Appendix

A. Proofs

A.1. Proof for parameterization gap

Proof A.1 *Recall the feasibility assumption 3 ensures the strong duality such that the primal and dual optimal objectives are equal, which means for* $\forall\ \tilde{\lambda} \in \mathbb{R}^+$, $\tilde{f}_s, \tilde{f}_v \in \mathcal{H}$, *the following saddle point condition holds*

$$\begin{aligned}\mathcal{R}(f_s^*, f_v^*, \tilde{\lambda}) \leq & \max_{\lambda} \min_{f_s, f_v} \mathcal{R}(f_s, f_v, \lambda) = D^* = P^* \\ & = \min_{f_s, f_v} \max_{\lambda} \mathcal{R}(f_s, f_v, \lambda) \leq \mathcal{R}(\tilde{f}_s, \tilde{f}_v, \lambda^*)\end{aligned} \tag{14}$$

Recall the definition of $\mathcal{D}_\epsilon^*$ *and by the inclusion relation* $\mathcal{H}_\theta \subseteq \mathcal{F}$, *we can derive the lower bound as*

$$\begin{aligned}\mathcal{D}_\epsilon^*(\gamma) \triangleq & \max_{\lambda} \min_{\theta, \phi \in \mathcal{H}} \mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{con}}(\theta, \phi) \\ & \geq \min_{\theta, \phi \in \mathcal{H}} \mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{con}}(\theta, \phi), \forall \lambda \in \mathbb{R}^+ \\ & \geq \min_{f_s, f_v \in \mathcal{F}} \mathcal{L}(f_s) + \lambda \mathcal{L}_{\text{con}}(f_s, f_v) = \mathcal{P}^*\end{aligned} \tag{15}$$

For the upper bound, we add and subtract $\mathcal{R} \equiv \mathcal{R}(f_s, f_v, \lambda) = \min_{f_s \in \mathcal{F}} \mathcal{L}(f_s) + \min_{f_s, f_v \in \mathcal{F}} \lambda \mathcal{L}_{con}(f_s, f_v)$ *from* $\mathcal{D}_\epsilon^\star(\gamma)$ *to get*

$$\begin{aligned}\mathcal{D}_\epsilon^\star(\gamma) &= \max_{\lambda} \min_{\theta, f_s, f_v} \mathcal{L}(\theta) + \lambda \mathcal{L}_{con}(\theta, \phi) + \mathcal{R} - \mathcal{R} \\ &= \max_{\lambda} \min_{\theta, f_s, f_v} \mathcal{R} + [\mathcal{L}(\theta) - \mathcal{L}(f_s)] + \lambda [\mathcal{L}_{con}(\theta, \phi) - \mathcal{L}_{con}(f_s, f_v)]\end{aligned} \tag{16}$$

With a slight abuse of notation, $\mathbb{E}$ *is short for* $\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}(X,Y),\tilde{\mathbf{x}}\sim\mathbb{P}(X)}$. *Then we consider the combination of the second and third term as*

$$\begin{aligned}&[1, \lambda] \begin{bmatrix} \mathcal{L}(\theta) - \mathcal{L}(f_s) \\ \mathcal{L}_{con}(\theta, \phi) - \mathcal{L}_{con}(f_s, f_v) \end{bmatrix} \\ \leq & (1 + \|\lambda\|_1) \left\| \begin{bmatrix} \mathcal{L}(\theta) - \mathcal{L}(f_s) \\ \mathcal{L}_{con}(\theta, \phi) - \mathcal{L}_{con}(f_s, f_v) \end{bmatrix} \right\|_\infty \\ = & (1 + |\lambda|) \max \{\mathcal{L}(\theta) - \mathcal{L}(f_s), \mathcal{L}_{con}(\theta, \phi) - \mathcal{L}_{con}(f_s, f_v)\} \\ = & (1 + |\lambda|) \max \{\mathbb{E}[\ell(h_s(\mathbf{x}; \theta); y) - \ell(f_s(\mathbf{x}); y)], \mathbb{E}[d(\mathbf{x}, D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi))) - d(\mathbf{x}, D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}})))]\} \\ \leq & (1 + |\lambda|) \mathbb{E} \{\max [[\ell(h_s(\mathbf{x}; \theta); y) - \ell(f_s(\mathbf{x}); y)], \mathbb{E}[d(\mathbf{x}, D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi))) - d(\mathbf{x}, D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}})))]]\} \\ \leq & (1 + |\lambda|) \mathbb{E} \{\max [L_\ell |h_s(\mathbf{x}; \theta) - f_s(\mathbf{x})|, L_d |D(h_s(\mathbf{x}; \theta), h_v(\tilde{\mathbf{x}}; \phi)) - D(f_s(\mathbf{x}), f_v(\tilde{\mathbf{x}}))|]\} \\ \leq & (1 + |\lambda|) \mathbb{E} \max \{L_\ell \epsilon_s, L_d \epsilon_g\} \\ = & (1 + |\lambda|) \max \{L_\ell \epsilon_s, L_d \epsilon_g\}\end{aligned} \tag{17}$$

where the first inequality is using Hölder's inequality (18) when $p = 1, q = \infty$ *and the second one is by the convexity of max-norm and Jensen's inequality. The third inequality is applying* L_ℓ *and* L_d *lipschitzness on* ℓ *and* d. *The fourth one is due to* ϵ_s *and* ϵ_g *parameterization on* f_s *and* D.

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^q \right)^{1/q} \tag{18}$$

Then Eq. (16) becomes

$$\mathcal{D}_\epsilon^\star(\gamma) \leq \max_{\lambda} \min_{f_x, f_v} \mathcal{R} + (1 + |\lambda|) \max\{L_\ell \epsilon_s, L_d \epsilon_g\} \triangleq D_p^* \tag{19}$$

In order to use strong duality to bound $\mathcal{D}_p^*$, we need to construct primal problem whose optimum is $\mathcal{P}_p^*$ from $\mathcal{D}_p^*$. Therefore, the remaining proof is to show $\mathcal{D}_p^*$ is the dual problem to a constraint statistical learning problem with perturbation function as $m = \max\{L_\ell \epsilon_s, L_d \epsilon_g\}$. Specifically, when $\lambda > 0$, D_p^* can be expanded and rearranged as

$$\begin{aligned}\mathcal{D}_p^* &= \max_{\lambda} \min_{f_s, f_v} \mathcal{L}(f_s) + \lambda \mathcal{L}_{\text{con}}(f_s, f_v) + (1 + |\lambda|)m \\ &= \max_{\lambda} \min_{f_s, f_v} \ell(f_s(\mathbf{x}; \theta); y) + m + \lambda[d(\mathbf{x}, D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))) - \gamma + m]\end{aligned} \tag{20}$$

By the feasibility assumption 3, Eq. (20) can be considered as the dual problem of the following one:

$$\begin{aligned}&\mathcal{P}_p^\star(\gamma) \triangleq \min_{f_s \in \mathcal{F}} \mathcal{L}(f_s) + m \\ &\text{s.t.} \quad d(\mathbf{x}, D(f_s(\mathbf{x}; \theta), f_v(\tilde{\mathbf{x}}; \phi))) \leq \gamma - m\end{aligned} \tag{21}$$

Denote f_s^* and f_v^* as the primal solutions to $\mathcal{P}_p^*$, λ^* and λ_p^* as the dual variable that give solutions $\mathcal{D}_p^*$, $\mathcal{P}_p^*$. With the regularity assumption 2, the saddle point condition shows for $\forall \tilde{f}_s, \tilde{f}_v \in \mathcal{H}$ and $\lambda, \lambda_p^* \in \mathbb{R}^+$, it holds that

$$\begin{aligned}\mathcal{R}(f_{s,p}^*, f_{v,p}^*, \tilde{\lambda}) + (1 + |\tilde{\lambda}|) &\leq \max_{\lambda} \min_{f_s, f_v} \mathcal{R}(f_s, f_v, \lambda) + (1 + |\lambda|)m = \mathcal{D}_p^* = \mathcal{P}_p^* \\ &= \min_{f_s, f_v} \max_{\lambda} \mathcal{R}(f_s, f_v, \lambda) + (1 + |\lambda|)m \\ &\leq \mathcal{R}(\tilde{f}_s, \tilde{f}_v, \lambda_p^*) + (1 + |\lambda_p^*|)m\end{aligned} \tag{22}$$

Recall that Eq. (22) holds for $\forall \tilde{f}_s, \tilde{f}_v \in \mathcal{H}$. Let $\tilde{f}_s = f_s^*, \tilde{f}_v = f_v^*$, then we use it to upper bound D_ϵ^* as

$$\mathcal{D}_\epsilon^*(\gamma) \leq \mathcal{P}_p^* \leq \mathcal{R}(f_s^*, f_v^*, \lambda^*) + (1 + |\lambda_p^*|)m = \mathcal{P}^*(\gamma) + (1 + |\lambda_p^*|)m \tag{23}$$

The last step is due to Eq. (14).

A.2. Proof for empirical gap

Proof A.2 *Similar to [16, 60], by KKT conditions and complementary slackness conditions [10] shows*

$$\begin{aligned}&\lambda_\epsilon^*(\mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim \mathbb{P}(X)}[d(\mathbf{x}, D(f_s(\mathbf{x}; \theta_\epsilon^*), f_v(\tilde{\mathbf{x}}; \phi_\epsilon^*))) - \gamma] = 0 \\ &\lambda_{\epsilon,n}^*(\sum_{i=1}^n \sum_{j \neq i}^n d(\mathbf{x_i}, D(f_s(\mathbf{x_i}; \theta_\epsilon^*), f_v(\mathbf{x_j}; \phi_\epsilon^*)) - \gamma) = 0\end{aligned} \tag{24}$$

where $(\theta_\epsilon^*, \phi_\epsilon^*, \lambda_\epsilon^*)$ and $(\theta_{\epsilon,n}^*, \phi_{\epsilon,n}^*, \lambda_{\epsilon,n}^*)$ are primal-dual pairs for achieving the optimum $\mathcal{D}_\epsilon^*(\gamma)$ and $\mathcal{D}_{\epsilon,n}^*(\gamma)$. Eq. (24) implies the constraint-related terms in the objectives to be zero, then consider the remaining term as

$$\begin{aligned}\mathcal{D}_\epsilon^*(\gamma) &= \mathbb{E}[\ell(f_s(\mathbf{x}; \theta); y)] \triangleq \mathcal{M}(\theta_\epsilon^*) \\ \mathcal{D}_{\epsilon,n}^*(\gamma) &= \sum_{i=1}^n \ell(f_s(\mathbf{x_i}), y_i) \triangleq \hat{M}(\theta_{\epsilon,n}^*)\end{aligned} \tag{25}$$

Thus the empirical gap reduces to

$$|\mathcal{D}_\epsilon^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| = |\mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*)| \tag{26}$$

Using the fact that θ_ϵ^* and $\theta_{\epsilon,n}^*$ are optimal for $\mathcal{M}(\theta_\epsilon^*)$ and $\hat{\mathcal{M}}(\theta_{\epsilon,n}^*)$, the following holds

$$\mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_\epsilon^*) \leq \mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*) \leq \mathcal{M}(\theta_{\epsilon,n}^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*) \tag{27}$$

Therefore, using the above lower and upper bound, we can bound Eq. (26) as

$$|\mathcal{D}_\epsilon^*(\gamma) - \mathcal{D}_{\epsilon,n}^*(\gamma)| \leq \max\{|\mathcal{M}(\theta_\epsilon^*) - \hat{\mathcal{M}}(\theta_\epsilon^*)|, |\mathcal{M}(\theta_{\epsilon,n}^*) - \hat{\mathcal{M}}(\theta_{\epsilon,n}^*)|\} \tag{28}$$

Then we resort to the classical VC-dimension bounds for the above two terms in Eq. (28) as follows

$$|\mathcal{M}(\theta) - \hat{\mathcal{M}}(\theta)| \leq 2B\sqrt{\frac{1}{n}\left[1 + \log\left(\frac{4(2n)^{d_{\mathrm{VC}}}}{\delta}\right)\right]} \tag{29}$$

holds with probability $1 - \delta$ *when* d_{VC} *is the VC dimension for all* θ*.*

Combing Eq. (28) and (29) completes the proof.

A.3. Proof for empirical duality gap

Proof A.3 *Simply combining the results in the above lemmas, i.e. parameterization gap and empirical gap, via applying the triangle inequality completes the proof.*

$$\begin{aligned}
\left|\mathcal{P}^\star - \mathcal{D}^\star_{\varepsilon,n}(\gamma)\right| &= \left|\mathcal{P}^\star + \mathcal{D}^\star_\epsilon(\gamma) - \mathcal{D}^\star_\epsilon(\gamma) - \mathcal{D}^\star_{\epsilon,n}(\gamma)\right| \\
&\leq |\mathcal{P}^\star - \mathcal{D}^\star_\varepsilon(\gamma)| + |\mathcal{D}^\star_\epsilon(\gamma) - \mathcal{D}^\star_{\epsilon,n}(\gamma)| \\
&\leq (1 + |\lambda|)m + 2B\sqrt{\frac{1}{n}\left[1 + \log\left(\frac{4(2n)^{d_{vc}}}{\delta}\right)\right]}
\end{aligned}$$

B. Domain Generalization by Learning on Fictitious Distributions

This section gives a justification for disentanglement from a different perspective by connecting the dots with classical domain adaptation. Specifically, we construct a fictitious distribution to extend it to the DG setting and decompose the target learning objective into empirical learning errors, domain divergence and source domain data diversity. Moreover, we show that learning disentangled representations gives a tighter risk upper bound.

With a slight abuse of notation, let $\mathcal{H}$ be a hypothesis space and denote $\tilde{\mathcal{D}}$ as the induced distribution over feature space $\mathcal{Z}$ for every distribution $\mathcal{D}$ over the raw space. Define $\mathcal{D}_S^i$ as the source distribution over $\mathcal{X}$, which enables a mixture construction of source domains as $\mathcal{D}_S^\alpha = \sum_{i=1}^{N_s} \alpha_i \mathcal{D}_S^i(\cdot)$. Denote a fictitious distribution $\mathcal{D}_U^\alpha = \sum_{i=1}^{N_s} \alpha_i^* \mathcal{D}_S^i(\cdot)$ as the convex combination of source domains which is the closest to $\mathcal{D}_U$, where $\alpha_1^*, ..., \alpha_{N_S}^* = \arg\min_{\alpha_1,...,\alpha_{N_s}} d_{\mathcal{H}}(\mathcal{D}_U, \sum_{i=1}^{N_s} \alpha_i \mathcal{D}_S^i(\cdot))$. The fictitious distribution induces a feature space distribution $\tilde{\mathcal{D}}_U^\alpha = \sum_{i=1}^{N_s} \alpha_i^* \tilde{\mathcal{D}}_S^i(\cdot)$. The following inequality holds for the risk $\epsilon_U(h)$ on any unseen target domain $\mathcal{D}_U$.

$$\epsilon_U(h) \leq \lambda_\alpha + \underbrace{\sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h)}_{\text{① Empirical}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha)}_{\text{② Divergence}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha)}_{\text{③ Diversity}} \tag{30}$$

where λ_α is the risk of the optimal hypothesis on the mixture source domain $\mathcal{D}_S^\alpha$ and $\mathcal{D}_U$.

We define the symmetric difference hypothesis space as $\mathcal{H}\Delta\mathcal{H} = \{h(\mathbf{x}) \oplus h'(\mathbf{x}) : h, h' \in \mathcal{H}\}$, where $\oplus$ is the XOR operator. Applying [6], we have

$$\begin{aligned}
\epsilon_U(h) &\leq \lambda_U + \Pr_{\mathcal{D}_U}[\mathcal{Z}_h \triangle \mathcal{Z}_h^*] \\
&\leq \lambda_U + \Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \triangle \mathcal{Z}_h^*] + |\Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \triangle \mathcal{Z}_h^*] - \Pr_{\mathcal{D}_U}[\mathcal{Z}_h \triangle \mathcal{Z}_h^*]| \\
&\leq \lambda_U + \Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \triangle \mathcal{Z}_h^*] + d_{\mathcal{H}}(\tilde{D}_U, \tilde{\mathcal{D}}_S^\alpha) \\
&\leq \lambda_U + \Pr_{\mathcal{D}_S^\alpha}[\mathcal{Z}_h \triangle \mathcal{Z}_h^*] + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha) \\
&\leq \lambda_\alpha + \underbrace{\sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h)}_{\text{① Empirical}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U^\alpha, \tilde{\mathcal{D}}_S^\alpha)}_{\text{② Divergence}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}_U^\alpha)}_{\text{③ Diversity}}
\end{aligned} \tag{31}$$

The fourth inequality holds because of the triangle inequality. We provide the explanation for our bound in the Eq. (30) and Eq. (31). The second term is the empirical loss for the convex combination of all source domains. The third term corresponds to "To what extent can the convex combination of the source domain approximate the target domain". The minimization of the third term requires diverse data or strong data augmentation, such that the unseen distribution lies within the convex combination of source domains. For the fourth term,the following equation holds for any two distributions D_U', D_U'', which are the convex combinations of source domains [12]

$$d_{\mathcal{H}[D_U', D_U'']} \leq \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \alpha_l \alpha_k d_{\mathcal{H}}[\mathcal{D}_{S,l}, \mathcal{D}_{S,k}] \tag{32}$$

Such an upper bound will be minimized when $d_{\mathcal{H}}[\mathcal{D}_{S,l}, \mathcal{D}_{S,k}] = 0, \forall\, l, k \in \{1, ..., N_S\}$. Namely projecting the source domain data into a feature space, where the source domain labels are hard to distinguish.

The above ③ Diversity term is also supported by the evidence that compositional generalization and extrapolation can be improved if the training domain data are rich enough [13, 32, 54]. To this end, one can obviously simulate data points with predetermined data augmentation methods such as rotating, cropping, Gaussian blur, color jitter, etc. However, their developments require prior knowledge and domain-specific expertise like translation-invariance on images, which is likely to fail in the unseen domain due to distribution shifts. It motivates learning disentangled representations that are transferable across various domains [21]. Thus we discuss the benefits of disentanglement on the domain generalization gap in the following section. Assume that the semantic and the variation factors are disentangled in the latent space $\mathcal{S}$ and $\mathcal{V}$, then the errors [64] on the disentangled source and target domain with a hypothesis h are

$$\epsilon_{S,i}(h) = \epsilon_{S,i}^s(h) + \epsilon_{S,i}^v(h), \epsilon_U(h) = \epsilon_U^s(h) + \epsilon_U^v(h) \tag{33}$$

Given $h^* = \arg\min_{h\in\mathcal{H}} \left(\epsilon^s_{S,i}(h), \epsilon^v_{S,i}(h)\right)\ \forall i \in \{1,\dots,N_S\}$, since $\epsilon_U(h) = \epsilon^s_U(h) + \epsilon^v_U(h)$, combining Eq. (30) and we have

$$\epsilon^s_U(h) + \epsilon^v_U(h) \leq \lambda_\alpha + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h) + d_{\mathcal{H}}(\tilde{\mathcal{D}}^\alpha_U, \tilde{\mathcal{D}}^\alpha_S) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}^\alpha_U) \tag{34}$$

where $\lambda_\alpha = \epsilon_U(h^*) + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h^*) = \epsilon^s_U(h^*) + \epsilon^v_U(h^*) + \sum_{i=1}^{N_S} \alpha_i \epsilon^s_{S,i}(h^*) + \sum_{i=1}^{N_S} \alpha_i \epsilon^v_{S,i}(h^*)$. Then the upper bound for the unseen domain can be further derived as follow,

$$\epsilon^s_U(h) \leq \lambda_\alpha + \sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h) + d_{\mathcal{H}}(\tilde{\mathcal{D}}^\alpha_U, \tilde{\mathcal{D}}^\alpha_S) + d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}^\alpha_U) - \epsilon^v_U(h) \tag{35}$$

Combining Eq. (30) and Eq. (33), we have

$$\epsilon^s_U(h) \leq \lambda_\alpha + \underbrace{\sum_{i=1}^{N_S} \alpha_i \epsilon_{S,i}(h)}_{\text{① Empirical}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}^\alpha_U, \tilde{\mathcal{D}}^\alpha_S)}_{\text{② Divergence}} + \underbrace{d_{\mathcal{H}}(\tilde{\mathcal{D}}_U, \tilde{\mathcal{D}}^\alpha_U)}_{\text{③ Diversity}} - \underbrace{\epsilon^v_U(h)}_{\text{④}} \tag{36}$$

where ① denotes the empirical loss over every source domain, ② means divergence minimization among source domains and ③ encourages the diversity and coverage of the mixture of source domains. The above result shows the benefits of disentanglement over representation spaces. Thus we propose to use two separate encoders representing semantic and variation subspaces respectively in the following section. Such a formulation combined with the disentanglement term ④ in Eq. (36) implies that we should perform ERM over the semantic space only. The above analysis essentially justifies the design of DDG from a classical domain adaptation perspective: optimizing empirical risk over semantic space while promoting diversity and divergence by taking disentanglement as a constraint.

C. Experimental Settings

C.1. Other Training Details

We optimize all models using Adam [40]. For all the detailed hyperparameter settings, please refer to our code which is publicly available at https://github.com/hlzhang109/DDG.

C.2. Dataset Statistics and Visualization

We show some images of the datasets in Fig. 7 to give an intuitive comparison among these image datasets. One can observe that these images have a diverse set of styles, making it very challenging to transfer knowledge from one to another.

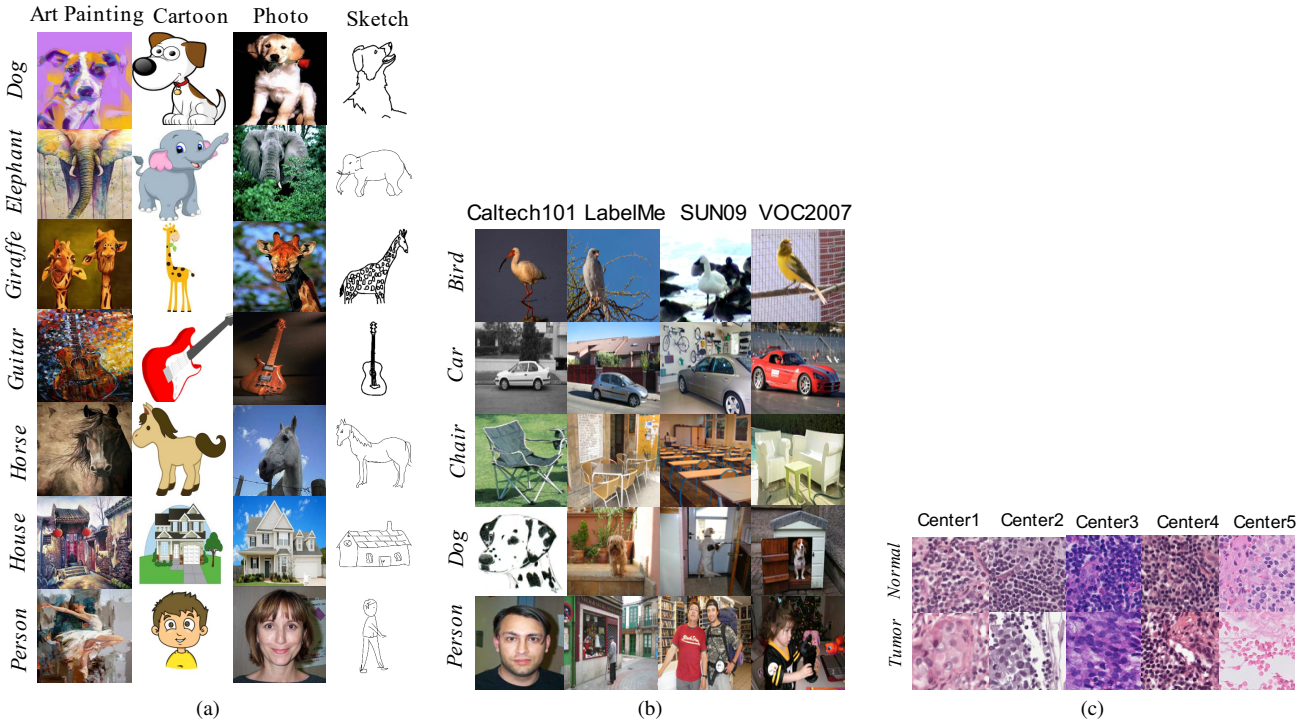


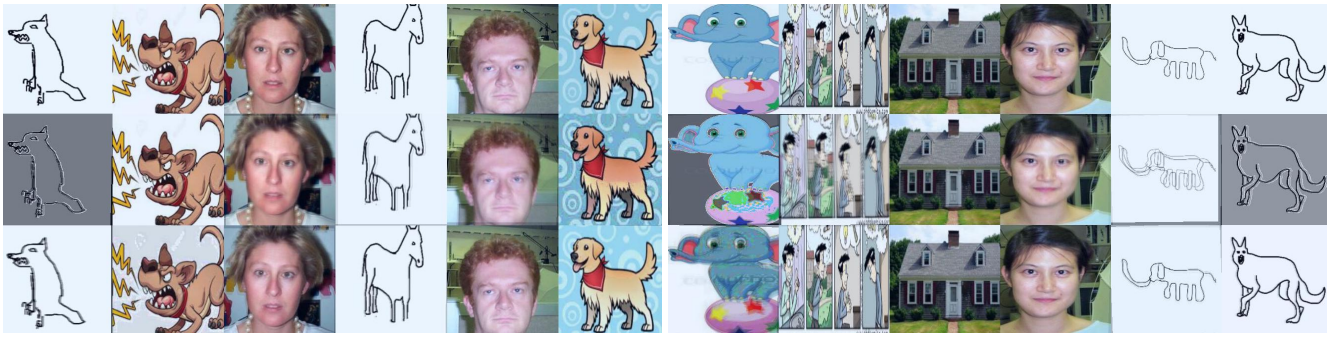
Figure 7. **Samples of DG datasets.** The training data (a) PACS and (b) VLCS (c) Wilds are shown. PACS has four domains *art* (A), *cartoons* (C), *photos* (P), *sketches* (S). VLCS contains four domains *Caltech101* (C), *LabelMe* (L), *SUN09* (S), *VOC2007* (V). The WILDS dataset includes data from five different medical centers as domains.

D. Additional Experimental Results

Qualitative comparison with AugMix. Looking into the Fig. 8, it is harder for the heuristic-based method AugMix to generate diverse samples via interpolation for training compared with DDG as shown in Fig. 3 and Fig. 10b.

More qualitative results via interpolation. Fig. 9 showcases the results of combining the semantic code of one image and the mixture of two variation codes. Results show that the model can generate samples with intermediate variation states.

More qualitative results via swapping variation and semantic factors. We showcase the qualitative results of swapping variation and semantic factors with PACS in Fig. 10b, MNIST in Fig. 10a, and WILDS in Fig. 10c. The results demonstrates the strong disentangled capability of DDG. Some interesting observations are DDG learns both intra- (e.g. thickness) and inter-domain (e.g. rotated angle) variations over RotatedMNIST. DDG also maintains semantic information like the color of distinct features across variation-rich data like PACS.

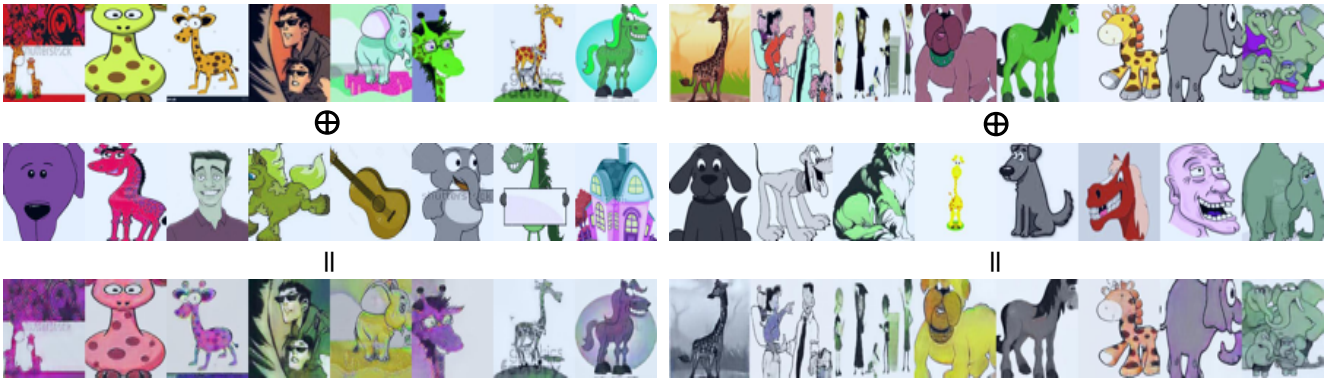


(a) (b)

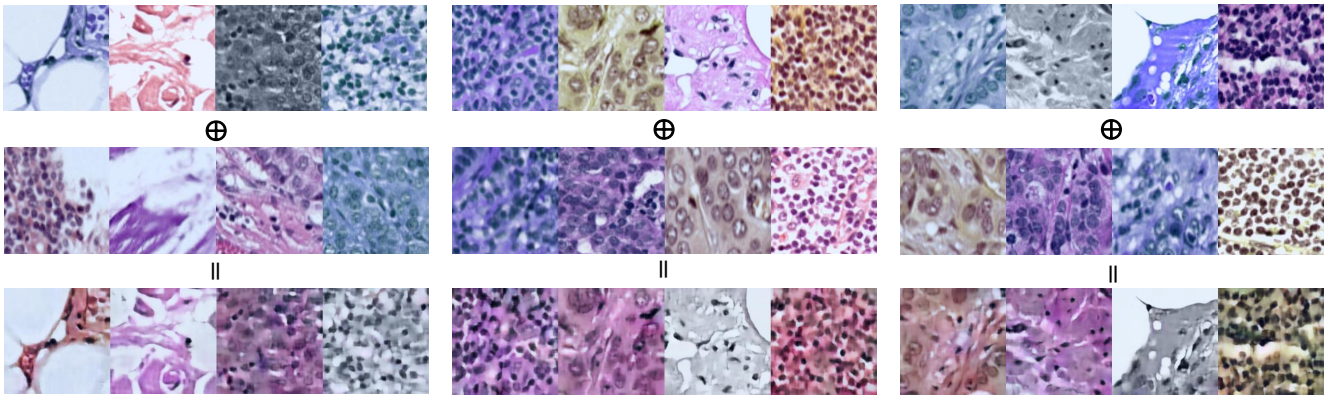
Figure 8. **The augmented samples from AugMix [30].** The second and third rows are generated by applying AugMix to the first row.

Figure 9. **Interpolation via mixing results on PACS.**

(a) Rotated MNIST



(b) PACS



(c) Wilds

Figure 10. **Qualitative disentanglement results on RotatedMNIST, PACS and Wilds.** In every panel, the training data in the first row manifests the semantic factors.

Numerical comparison with MBDG. We quantatively compare DDG with MBDG [60] as in Table 2. For a consistent comparison, we run the source code of authors under a test domain validation protocal [27] on PACS with the results as:

	A	C	P	S	Avg
MBDG	82.0 ± 0.0	78.4 ± 0.01	93.9 ± 0.0	85.0 ± 0.0	85.8
MBDG$_{Reg}$	84.9 ± 0.0	84.9 ± 0.0	93.9 ± 0.0	$\mathbf{85.6 \pm 0.0}$	87.3
DDG	$\mathbf{88.9 \pm 0.6}$	$\mathbf{85.0 \pm 1.9}$	$\mathbf{97.2 \pm 1.2}$	84.3 ± 0.7	**88.9**

Table 2. Numerical results for comparing DDG and MBDG.

Though adopting similar PACCL frameworks and primal-dual algorithms, DDG can consistently outperform MBDG and its variant except over domain S. The primary reason behind the clear performance gain of DDG can be that DDG is better at capturing variations within data via random sampling without relying on domain labels. Specifically, parameterizing h_s, h_v, D

and constrain them based on disentanglement makes the model more robust to both inter- and intra-domain nuisance factors compared to MBDG that only use a pretrained generator to simulate inter-domain variations. Moreover, the performance gain is also partly due to the three major differences between our approach and MBDG we highlight in the Related Work: First, our upper bound of the parameterization gap is tighter under mild conditions, whereas MBDG requires unrealistic assumptions on the distance metric, i.e., $d(\cdot, \cdot)$ satisfies Lipschitz-like inequality on both arguments, which is stronger than our normal L_d Lipschitzness assumption; second, MBDG consumes additional domain labels, which are prohibitively expensive or even infeasible to obtain in safety-critical applications or those containing sensitive demographics; third, DDG enforces invariance constraints via parameterizing semantic and variation encoders, which does not belong to a model-based approach. In contrast, MBDG requires a pre-trained domain transformation model (e.g., CycleGAN) during training, which may result in sub-optimal solutions and parameter inefficiency, while DDG is more flexible by treating this as a design choice.