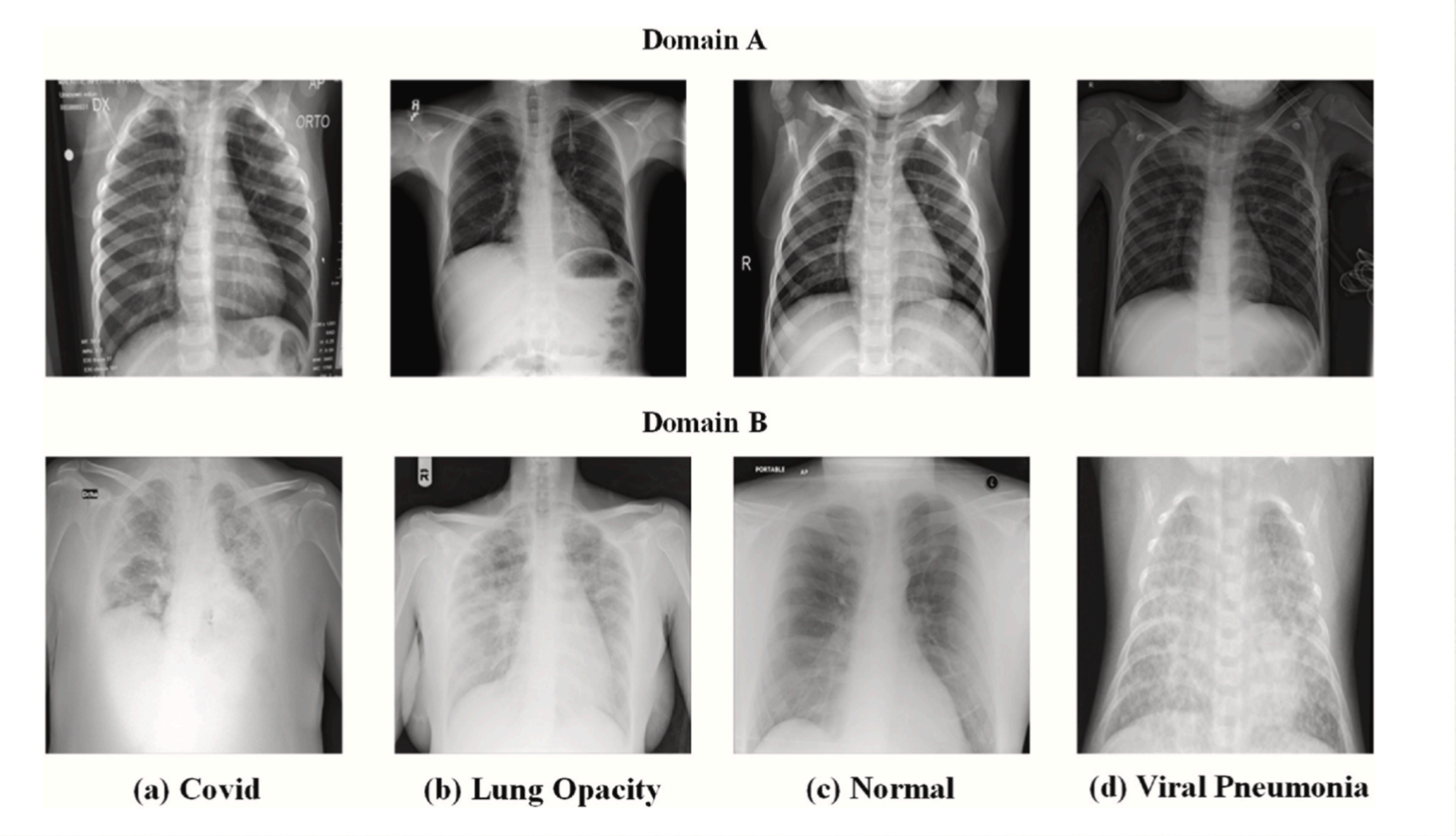
Detection of Covid-19: A Case Study of Domain Generalization

Related Reference:

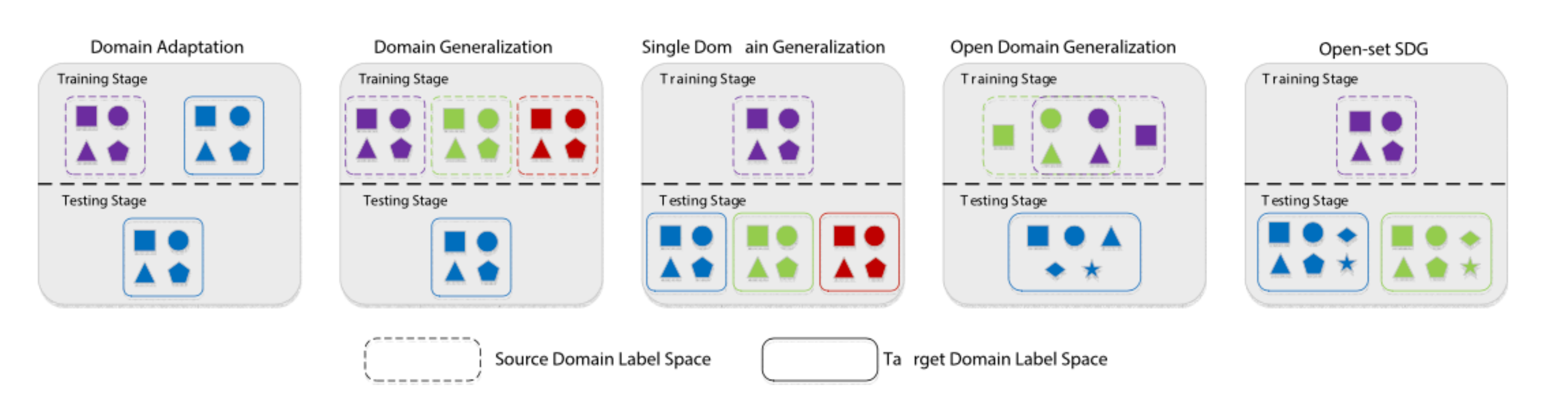
- Volpi R, Namkoong H, Sener O, et al. Generalizing to unseen domains via adversarial data augmentation[J]. Advances in neural information processing systems, 2018, 31.
- Zheng K, Wu J, Yuan Y, et al. From single to multiple: Generalized detection of Covid-19 under limited classes samples[J]. Computers in Biology and Medicine, 2023: 107298.

Background

The issue of **domain shift**,
where **the distribution of samples in the testing and training sets differ**, arises in practical applications

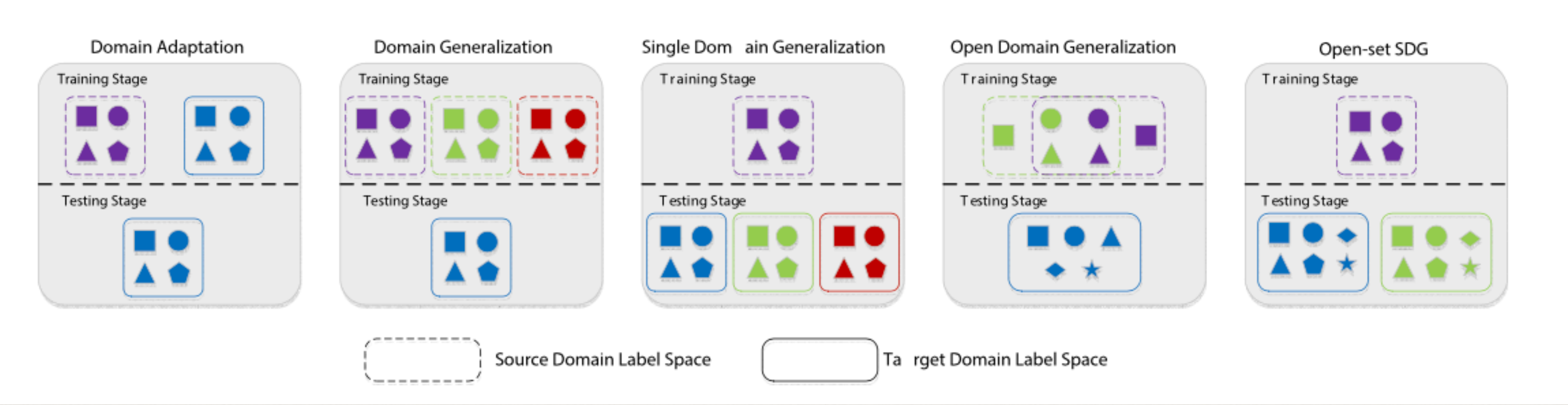


Different Domain Shift



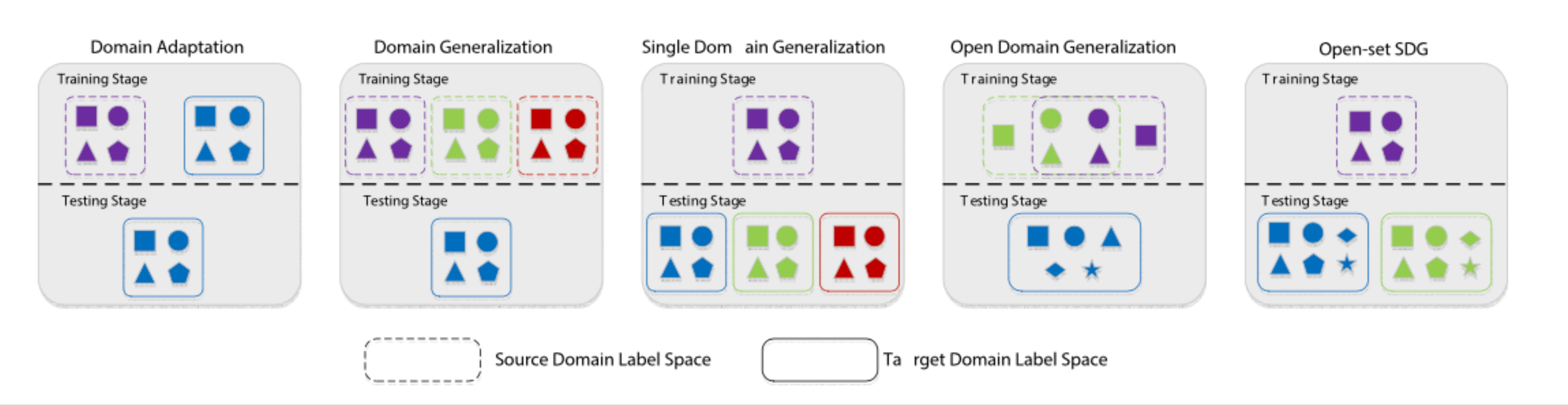
- Domain Adaptation: Have some samples from the target domain during the training phase

Different Domain Shift



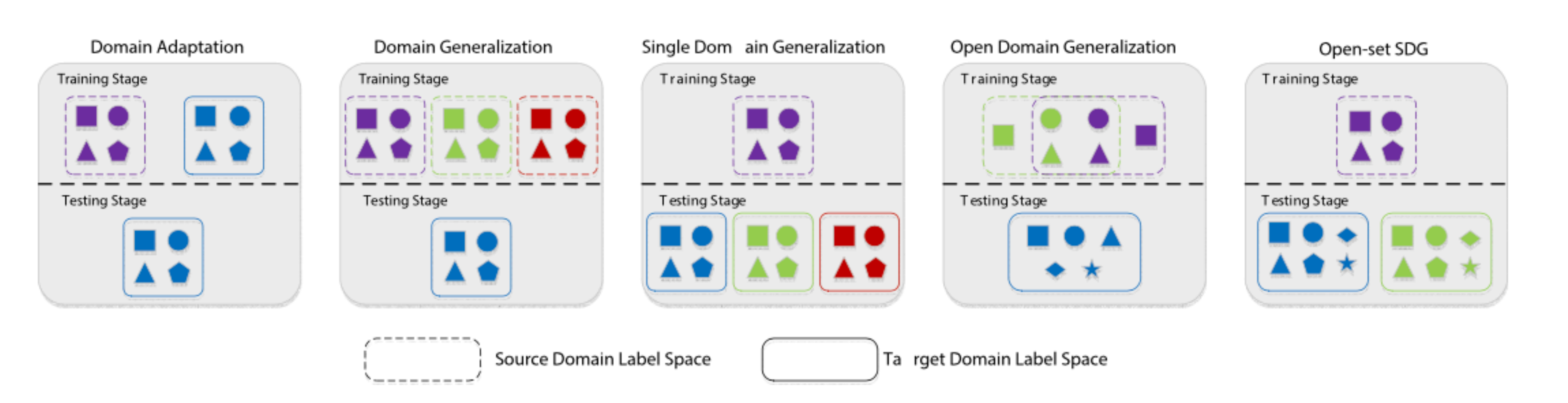
- Domain Generalization: Target domain's data is **inaccessible** during the training phase

Different Domain Shift



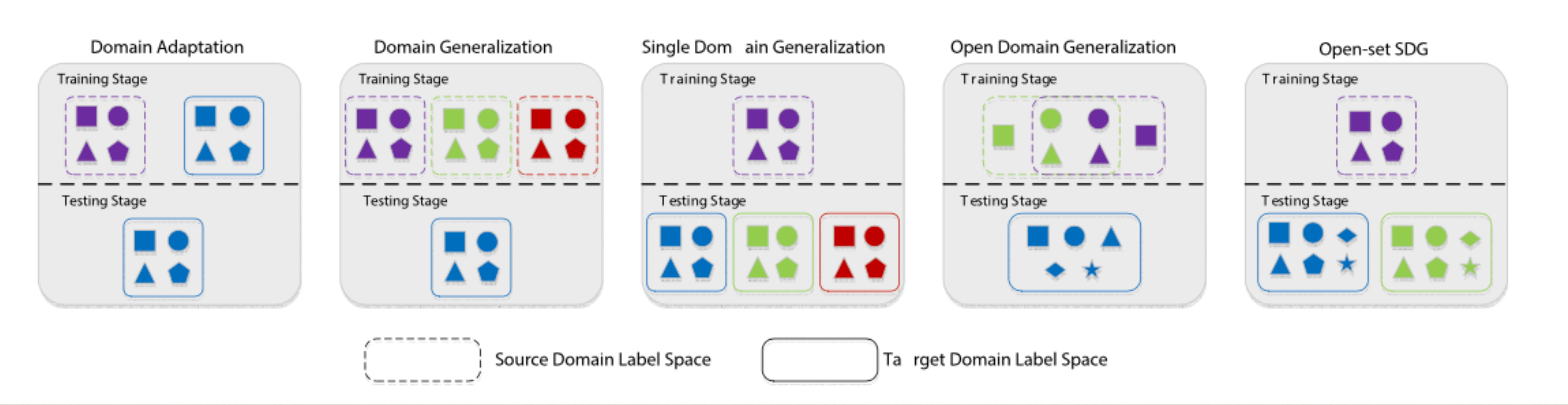
- Single Domain Generalization: learn as reliable features as possible from a **single source domain**

Different Domain Shift



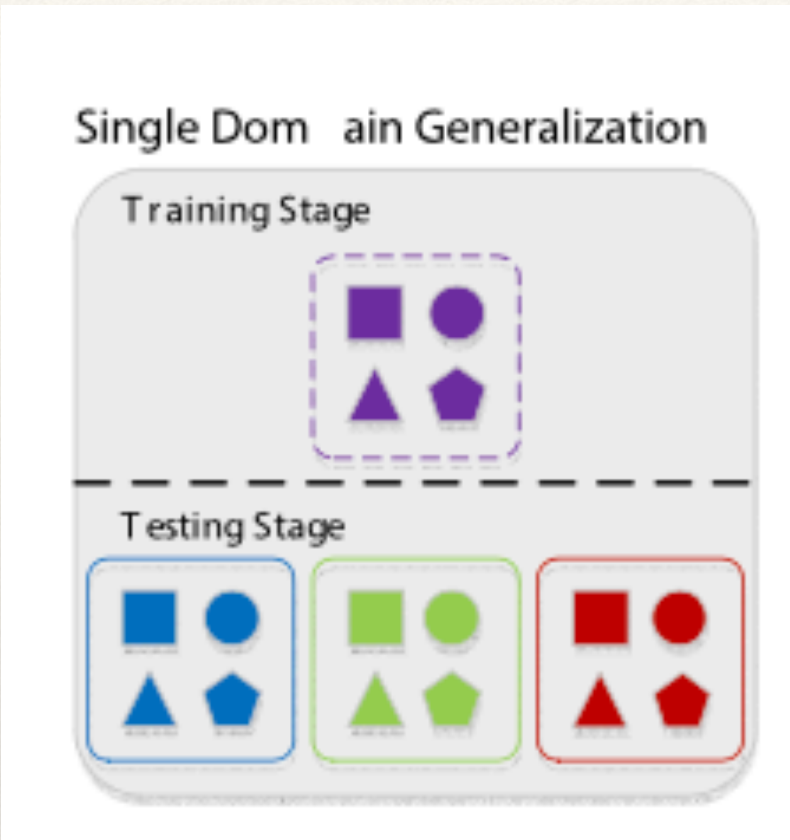
Open Domain Generalization: label space of the source domain is actually a proper subset of the target domain.

Different Domain Shift



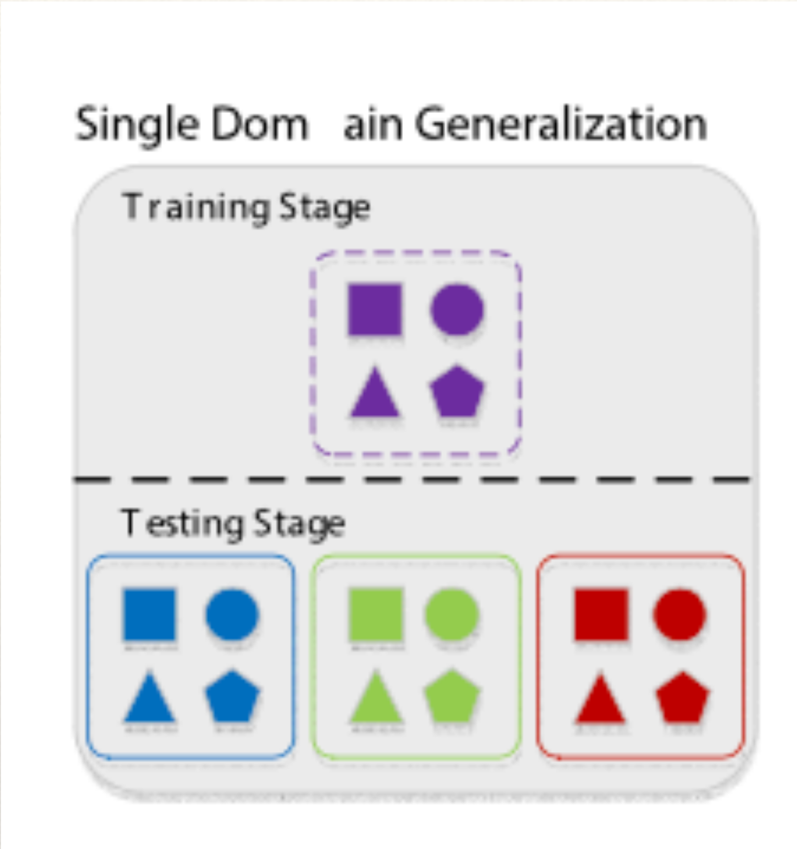
- Open-set Single Domain Generalization: Single Domain + limited label space

Single Domain Generalization



$$\min_{\theta} \sup_{\mathcal{D}_{\text{Target}}} \left\{ \mathbb{E}[\mathcal{L}(\theta; \mathcal{D}_{\text{Target}})] : D(\mathcal{D}_{\text{Target}}, \mathcal{D}_{\text{Source}}) \leq \rho \right\}$$

Single Domain Generalization



$$\min_{\theta} \sup_{\mathcal{D}_{\text{Target}}} \left\{ \mathbb{E}[\mathcal{L}(\theta; \mathcal{D}_{\text{Target}})] - \gamma D(\mathcal{D}_{\text{Target}}, \mathcal{D}_{\text{Source}}) \right\}$$

Wasserstein distance on the semantic space On the space $\mathbb{R}^p \times \mathcal{Y}$, consider the following transportation cost c—cost of moving mass from (z, y) to (z', y')

$$c((z, y), (z', y')) := \frac{1}{2} \|z - z'\|_2^2 + \infty \cdot \mathbf{1}\{y \neq y'\}.$$

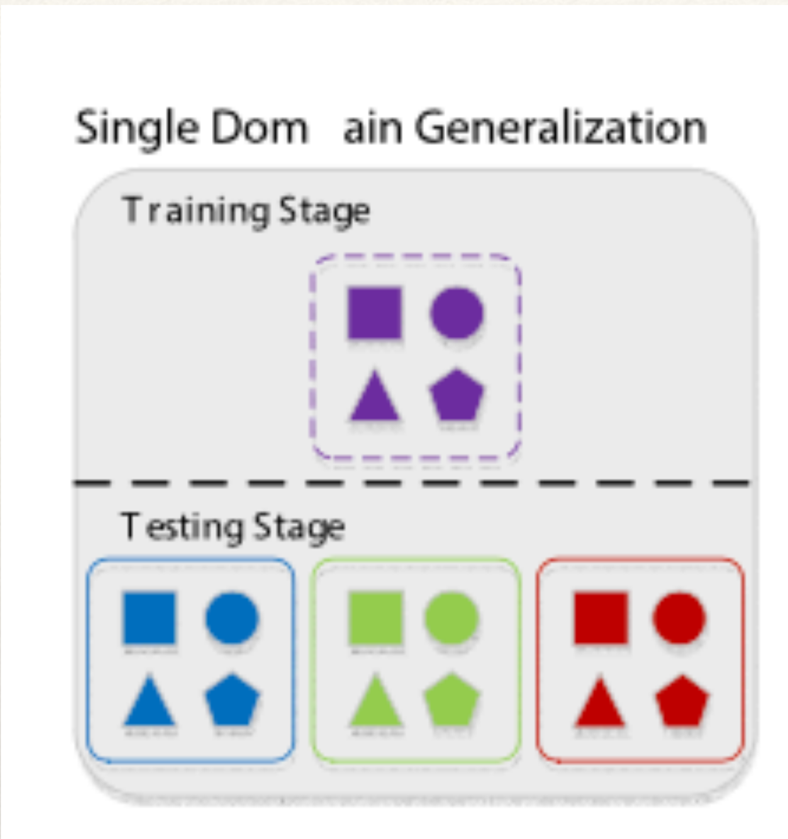
The transportation cost takes value ∞ for data points with different labels, since we are only interested in perturbation to the marginal distribution of Z. We now define our notion of distance on the semantic space. For inputs coming from the original space $\mathcal{X} \times \mathcal{Y}$, we consider the transportation cost c_θ defined with respect to the output of the last hidden layer

$$c_\theta((x, y), (x', y')) := c((g(\theta_f; x), y), (g(\theta_f; x'), y'))$$

so that c_θ measures distance with respect to the feature mapping $g(\theta_f; x)$. For probability measures P and Q both supported on $\mathcal{X} \times \mathcal{Y}$, let $\Pi(P, Q)$ denote their couplings, meaning measures M with $M(A, \mathcal{X} \times \mathcal{Y}) = P(A)$ and $M(\mathcal{X} \times \mathcal{Y}, A) = Q(A)$. Then, we define our notion of distance by

$$D_\theta(P, Q) := \inf_{M \in \Pi(P,Q)} \mathbb{E}_M[c_\theta((X, Y), (X', Y'))]. \tag{3}$$

Optimization Procedure



$$\min_{\theta} \sup_{\mathcal{D}_{\text{Target}}} \left\{ \mathbb{E}[\mathcal{L}(\theta; \mathcal{D}_{\text{Target}})] - \gamma D(\mathcal{D}_{\text{Target}}, \mathcal{D}_{\text{Source}}) \right\}$$

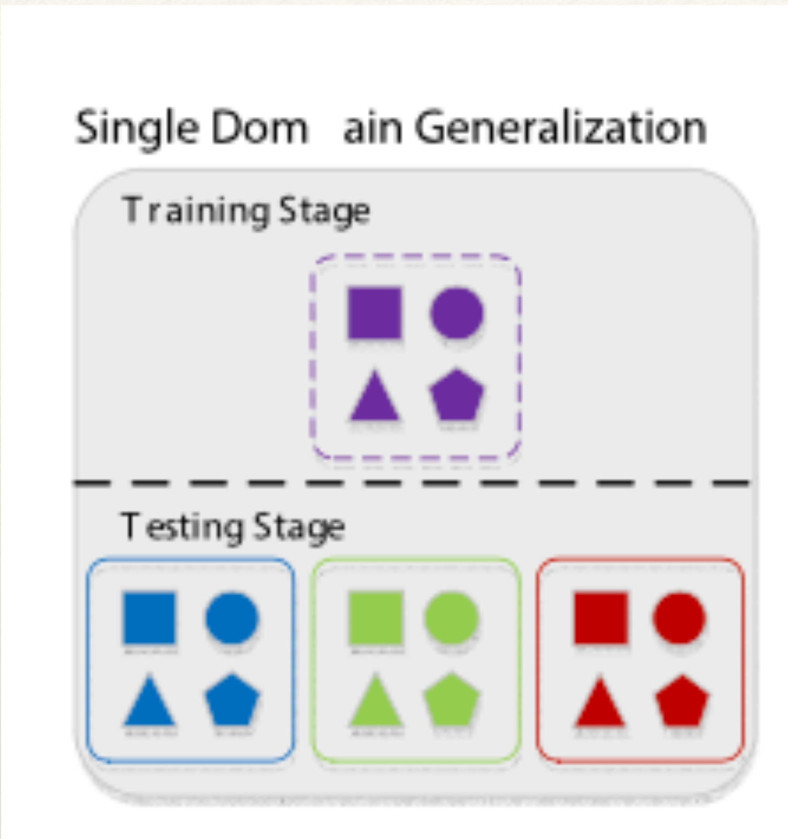
Taking the dual reformulation of the penalty relaxation (4), we can obtain an efficient solution procedure. The following result is a minor adaptation of [2, Theorem 1]; to ease notation, let us define the robust surrogate loss

$$\phi_\gamma(\theta; (x_0, y_0)) := \sup_{x \in \mathcal{X}} \{\ell(\theta; (x, y_0)) - \gamma c_\theta((x, y_0), (x_0, y_0))\}. \tag{5}$$

Lemma 1. *Let* $\ell : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ *be continuous. For any distribution* Q *and any* $\gamma \geq 0$*, we have*

$$\sup_{P} \{\mathbb{E}_P[\ell(\theta; (X, Y))] - \gamma D_\theta(P, Q)\} = \mathbb{E}_Q[\phi_\gamma(\theta; (X, Y))]. \tag{6}$$

Theoretical Motivation



$$\min_{\theta} \sup_{\mathcal{D}_{\text{Target}}} \left\{ \mathbb{E}[\mathcal{L}(\theta; \mathcal{D}_{\text{Target}})] - \gamma D(\mathcal{D}_{\text{Target}}, \mathcal{D}_{\text{Source}}) \right\}$$

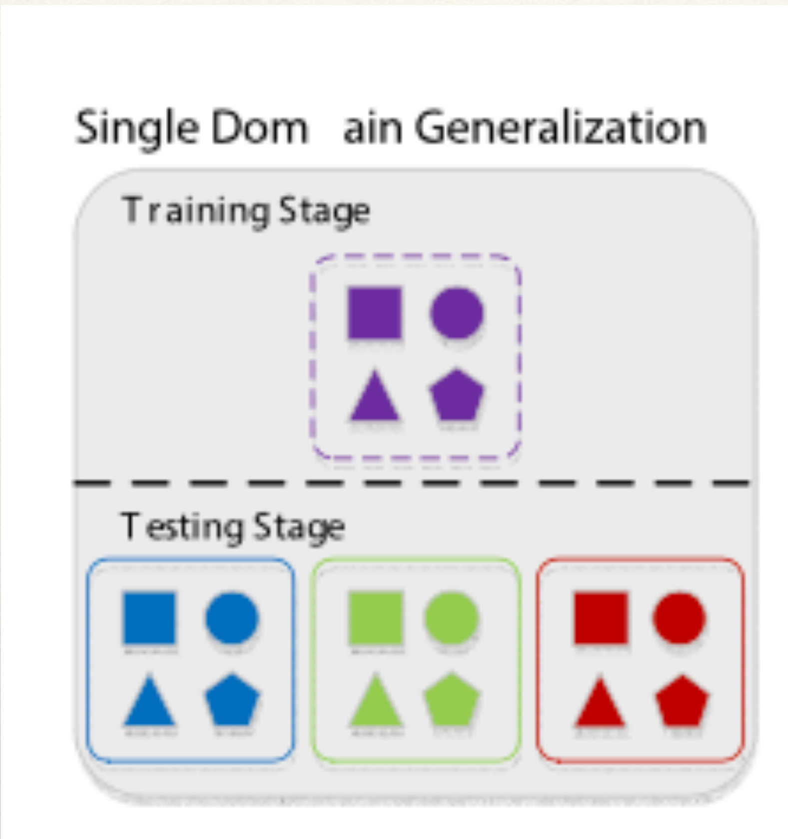
We now give an interpretation for the augmented data points in the maximization phase (8). Concretely, we fix $\theta \in \Theta$, $x_0 \in \mathcal{X}$, $y_0 \in \mathcal{Y}$, and consider an ϵ-maximizer

$$x_\epsilon^\star \in \epsilon\text{-}\arg\max_{x \in \mathcal{X}} \left\{ \ell(\theta; (x, y_0)) - \gamma c_\theta((x, y_0), (x_0, y_0)) \right\}.$$

We let $z_0 := g(\theta_f; x_0) \in \mathbb{R}^p$, and abuse notation by using $\ell(\theta; (z_0, y_0)) := \ell(\theta; (x_0, y_0))$. In what follows, we show that the feature mapping $g(\theta_f; x_\epsilon^\star)$ satisfies

$$g(\theta_f; x_\epsilon^\star) = \underbrace{g(\theta_f; x_0) + \frac{1}{\gamma}\left(I - \frac{1}{\gamma}\nabla_{zz}\ell(\theta; (z_0, y_0))\right)^{-1} \nabla_z \ell(\theta; (z_0, y_0))}_{=: \widehat{g}_{\text{newton}}(\theta_f; x_0)} + O\left(\sqrt{\frac{\epsilon}{\gamma}} + \frac{1}{\gamma^2}\right). \tag{10}$$

Theoretical Motivation



$$\min_{\theta} \sup_{\mathcal{D}_{\text{Target}}} \left\{ \mathbb{E}[\mathcal{L}(\theta; \mathcal{D}_{\text{Target}})] - \gamma D(\mathcal{D}_{\text{Target}}, \mathcal{D}_{\text{Source}}) \right\}$$

For classification problems, we show that the robust surrogate loss (5) corresponds to a particular data-dependent regularization scheme. Let $\ell(\theta; (x, y))$ be the m-class softmax loss (2) given by

$$\ell(\theta; (x, y)) = -\log p_y(\theta, x) \;\; \text{where} \;\; p_j(\theta, x) := \frac{\exp(\theta_{c,j}^\top g(\theta, x))}{\sum_{l=1}^m \exp(\theta_{c,l}^\top g(\theta_f; x))}.$$

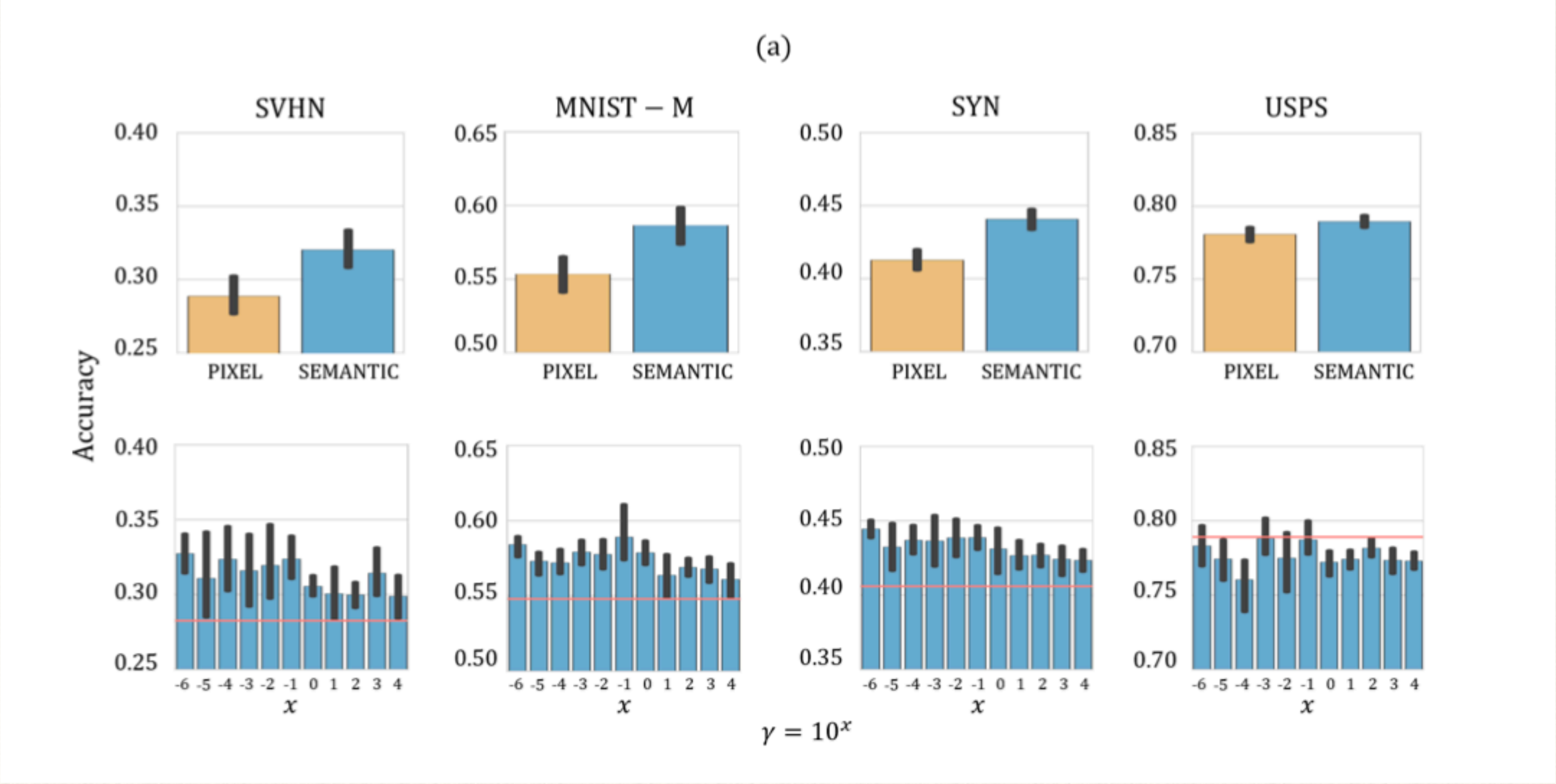
where $\theta_{c,j} \in \mathbb{R}^p$ is the j-th row of the classification layer weight $\theta_c \in \mathbb{R}^{p \times m}$. Then, the robust surrogate ϕ_γ is an approximate regularizer on the classification layer weights θ_c

$$\phi_\gamma(\theta; (x, y)) = \ell(\theta; (x, y)) + \frac{1}{\gamma} \left\| \theta_{c,y} - \sum_{j=1}^m p_j(\theta, x) \theta_{c,j} \right\|_2^2 + O\left(\frac{1}{\gamma^2}\right). \tag{11}$$

The expansion (11) shows that the robust surrogate (5) is roughly equivalent to data-dependent regularization where we minimize the distance between $\sum_{j=1}^m p_j(\theta, x)\theta_{c,j}$, our "average estimated linear classifier", to $\theta_{c,y}$, the linear classifier corresponding to the true label y. Concretely, for any fixed

Numerical Study

- Train on MNIST dataset
- Test on MNIST-M, SVHN, SYN, and USPS.



Numerical Study

Semantic scene segmentation We use the SYTHIA[31] dataset for semantic segmentation. The dataset contains images from different locations (we use *Highway*, *New York-like City* and *Old European Town*), and different weather/time/date conditions (we use *Dawn*, *Fog*, *Night*, *Spring* and *Winter*. We train models on a source domain and test on other domains, using the standard mean Intersection Over Union *(mIoU)* metric to evaluate our performance [8]. We arbitrarily chose images

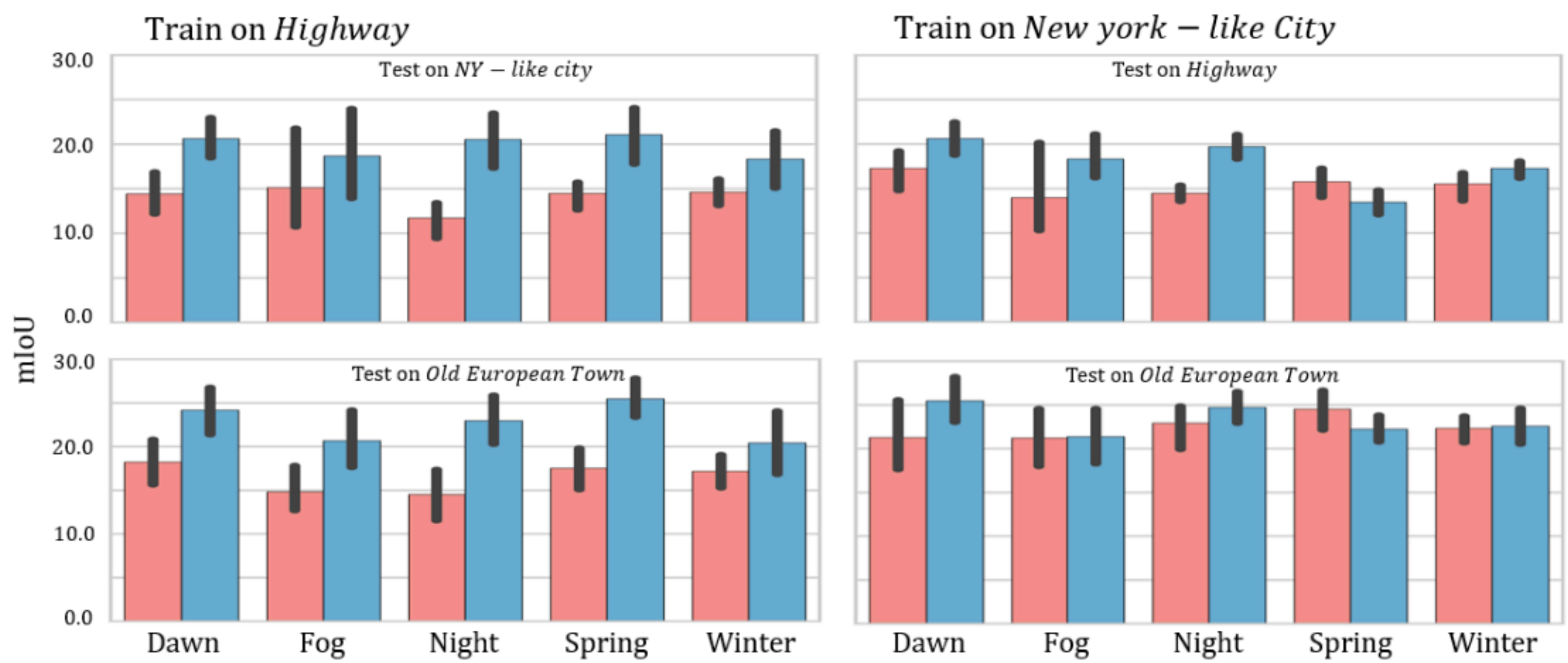


Figure 2. Results obtained with semantic segmentation models trained with ERM *(red)* and our method with $K = 1$ and $\gamma = 1.0$ *(blue)*. *Leftmost* panels are associated with models trained on *Highway*, *rightmost* panels are associated with models trained on *New York-like City*. Test datasets are *Highway*, *New York-like City* and *Old European Town*.

Numerical Study