Supplementary Material
Promoting Semantic Connectivity: Dual Nearest Neighbors Contrastive Learning for Unsupervised Domain Generalization

A. Proof of the Proposition

A.1 Proof of Proposition 1

Proposition 1. *For stronger augmentations $\hat{A}$, i.e., $A \subseteq \hat{A}$, augmented views have smaller intra-domain connectivity as $\hat{C}_\alpha := \mathbb{E}_{d\sim P_D^S}\mathbb{E}_{x_i,x_j\sim P_d^{S_{\mathrm{UL}}}}[\hat{\mathcal{A}}(x_i^+|x_i)\hat{\mathcal{A}}(x_j^+|x_j)]$.*

Proof. Without loss of generality, we consider two given samples x_i and x_j belonging to the same domain, *i.e.*, $d_i = d_j$. For a given data augmentation set A, we first define the augmented distance between two samples as the maximum distance between their augmented views as

$$d_A\left(x_i^+, x_j^+\right) = \max_{x_i^+\in A(x_i), x_j^+\in A(x_j)} \left\|x_i^+ - x_j^+\right\| \tag{S-1}$$

Since vanilla augmentations are included in strong augmentations, *i.e.*, $A \subseteq \hat{A}$, we have the inequality as $d_{\hat{A}}\left(x_i^+, x_j^+\right) \geq d_A\left(x_i^+, x_j^+\right)$. Correspondingly, we have the supremum of the distance of two augmented view sets as

$$\mathcal{V}_A \triangleq \sup \rho(A(x_i), A(x_j)) = d_A(x_i^+, x_j^+) \tag{S-2}$$

$$\mathcal{V}_{\hat{A}} \triangleq \sup \rho(\hat{A}(x_i), \hat{A}(x_j)) = d_{\hat{A}}(x_i^+, x_j^+) \tag{S-3}$$

Since $d_{\hat{A}}\left(x_i^+, x_j^+\right) \geq d_A\left(x_i^+, x_j^+\right)$, we have $\mathcal{V}_{\hat{A}} \geq \mathcal{V}_A$. Then we define the overlap of two distributions as

$$\phi_{\mathcal{A}} \triangleq \mathrm{Supp}(\mathcal{A}(x_i^+ \mid x_i)) \bigcap \mathrm{Supp}(\mathcal{A}(x_j^+ \mid x_j)) \tag{S-4}$$

$$\phi_{\hat{\mathcal{A}}} \triangleq \mathrm{Supp}(\hat{\mathcal{A}}(x_i^+ \mid x_i)) \bigcap \mathrm{Supp}(\hat{\mathcal{A}}(x_j^+ \mid x_j)) \tag{S-5}$$

Since $\mathcal{V}_{\hat{A}} \geq \mathcal{V}_A$, we have $\phi_{\hat{\mathcal{A}}} \subseteq \phi_{\mathcal{A}}$. Then for a given data augmentation set A, we define the minimum product of two augmented samples as

$$e_A\left(x_i^+, x_j^+\right) = \min_{x_i^+\in A(x_i), x_j^+\in A(x_j)} x_i^+ x_j^+ \tag{S-6}$$

Since vanilla augmentations are included in strong augmentations, *i.e.*, $A \subseteq \hat{A}$, we have the inequality as $e_{\hat{A}}\left(x_i^+, x_j^+\right) \leq e_A\left(x_i^+, x_j^+\right)$. We assume the mean of the product value in the overlap part of distributions as a constant multiple of the minimum product. Then we have $\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)$ and $\hat{\mathcal{A}}(x_i^+|x_i)\hat{\mathcal{A}}(x_j^+|x_j)$ as

$$\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j) = \begin{cases} 0 & x_i^+, x_j^+ \notin \phi_{\mathcal{A}} \\ C\cdot e_A\left(x_i^+, x_j^+\right) & x_i^+, x_j^+ \in \phi_{\mathcal{A}} \end{cases} \tag{S-7}$$

$$\hat{\mathcal{A}}(x_i^+|x_i)\hat{\mathcal{A}}(x_j^+|x_j) = \begin{cases} 0 & x_i^+, x_j^+ \notin \phi_{\hat{\mathcal{A}}} \\ C\cdot e_{\hat{A}}\left(x_i^+, x_j^+\right) & x_i^+, x_j^+ \in \phi_{\hat{\mathcal{A}}} \end{cases} \tag{S-8}$$

Since $e_{\hat{A}}\left(x_i^+, x_j^+\right) \leq e_A\left(x_i^+, x_j^+\right)$ and $\phi_{\hat{\mathcal{A}}} \subseteq \phi_{\mathcal{A}}$, $\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j) \geq \hat{\mathcal{A}}(x_i^+|x_i)\hat{\mathcal{A}}(x_j^+|x_j)$. Consequently, we have

$$\begin{aligned}&\mathbb{E}_{d\sim P_D^S}\mathbb{E}_{x_i,x_j\sim P_d^{S_{\mathrm{UL}}}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)] \\ &\qquad \geq \mathbb{E}_{d\sim P_D^S}\mathbb{E}_{x_i,x_j\sim P_d^{S_{\mathrm{UL}}}}[\hat{\mathcal{A}}(x_i^+|x_i)\hat{\mathcal{A}}(x_j^+|x_j)]\end{aligned} \tag{S-9}$$

Thus, we draw the conclusion $\hat{C}_\alpha < C_\alpha$. □

A.2 Proof of Proposition 2

Proposition 2. *Dual nearest neighbors can increase the intra-class connectivity as $\hat{C}_\beta := \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P_y^{S_{\mathrm{UL}}}}$ $[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]$, where $\hat{C}_\beta > C_\beta$. More accurate cross domain NN and more diverse in-domain NN can further increase intra-class connectivity.*

Proof. To calculate the intra-class connectivity, we first divide all the samples into two parts: intra-domain intra-class samples and cross-domain intra-class samples. Correspondingly, the intra-class connectivity can be calculated as the sum of cross-domain intra-class connectivity and intra-domain intra-class connectivity.

$$\begin{aligned}C_\beta &= \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P_y^{S_{\mathrm{UL}}}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)] \\ &= \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P_{y,d_i\neq d_j}^{S_{\mathrm{UL}}}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)] \\ &+ \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P_{y,d_i= d_j}^{S_{\mathrm{UL}}}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)]\end{aligned} \tag{S-10}$$

Without loss of generality, we consider two given samples x_i and x_j belonging to the different domains with the same semantic class, *i.e.*, $d_i \neq d_j$ and $y_i = y_j$. Given a data augmentation set A, we define the overlap of two distributions as

$$\phi_{d_i \neq d_j} \triangleq \text{Supp}(\mathcal{A}(x_i^+ \mid x_i)) \bigcap \text{Supp}(\mathcal{A}(x_j^+ \mid x_j)) \tag{S-11}$$

For a given data augmentation set A, transformations cannot overcome significant distribution shifts across different domains, *e.g.*, one can hardly transform a cat in sketch to photo. Thus, we have $\phi_{d_i \neq d_j} \simeq \varnothing$.

While we search for cross domain nearest neighbors (NN) in the latent embedding space as the positive sample. Denote the nearest neighbors of x_j in domain i as $N_i(x_j)$. We have the overlap of distributions as

$$\hat{\phi}^N_{d_i \neq d_j} \triangleq \text{Supp}(\mathcal{A}(x_i^+ \mid x_i)) \bigcap \text{Supp}(\mathcal{A}(N(x_j)^+ \mid N(x_j))) \tag{S-12}$$

Since $N_i(x_j)$ is in the same domain with x_i with similar semantic information, the augmentation overlap exists. Then, we have $\hat{\phi}^N_{d_i \neq d_j} > \varnothing$ and $\hat{\phi}^N_{d_i \neq d_j} > \phi_{d_i \neq d_j}$. Thus, we have

$$\begin{aligned}&\mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i\neq d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]\\ &> \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i\neq d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)]\end{aligned} \tag{S-13}$$

Besides, we consider two given samples x_i and x_j belonging to the same domains with the same semantic class, i.e., $d_i = d_j$ and $y_i = y_j$. Similarly, we have the distribution overlap as $\hat{\phi}_{d_i=d_j}$. Though $\phi_{d_i=d_j} > \varnothing$, the overlap is limited by some intra-domain intra-class semantic variances. Comparably, our intra-domain nearest neighbors (NN) can overcome intra-domain variances with the increased overlap as $\hat{\phi}^N_{d_i=d_j} > \phi_{d_i=d_j}$. Thus, we have

$$\begin{aligned}&\mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i=d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]\\ &> \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i=d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)]\end{aligned} \tag{S-14}$$

Combined with Eq. (S-13), we have

$$\begin{aligned}&\mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i\neq d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]\\ &+ \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i=d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]\\ &> \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i\neq d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)]\\ &+ \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i=d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(x_j^+|x_j)]\end{aligned} \tag{S-15}$$

Totally, we draw the conclusion $\hat{C}_\beta > C_\beta$.

For more accurate cross domain NN, since the searched neighbors are more likely to belong to the same semantic class, the searched $N_i'(x_j)$ share more similar semantic information with x_i, which results in a larger augmentation overlap as $\hat{\phi}^{N'}_{d_i\neq d_j} > \hat{\phi}^N_{d_i\neq d_j}$. Thus, we have

$$\begin{aligned}&\mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i\neq d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}'(x_j)^+|\mathcal{NN}'(x_j))]\\ &> \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i\neq d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]\end{aligned} \tag{S-16}$$

For more diverse in-domain NN, since the searched neighbors are more likely to overcome more severe intra-domain variances, the searched $N_i'(x_j)$ can lead to a larger augmentation overlap as $\hat{\phi}^{N'}_{d_i=d_j} > \hat{\phi}^N_{d_i=d_j}$. Thus, we have

$$\begin{aligned}&\mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i=d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}'(x_j)^+|\mathcal{NN}'(x_j))]\\ &> \mathbb{E}_{y\sim P_Y^S}\mathbb{E}_{x_i,x_j\sim P^{S_{\text{UL}}}_{y,d_i=d_j}}[\mathcal{A}(x_i^+|x_i)\mathcal{A}(\mathcal{NN}(x_j)^+|\mathcal{NN}(x_j))]\end{aligned} \tag{S-17}$$

Combined with Eq. (S-16), we can draw the conclusion that more accurate cross domain NN and more diverse in-domain NN can further increase the intra-class connectivity. □

A.3 Proof of Proposition 3

Proposition 3. *Our proposed CD^{2}NN is more accurate than cross domain NN in the UDG setting.*

Proof. Denote e_{cr} as the error rate of the cross domain nearest neighbor and e_{in} as the error rate of the in-domain nearest neighbor. For simplicity, we assume the error rate of the second nearest neighbor is also e_{cr} and e_{in} for cross domain and in-domain, respectively. We assume when the nearest neighbor is wrong, it is equally likely to match to any one of the remaining $C-1$ classes, where C is the total number of classes.

Considering a given query z, the error rate of the vanilla cross domain NN is $P_{\text{vanilla}} = e_{cr}$.

For our proposed CD^{2}NN strategy shown in Figure 3, if $\mathcal{R}_1 \neq \varnothing$, *i.e.*, our CD^{2}NN selects the NN in $\mathcal{R}_1$. The selected NN is wrong only if the following two conditions are met: 1) The cross domain NN of z is wrong; 2) The in-domain NN $z_{nn}^{q_{in}}$ of z is right and cross domain NN of $z_{nn}^{q_{in}}$ is wrong **or** the in-domain NN $z_{nn}^{q_{in}}$ of z is wrong and the cross domain NN of $z_{nn}^{q_{in}}$ is not in the same class as z. Thus, the error rate of our CD^{2}NN is

$$\begin{aligned}P^{\mathcal{R}_1}_{\text{CD}^2\text{NN}} &= e_{cr}\cdot\left((1-e_{in})\cdot e_{cr} + e_{in}\cdot(1-e_{cr}+e_{cr}\cdot\frac{C-2}{C-1})\right)\\ &= e_{cr}\cdot\left((1-e_{in})\cdot e_{cr} + e_{in}\cdot(1-e_{cr}\cdot\frac{1}{C-1})\right)\\ &< e_{cr}\cdot((1-e_{in})\cdot e_{cr}+e_{in}).\end{aligned} \tag{S-18}$$

Then, we have

$$P^{\mathcal{R}_1}_{\text{CD}^2\text{NN}} < e_{cr}\cdot(e_{cr}+e_{in}-e_{cr}\cdot e_{in}). \tag{S-19}$$

Since $0 < e_{cr} < 1$ and $0 < e_{in} < 1$, we have

$$e_{cr} + e_{in} - e_{cr} \cdot e_{in} - 1 = (e_{cr} - 1) \cdot (1 - e_{in}) < 0. \quad \text{(S-20)}$$

Thus, $e_{cr} + e_{in} - e_{cr} \cdot e_{in} < 1$ and $e_{cr} \cdot (e_{cr} + e_{in} - e_{cr} \cdot e_{in}) < e_{cr}$. Since $P_{\text{vanilla}} = e_{cr}$, from Equation (S-19), we have:

$$P^{\mathcal{R}_1}_{\text{CD}^2\text{NN}} < e_{cr} \cdot ((1 - e_{in}) \cdot e_{cr} + e_{in}) < P_{\text{vanilla}}. \quad \text{(S-21)}$$

As shown in Figure 3, if $\mathcal{R}_1 = \varnothing$ and $\mathcal{R}_2 \neq \varnothing$, *i.e.*, our CD^{2}NN selects the NN in $\mathcal{R}_2$. The selected NN is wrong only if the following two conditions are met: 1) The cross domain NN of z is wrong; 2) The cross domain NN $z^{q_{cr}}_{nn}$ of z is right and in-domain NN of $z^{q_{cr}}_{nn}$ is wrong **or** the cross domain NN $z^{q_{cr}}_{nn}$ of z is wrong and the in-domain NN of $z^{q_{cr}}_{nn}$ is not in the same class as z. Thus, the error rate of our CD^{2}NN is

$$\begin{aligned} P^{\mathcal{R}_2}_{\text{CD}^2\text{NN}} &= e_{cr} \cdot \left((1 - e_{cr}) \cdot e_{in} + e_{cr} \cdot (1 - e_{in} + e_{in} \cdot \frac{C-2}{C-1}) \right) \\ &= e_{cr} \cdot \left((1 - e_{cr}) \cdot e_{in} + e_{cr} \cdot (1 - e_{in} \cdot \frac{1}{C-1}) \right) \\ &< e_{cr} \cdot ((1 - e_{cr}) \cdot e_{in} + e_{cr}) \end{aligned} \quad \text{(S-22)}$$

Similarly, we have

$$P^{\mathcal{R}_2}_{\text{CD}^2\text{NN}} < e_{cr} \cdot (e_{in} + e_{cr} - e_{cr} \cdot e_{in}), \quad \text{(S-23)}$$

Since $0 < e_{cr} < 1$ and $0 < e_{in} < 1$, we have

$$e_{in} + e_{cr} - e_{cr} \cdot e_{in} - 1 = (e_{cr} - 1) \cdot (1 - e_{in}) < 0 \quad \text{(S-24)}$$

Thus, $e_{in} + e_{cr} - e_{cr} \cdot e_{in} < 1$ and $e_{cr} \cdot (e_{in} + e_{cr} - e_{cr} \cdot e_{in}) < e_{cr}$. Since $P_{\text{vanilla}} = e_{cr}$, from Equation (S-23), we have: and

$$P^{\mathcal{R}_2}_{\text{CD}^2\text{NN}} < e_{cr} \cdot ((1 - e_{cr}) \cdot e_{in} + e_{cr}) < P_{\text{vanilla}}. \quad \text{(S-25)}$$

Totally, since $P^{\mathcal{R}_1}_{\text{CD}^2\text{NN}} < P_{\text{vanilla}}$ and $P^{\mathcal{R}_2}_{\text{CD}^2\text{NN}} < P_{\text{vanilla}}$, we have $P_{\text{CD}^2\text{NN}} < P_{\text{vanilla}}$. Thus, we show theoretically that Our proposed CD^{2}NN is more accurate than cross domain nearest neighbor in this specific domain generalization setting. □

B. More Visualizations

B.1 Intra-class connectivity of our method.

We add analysis from the connectivity perspective. This experiment is conducted without the in-domain cycle NN to evaluate the performance of CD^{2}NN. Specifically, we train the unsupervised model on PACS with three strategies, *i.e.*, in-domain NN, vanilla cross domain NN and our CD^{2}NN, respectively, and compute the corresponding intra-class connectivity. Fig. S-1 shows that our CD^{2}NN can increase the intra-class connectivity as the training proceeds. Vanilla cross domain NN selection strategy suffers the limited intra-class connectivity gain due to many wrong NN matches brought by domain shifts. In-domain NN selection strategy achieves satisfactory intra-class connectivity at the beginning of training by clustering more accurate in-domain neighbors. However, in-domain NN selection strategy fails to overcome distribution shifts across domains and cannot align intra-class samples from different domains, which suffers the degraded intra-class connectivity eventually.

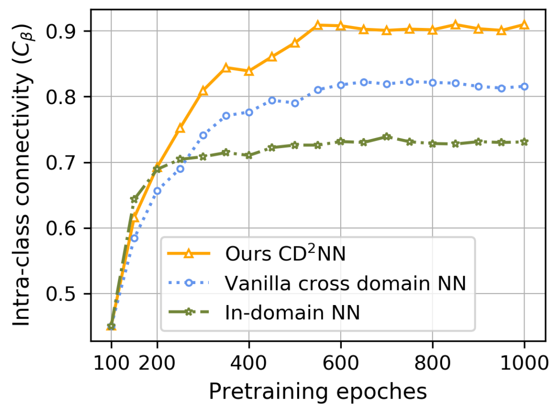


Figure S-1. Intra-class connectivity for the model trained with in-domain NN, vanilla cross domain NN and our CD^{2}NN strategy.

B.2 Nearest neighbors searched by our CD^{2}NN.

In Fig. S-2, we showcase the domain invariant capabilities of the feature representation learned without supervision using our DN^{2}A approach. Each example shows the top-5 nearest neighbors of a random query image (from the PACS dataset) searched in the entire set of images of each of the four different PACS domains: *Photo*, *Art painting*, *Cartoon* and *Sketch*. All images are encoded using our self-supervised model trained on three domains (*Art painting*, *Cartoon* and *Sketch*) of PACS dataset.

C. More Experiments

C.1 Experiments on Open Domain Generalization

We follow the open domain generalization setting, *i.e.*, the class split for each domain, in DAML [13] to conduct experiments on PACS dataset. We train the model on the unlabeled source data using SimCLR and our DN^{2}A with the same experimental setting in the main text. Besides, we also conduct unsupervised pre-training based on ImageNet initialization (as seen in the bottom half of Table S-1). Then we use the parameters of the trained model as the initialization for the SOTA open domain generalization method DAML [13].

As shown in Table S-1, DAML benefits from unsupervised pre-training. Compared with the random initializa-

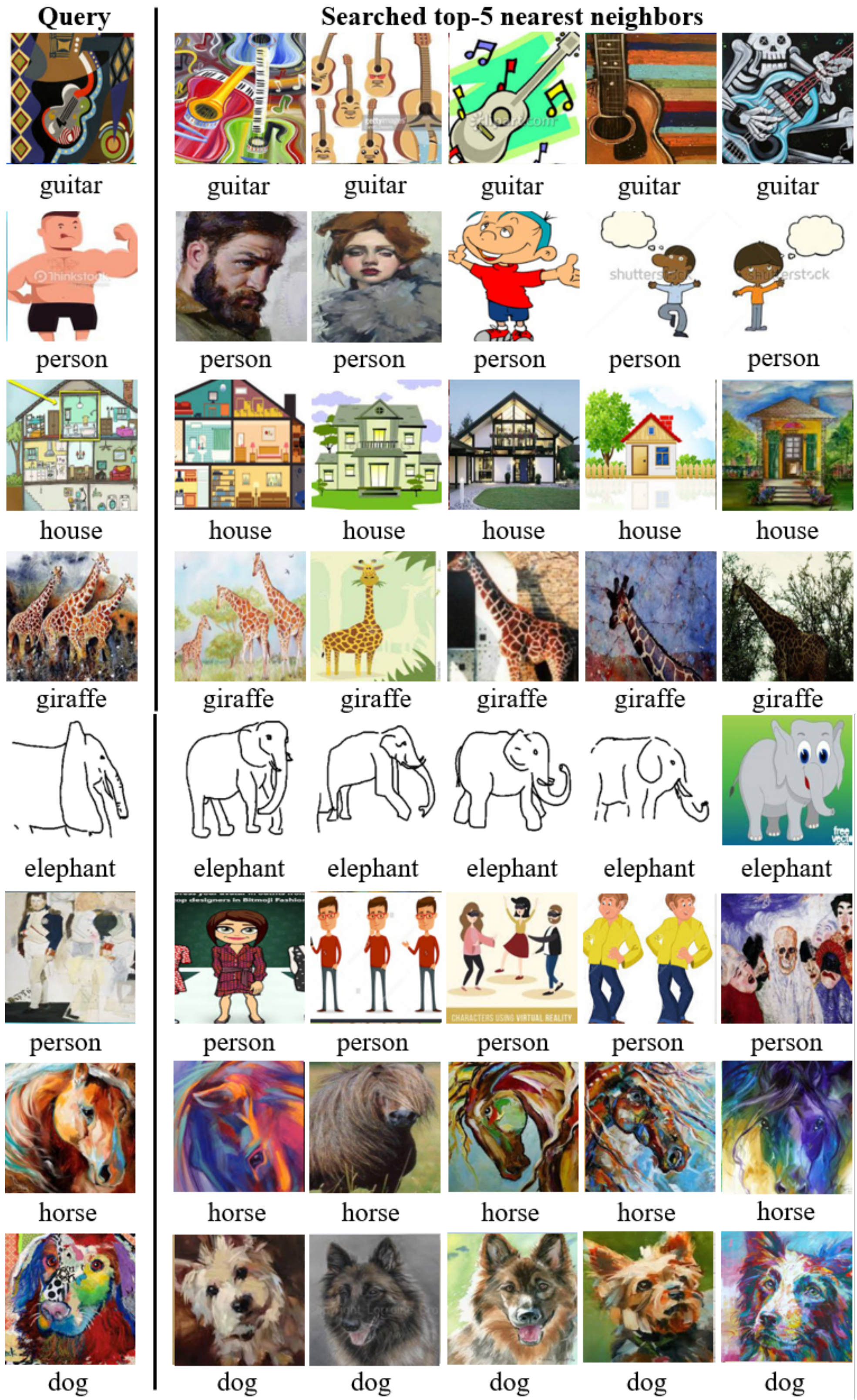


Figure S-2. Nearest neighbors searched by our DN^{2}A method.

Method	Art		Sketch		Photo		Cartoon		Avg	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
DAML(random init.) [13]	26.20	17.04	24.81	21.04	21.89	16.42	39.92	20.26	28.21	18.69
SimCLR + DAML	35.07	25.91	42.61	31.53	28.33	23.86	48.65	31.66	38.67	28.24
Ours + DAML	**43.86**	**37.30**	**56.42**	**51.09**	**40.67**	**32.37**	**62.81**	**46.54**	**50.94**	**41.83**
DAML(ImageNet init.) [13]	54.10	43.02	58.50	56.73	75.69	53.29	73.65	54.47	65.49	51.88
Ours† + DAML	**62.75**	**49.16**	**69.96**	**61.91**	**76.06**	**59.11**	**76.26**	**61.59**	**71.26**	**57.94**

Table S-1. Results with different initialization methods under the open-domain setting on PACS. † indicates that the unsupervised pre-training is based on ImageNet initialization.

Method	Office: Target Acc. on 1-shot / 3-shots						
	A→D	A→W	D→A	D→W	W→A	W→D	Avg
CDS [9]	48.3 / 65.9	49.2 / 65.5	61.4 / 64.4	77.5 / 90.4	57.4 / 64.4	71.5 / 93.0	60.9 / 73.9
PCS [16]	60.2 / 78.2	69.8 / 82.9	76.1 / 76.4	90.6 / 94.1	71.2 / 76.3	91.8 / 96.0	76.6 / 84.0
PCS w/o APCU & MIM	47.2 / 71.1	52.7 / 70.6	59.0 / 75.5	76.4 / 90.3	58.5 / 74.1	66.9 / 91.8	60.1 / 78.9
Ours	50.8 / 72.4	54.9 / 71.2	65.1 / 69.7	77.6 / 90.8	62.6 / 71.9	71.5 / 93.1	63.8 / 78.2

Table S-2. Target accuracy (%) on few-shot domain adaptation with source 1-shot and 3-shots labels per class on the Office dataset.

tion and ImageNet initialization, our method can improve DAML for 22.73% and 6.50% average accuracy, respectively. Compared with SimCLR, our DN^{2}A provides a much stronger initialization and boosts the generalization ability of DAML with a 12.27% improvement in average accuracy and a 13.59% improvement in average H-score, demonstrating the effectiveness of our method in the open-set DG setting.

Open-set DG setting assumes different source domains contain private classes and shared classes. With our proposed DN^{2}A, samples in different domains from the shared classes can be aligned. For open-set samples in private classes of some domains (no cross domain neighbors of the same class), our proposed cross domain double-lock NN selection strategy can filter out these untrustworthy noisy neighbors not used as positive samples, *i.e.*, $\mathcal{R} = \varnothing$. With positive samples generated by strong augmentation to suppress the domain information, our method learns class-semantic similarity by separating visually dissimilar images, and eventually separates the private classes from the shared classes. Totally, our method can align shared classes in different source domains, while separating shared classes from private classes. Thus, our proposed method can achieve good performance for the challenging open-set DG setting.

C.2 Experiments on Few-shot Domain Adaptation

We follow the few-shot domain adaptation protocol defined in [16] with the same data split, where the source domain has a single or three labeled images per class and the remaining images are provided as unlabeled. Following [9, 16], we use the Resnet-50 pretrained on ImageNet as the backbone, and use 1 or 3 source domain samples per class for the source-only training.

As shown in Table S-2, our method outperforms CDS by 2.7% average accuracy for 1-shot adaptation. CDS assumes samples of the same class are closer than other samples of different classes across different domains, and directly applies the cross domain matching, which suffers from false matches and introduces the noise to compromise the final performance. As an end-to-end framework proposed for domain adaptation, PCS aims to learn a model that could achieve high accuracy on the target domain. Thus, PCS achieves the highest accuracy with adaptive prototypical classifier learning (consisting of Adaptive Prototype-Classifier Update (APCU) and Mutual Information Maximization (MIM)) for the target domain. We also take the result without APCU and MIM from [16] for a relatively fair comparison of the cross domain self-supervised learning strategy itself. Our method outperforms by 3.7% average accuracy for 1-shot adaptation. The proposed instance-prototype cross domain matching [16] also suffers from the matching noise and degrades the performance.

C.3 Comparison with MIM-based Methods

Recently, mask image modeling-based methods [4, 8, 15] have made growing progress. Table S-4 shows DN^{2}A significantly outperforms MIM-based models with various portions of labeled data. For example, with 10% labeled data, our DN^{2}A outperforms MAE by 33.41% accuracy and DiMAE by 6.55%, respectively. With 1% labeled data, our DN^{2}A outperforms MAE by 30.93% accuracy and DiMAE

Operation	ShearX(Y)	TranslateX(Y)	Rotate	AutoContrast	Identity	Equalize
Mag Range	[-0.3,0.3]	[-0.3,0.3]	[-30,30]	0 or 1	0 or 1	0 or 1
Operation	Solarize	Posterize	Contrast	Color	Brightness	Sharpeness
Mag Range	[0,256]	[4,8]	[0.05,0.95]	[0.05,0.95]	[0.05,0.95]	[0.05,0.95]

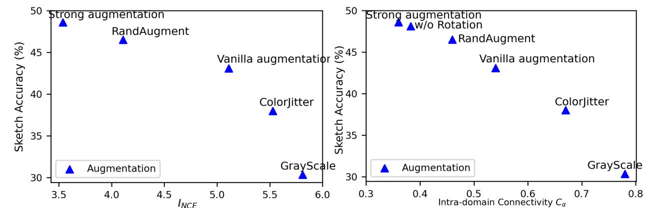
Table S-3. Various augmentations we applied to strongly augment the training images.

Label Fraction	MAE [8]	DiMAE [15]	DN^{2}A (Ours)
1%	24.89	34.23	**55.82**
5%	28.77	40.91	**62.89**
10%	31.79	58.65	**65.20**

Table S-4. Accuracy on PACS compared with MAE and DiMAE.

	Photo	Art.	Cartoon	Sketch	Avg.
Baseline+SA	53.77	34.08	40.64	48.58	44.27
+GT Positive	68.19	50.24	56.52	61.17	59.03
+GT Negative	57.82	38.03	44.89	51.06	47.95
Ours DN^{2}A	67.84	44.06	53.98	57.43	55.82
+GT Negative	69.14	45.91	55.83	58.22	57.27
+FNE	68.26	44.38	53.95	57.72	56.08

Table S-5. Ablation study on the impact of negative samples.



(a) I_{NCE} v.s Accuracy (b) C_α v.s Accuracy

Figure S-3. I_{NCE} and C_α v.s accuracy with data augmentations.

T (epoch)	Photo	Art.	Cartoon	Sketch	Avg.
0	61.32	39.59	46.17	51.92	49.75
50	66.09	42.35	51.83	55.61	53.97
100	67.84	44.06	53.98	57.43	55.82
200	67.41	44.17	52.81	56.59	55.25

Table S-6. Ablation study on T with kNN accuracy.

by 21.59%, respectively. Experimental results demonstrate that our DN^{2}A is more effective than MIM-based methods in learning domain-invariant features using unlabeled data.

C.4 Discussion on Noisy Negative Samples

To evaluate the impact, we select negatives from truly different classes using ground-truth (GT) labels and show in Table S-5 that GT Negative improves performance by mitigating noise. However, the impact of GT Positive is much greater than GT Negative. In fact, the success of contrastive learning relies heavily on positives [6, 14] rather than negatives, where positives are crucial for learning semantic invariance while negatives serve to avoid model collapse. Thus, we focus on positive selection. Moreover, our dual NN can improve the robustness by making same-class embeddings closer (GT Negative achieves a slight gain). To further mitigate the noise of negatives, we utilize our dual NNs as queries to search their in-domain NNs in the mini-batch as False Negatives and Eliminate them from computing the loss. Table S-5 shows FNE yields performance gain.

C.5 Discussion on Strong Augmentation

Fig. S-3(b) shows intra-domain connectivity decreases as we strengthen the family of augmentations by including more functions, which is consistent with Proposition 1. Thus, we use all functions in the PIL library to build strong augmentation. We design augmentations to ensure low intra-domain connectivity to facilitate contrastive learning rather than rely on domain shift that is subtle and unknown in target domain. Augmentations like Color and Sharpness destroy the color and texture information and are closely related to domain bias, whereas less-related augmentations like Rotation also reduce the connectivity for better performance (0.47% accuracy loss w/o Rotation in Fig. S-3(b)).

Besides, Mutual information I can measure the amount of information shared by positives. We use I_{NCE} as a neural proxy to estimate I. Fig. S-3(a) shows the amount of shared information decreases as we strengthen the family of augmentations. Thus, we deem strong augmentations as augmentation with a certain low level of shared information that can prevent the failure of contrastive learning. In fact, intra-domain connectivity shares the same trend with I_{NCE} in indicating strong augmentations. For general purposes, we use all label-preserving augmentations in the PIL library. As mentioned in Limitations, specific augmentations related to the dataset (e.g., style transfer) may further reduce the shared information for better performance.

C.6 Ablation on Hyperparameters

Hyperparameter T controls the epoch to introduce the contrastive loss of positives generated by our dual NNs. Table S-6 shows the performance is best at $T = 100$ but degraded at $T = 0$ due to noisy neighbors in random initialization and at $T = 200$ due to late introduction of NN as positives.

D. Related Works

D.1 Unsupervised Domain Adaptation (UDA)

UDA aims to transfer the knowledge from a labeled source domain to an unlabeled target domain. Haeusser et al. [5] propose the association loss as a discrepancy measure to enforce associations between source and target data for producing statistically domain invariant embeddings. Li et al. [11] propose domain consensus clustering to learn the intrinsic structure of the target domain via encouraging discriminative target clusters. Chen et al. [2] achieve the feature alignment via mutual nearest neighbors contrast and exploit domain discrimination knowledge by hybrid prototype self-training.

D.2 Self-supervised Learning for UDA

Recently, self-supervised learning is introduced into domain adaptation. CDS [9] is proposed to perform self-supervised learning (SSL) not only within a single domain but also across two domains for better domain adaptation performance. PCS [16] further extends the instance-wise SSL in CDS to prototypical SSL, and proposes a powerful end-to-end framework for domain adaptation. Our DN^{2}A is different from CDS and PCS in the starting point. CDS and PCS are proposed for domain adaptation, where there are two domains and the goal is the target domain alignment. While our method is proposed for domain generalization with multiple domains (more than two domains) to learn domain invariant features. Notice that the cross-domain matching strategy, which is the key component of CDS and PCS for domain alignment, cannot be easily extended to multiple domains. Directly matching each pair of multiple domains may cause a negative transfer, especially for open-set samples. While our method can flexibly find the neighbors in the right domain among multiple domains (also applicable for two domains) for domain-invariant learning. Secondly, CDS assumes that samples of the same class are closer than other samples of different classes across different domains, and uses entropy minimization to implicitly discover and enforce the similarity between cross domain pairs, which suffers from the match noise brought by the domain gap and can be deemed as the vanilla cross domain NN selection counterpart of our method. Though PCS proposes the instance-prototype matching to mitigate the noise, the performance is undermined, especially for open-set samples, where there could be no positive matches from the same class. PCS indiscriminately pushes these negative matches together, while our cross domain double-lock NN (CD^{2}NN) can avoid this situation by excluding the untrustworthy negative matches from training. Thus, our proposed CD^{2}NN strategy is more flexible, effective and robust for cross domain matching, and can be used in CDS and PCS as a superior alternative of their cross domain SSL strategy to boost the performance for domain adaptation tasks. Besides, our CD^{2}NN strategy can extend CDS and PCS to multi-source domain adaptation tasks, which could be interesting future work.

E. Limitations and Future Work

While our work shows promising results, there are still some limitations including: i) Pre-defined data augmentations such as Rotate, Contrast, Color, and Sharpness might not be sufficient to eliminate domain information and limit intra-domain connectivity. We will consider leveraging more complicated data augmentation methods related to style transfer in future work. ii) For the extreme case, where there are no shared classes between any domains, our work fails to use cross domain nearest neighbors for learning the domain-invariant feature space. One possible way to address this issue is to use generative-based methods to generate fictitious cross domain samples potentially belonging to the same class as nearest neighbors. iii) Our work can be further improved with adaptive prototypical classifier learning to achieve better performance for domain adaptation task and multi-source domain adaptation task.

F. Datasets and Implementation Details

F.1 Datasets

DomainNet [12] is a recently proposed large-scale dataset with 0.6 million images of 345 classes distributed on 6 domains, *i.e.*, *Real*, *Clipart*, *Infograph*, *Painting*, *Quickdraw* and *Sketch*. We follow the training/testing split released by [12] and follow [1] to partition the training split at a ratio of 9:1 into the training and validation splits for model selection. **PACS** [10] consists of four domains, *i.e.*, *Photo*, *Art painting*, *Cartoon* and *Sketch*, with diverse image styles. It contains seven classes and 9,991 images totally. We use the original training/validation split provided by [10].

F.2 Implementation Details.

Specifically, our strong augmentation strategy consists of 14 types of augmentations: ShearX/Y, TranslateX/Y, Rotate, AutoContrast, Identity, Equalize, Solarize, Posterize, Contrast, Color, Brightness, Sharpness. The magnitude of each augmentation is significant enough to produce as strong augmentations as possible. More details of different transformations are listed in Table S-3. Specifically, to transform an image, we randomly select 5 augmentations from the above 14 types of transformations, which creates powerful $\hat{A}$ with $\binom{14}{5}$ possible combinations, and apply them to the image sequentially.

The UDG experiments consist of three steps: 1) unsupervised training on the source domains; 2) using a small subset of labeled source domain images to train the unsupervised model (linear probing or fine-tuning); 3) testing the trained model on the target domain, which is unseen during the whole training process.

For unsupervised training, based on SimCLR [3], we adopt ResNet-18 as the backbone, and use the projection head with two MLP layers mapping the features to 128-d and with ℓ_2-norm on top. We strictly follow the protocol of existing UDG methods [7, 17], including same backbone, same number of epochs, and same subset of classes used for training and testing. We use batches of size 128, Adam optimizer with lr $3e^{-4}$ and cosine LR-schedule for 1000 epochs training. We set the temperature as $\tau = 0.07$ and warm up epoch as $T = 100$. For **DomainNet**, we train on *Painting*, *Real* and *Sketch* and test on *Clipart*, *Infograph* and *Quickdraw*, and vice versa. For **PACS**, we evaluate our method in the leave-one-domain-out way, *i.e.*, train on three domains and test on the remaining domain.

For *all correlated* setting, we evaluate with linear probing and KNN accuracy. For linear probing, we train a linear classifier with a learning rate of 30 for 30 epochs and use the source validation set for model selection. Besides, we provide KNN (K=1) accuracy for our method, where we directly use our unsupervised features without any additional training. For *domain correlated* setting, due to category shift, we evaluate the model after finetuning 30 epochs with learning rate $1e^{-3}$, and use the source validation set for model selection.

F.3 Surrogate Metrics for Connectivity.

We propose to define the Overlap Ratio (OR) metric as a surrogate measure for the degree of connectivity. Given an unlabeled dataset S_{UL} with N_{UL} samples, we randomly augment each raw image $x_i \in S_{\mathrm{UL}}$ for C times, and get an augmented set $\tilde{S}_{\mathrm{UL}} = \{x_{ij}, i \in [N_{\mathrm{UL}}], j \in [C]\}$, which is the experimental approximation to the distribution of augmentations $\mathcal{A}(\cdot|x)$. Then, for each $x_{ip} \in \tilde{S}_{\mathrm{UL}}$that is an augmented view of $x_i \in S_{\mathrm{UL}}$, denoting its k-nearest neighbors in $\tilde{S}_{\mathrm{UL}}$ in the embedding space of the encoder f as $N(x_{ip}, \tilde{S}_{\mathrm{UL}} \backslash x_{ip}, k)$, other augmented views from the same domain as $\mathcal{C}_\alpha(x_{ip}) = \{x_{jl}, d_i = d_j, l \in [C]\}$, and other augmented views from the same category as $\mathcal{C}_\beta(x_{ip}) = \{x_{jl}, y_i = y_j, l \in [C]\}$, we can define the intra-domain overlap ratio and intra-class overlap ratio as the ratio of augmented views from the same domain and category in its k-nearest neighbors, respectively.

$$\mathrm{OR}_\alpha\left(x_{ip}\right) = \frac{\#[N(x_{ip}, \tilde{S}_{\mathrm{UL}} \backslash x_{ip}, k) \cap \mathcal{C}_\alpha\left(x_{ip}\right)]}{\#N(x_{ip}, \tilde{S}_{\mathrm{UL}} \backslash x_{ip}, k)} \in [0, 1] \tag{S-26}$$

$$\mathrm{OR}_\beta\left(x_{ip}\right) = \frac{\#[N(x_{ip}, \tilde{S}_{\mathrm{UL}} \backslash x_{ip}, k) \cap \mathcal{C}_\beta\left(x_{ip}\right)]}{\#N(x_{ip}, \tilde{S}_{\mathrm{UL}} \backslash x_{ip}, k)} \in [0, 1] \tag{S-27}$$

We can define its average as Average Overlap Ratio (AOR) on the whole dataset:

$$\mathrm{AOR}_\alpha = \mathbb{E}_{x_{ip} \sim \tilde{S}_{\mathrm{UL}}} \mathrm{OR}_\alpha\left(x_{ip}\right) \tag{S-28}$$

$$\mathrm{AOR}_\beta = \mathbb{E}_{x_{ip} \sim \tilde{S}_{\mathrm{UL}}} \mathrm{OR}_\beta\left(x_{ip}\right) \tag{S-29}$$

Here AOR_α and AOR_β are surrogate metrics for intra-domain and intra-class connectivity, respectively. In specific, we use $C = 10$ augmentations for each image and take $k = 1$ by default. The encoder f is ResNet-18 trained by 10 epochs for warm-up in an unsupervised manner.

References

[1] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020. 7

[2] Liang Chen, Qianjin Du, Yihang Lou, Jianzhong He, Tao Bai, and Minghua Deng. Mutual nearest neighbor contrast and hybrid prototype self-training for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6248–6257, 2022. 7

[3] 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. 8

[4] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022. 5

[5] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2765–2773, 2017. 7

[6] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021. 6

[7] Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, et al. Unsupervised domain generalization by learning a bridge across domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5280–5290, 2022. 8

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 5, 6

[9] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9123–9132, 2021. 5, 7

[10] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5542–5550, 2017. 7

[11] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9757–9766, 2021. 7

[12] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 7

[13] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021. 3, 5

[14] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022. 6

[15] Haiyang Yang, Shixiang Tang, Meilin Chen, Yizhou Wang, Feng Zhu, Lei Bai, Rui Zhao, and Wanli Ouyang. Domain invariant masked autoencoders for self-supervised learning from multi-domains. In *European Conference on Computer Vision*, pages 151–168. Springer, 2022. 5, 6

[16] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 5, 7

[17] Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4910–4920, 2022. 8