pair. However, there is a fatal problem in the threshold judgment method that the optimal results of different domain pairs are not the same threshold. Figure 5 shows the effect of different thresholds on two UDA tasks.

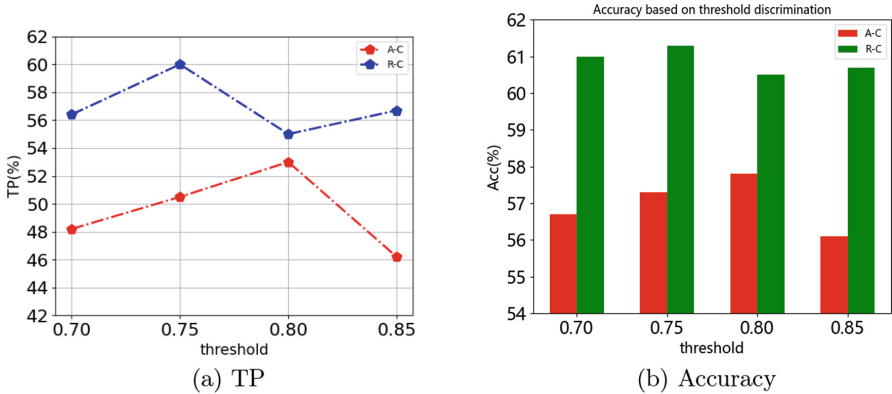


Fig. 5. Figure (a) is the probability of learning true positive pairs under the conditions of setting different thresholds, and Figure (b) is the accuracy rate generated by setting different thresholds. As can be seen from the figure, different domain pairs apply to different thresholds.

5 Conclusion

In this paper, we propose an energy representation to further improve the accuracy of UDA tasks. Specifically, we extract information of different dimensions of features through multiple MLP layers, and then represent the difference between the source and target domains through the combination of entropy and free energy, and mitigate the effect of domain shift by reducing this gap. At the same time, we also achieve instance-level alignment across domains through

contrastive learning. Furthermore, our method is compared with many previous state-of-the-art on three datasets, which demonstrates the effectiveness of our method.

Acknowledgement. This work was supported by the Natural Science Foundation of China under Grant 41906177 and 41927805, the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City under Grants 2021JJLH0061, The National Key Research and Development Program of China under Grants 2018AAA0100605.

References

1. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: PMLR, pp. 97–105 (2015)
2. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: PMLR, pp. 2208–2217 (2017)
3. Kang, G., Zheng, L., Yan, Y., Yang, Y.: Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 420–436. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_25
4. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: PMLR, pp. 1180–1189 (2015)
5. Hoffman, J., et al.: Cycada: cycle-consistent adversarial domain adaptation. In: ICML. PMLR, pp. 1989–1998 (2018)
6. Shao, H., Yuan, Z., Peng, X., Wu, X.: Contrastive learning in frequency domain for non-i.I.D. image classification. In: International Conference on Multimedia Modeling (2021)
7. Cui, S., Jin, X., Wang, S., He, Y., Huang, Q.: Heuristic domain adaptation. In: NeurIPS vol. 33, pp. 7571–7583 (2020)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: PMLR, vol. 10, pp. 1597–1607 (2020)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
10. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR, pp. 5018–5027 (2017)
11. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 213–226. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_16
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
13. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR, pp. 3723–3732 (2018)
14. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: ICML. PMLR, pp. 7404–7413 (2019)
15. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NeurIPS, vol. 19 (2006)

16. Chen, Y.-C., Gao, C., Robb, E., Huang, J.-B.: NAS-DIP: learning deep image prior with neural architecture search. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12363, pp. 442–459. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_26
17. Saito, K., Saenko, K.: OVANet:: one-vs-all network for universal domain adaptation. In: *ICCV*, pp. 9000–9009 (2021)
18. Singh, A.: CLDA: contrastive learning for semi-supervised domain adaptation. In: *NerulPS*, vol. 34 (2021)
19. Cui, S., Wang, S., Zhuo, J., Su, C., Huang, Q., Tian, Q.: Gradually vanishing bridge for adversarial domain adaptation. In: *CVPR*, pp. 12 455–12 464 (2020)
20. Hu, L., Kan, M., Shan, S., Chen, X.: Unsupervised domain adaptation with hierarchical gradient synchronization. In: *CVPR*, pp. 4043–4052 (2020)
21. Wei, G., Lan, C., Zeng, W., Chen, Z.: Metaalign: coordinating domain alignment and classification for unsupervised domain adaptation. In: *CVPR*, pp. 16 643–16 653 (2021)
22. James, S., et al.: Sim-to-real via sim-to-sim: data-efficient robotic grasping via randomized-to-canonical adaptation networks. In: *CVPR*, pp. 12 627–12 637 (2019)
23. Wang, Y., et al.: Clusterscl: cluster-aware supervised contrastive learning on graphs. In: *Proceedings of the ACM Web Conference 2022*, 1611–1621 (2022)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11) (2008)
25. Goodfellow, I.J., et al.: Generative adversarial nets. In: *NeurIPS* (2014)
26. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: *ICCV*, pp. 4893–4902 (2019)
27. Zhong, H., et al.: Graph contrastive clustering. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9224–9233 (2021)
28. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475 (2020)
29. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. In: *Predicting Structured Data*, vol. 1, no. 0 (2006)
30. Wang, Y., et al.: Revisiting the transferability of supervised pretraining: an MLP perspective. *arXiv preprint [arXiv:2112.00496](https://arxiv.org/abs/2112.00496)* (2021)
31. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. In: *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(10), 8547–8555 (2021)
32. Gao, Y., Chaudhari, P.: A free-energy principle for representation learning. In: *International Conference on Machine Learning*. PMLR, pp. 3367–3376 (2020)
33. Vu, T.-H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: adversarial entropy minimization for domain adaptation in semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526 (2019)
34. Deng, J., et al.: ImageNet: a large-scale hierarchical image database. In: *CVPR*, pp. 248–255 (2009)
35. Wang, F., Ding, Y., Liang, H., Wen, J.: Discriminative and selective pseudo-labeling for domain adaptation. In: *International Conference on Multimedia Modeling* (2021)