Generalizing to Unseen Domains: A Survey on Domain Generalization

Jindong Wang[1], Cuiling Lan[1], Chang Liu[1], Yidong Ouyang[2], Wenjun Zeng[1], Tao Qin[1]

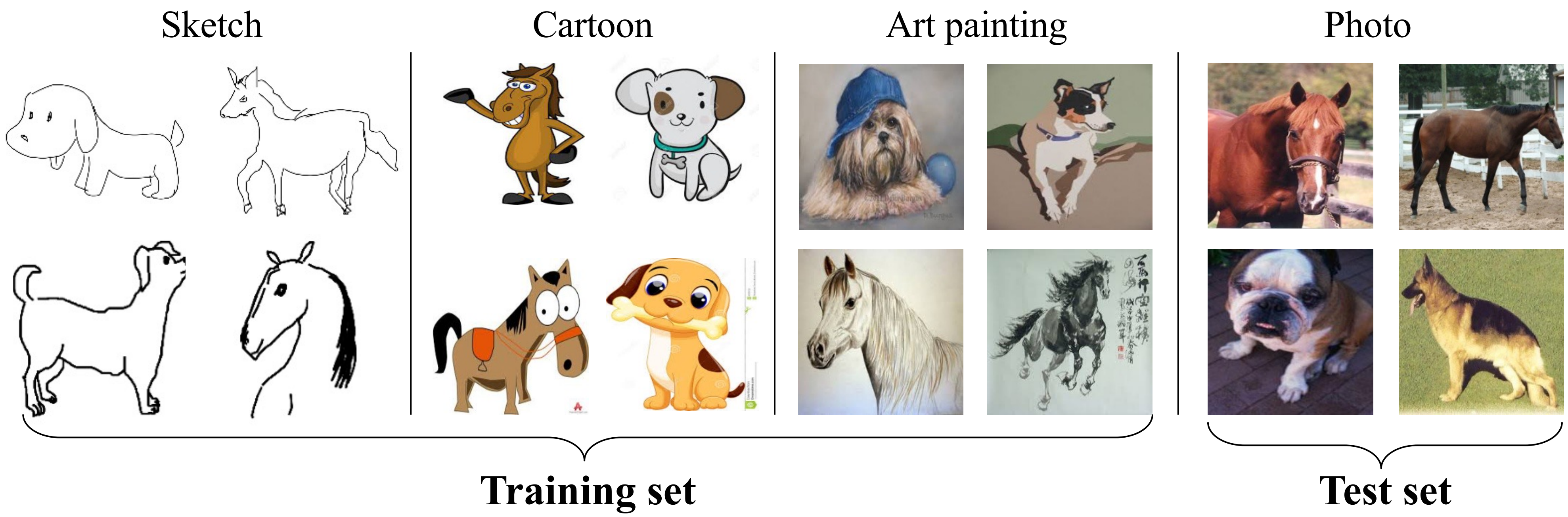
[1] Microsoft Research Asia, Beijing, China
[2] Central University of Finance and Economics, Beijing, China

jindong.wang@microsoft.com

https://arxiv.org/abs/2103.03097 https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG

Background

- Domain adaptation



- $P_{Train}(x, y) \neq P_{Test}(x, y)$
- Source: Train, Target: test

Images are copied from PACS dataset [Li et al.'17]

Domain adaptation

- Basic theory of DA [Ben-David et al'07]

$$\underbrace{\epsilon^t(h)}_{\text{Target risk}} \leq \underbrace{\epsilon^s(h)}_{\text{Source risk}} + \underbrace{d_{\mathcal{H}\Delta\mathcal{H}}\left(P_X^s, P_X^t\right)}_{\text{Source-target distribution divergence}} + \underbrace{\lambda_{\mathcal{H}}}_{\text{Complexity of } \mathcal{H}}$$

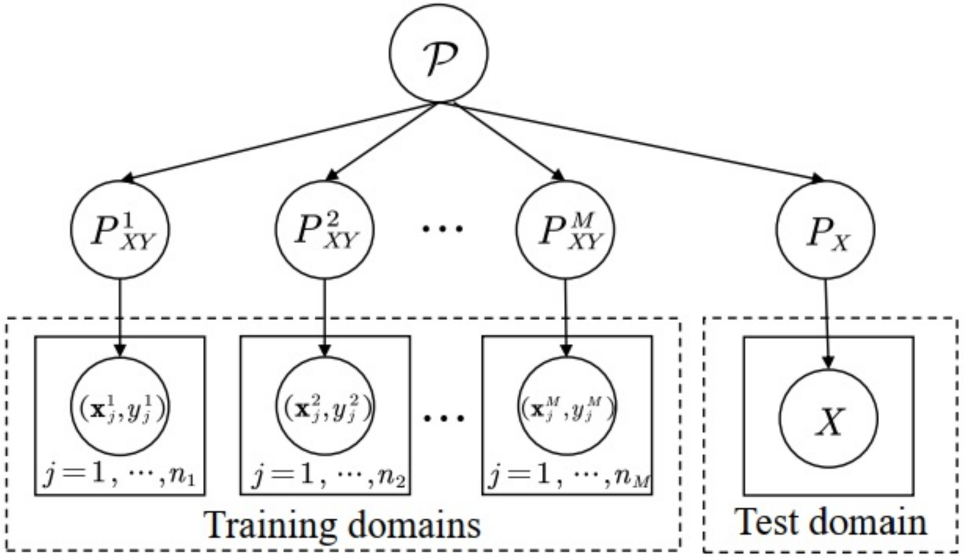
- To solve DA, we need to:
  - Reweight instances to select a subset of two domains where their $d(\cdot,\cdot)$ is small
    - **Tradaboost** [Dai et al'07], **KMM** [Sugiyama et al'08], **Distant TL** [Tan et al'17]
  - Learn domain-invariant feature representations to reduce $d(\cdot,\cdot)$
    - **TCA** [Pan et al'11], **DANN** [Ganin et al'15], **DDC**[Tzeng et al'14] and their extensions as of today
- How about <u>testing phase</u>? <u>Model selection</u>?

DA requires direct access to target domain in training!

Domain generalization

- Definition
  - Given: M training domains $\mathcal{S} = \{\mathcal{S}_i \mid i = 1, \cdots, M\}$, where $\mathcal{S}_i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$
  - Condition:
    - Joint distributions are different, i.e., $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq M$
    - Test domain **cannot be accessed** in training
  - Goal:
    - Achieve minimum test error on test domain
    - $(P_{XY}^i \neq P_{XY}^{test})$

$$\min_h \mathbb{E}_{(\mathbf{x},y)\in\mathcal{S}_{test}}[\ell(h(\mathbf{x}), y)]$$



Related area

Learning paradigm	Training data	Test data	Condition	Test access
Multi-task learning	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{Y}^i \neq \mathcal{Y}^j, 1 \leq i \neq j \leq n$	✓
Transfer learning	$\mathcal{S}^{src}, \mathcal{S}^{tar}$	$\mathcal{S}^{tar}$	$\mathcal{Y}^{src} \neq \mathcal{Y}^{tar}$	✓
Domain adaptation	$\mathcal{S}^{src}, \mathcal{S}^{tar}$	$\mathcal{S}^{tar}$	$\mathcal{X}^{src} \neq \mathcal{X}^{tar}$	✓
Meta-learning	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^{n+1}$	$\mathcal{Y}^i \neq \mathcal{Y}^j, 1 \leq i \neq j \leq n+1$	✓
Lifelong learning	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^i$ arrives sequentially	✓
Zero-shot learning	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^{n+1}$	$\mathcal{Y}^{n+1} \neq \mathcal{Y}^i, 1 \leq i \leq n$	×
Domain generalization	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^{n+1}$	$P(\mathcal{S}^i) \neq P(\mathcal{S}^j), 1 \leq i \neq j \leq n+1$	×

Theory

- Domain adaptation error bound

Theorem 1 (Domain adaptation error bound (non-asymptotic) [20] (Thm. 2))**.** *Let d be the Vapnik–Chervonenkis (VC) dimension [22] of $\mathcal{H}$, and $\mathcal{U}^s$ and $\mathcal{U}^t$ be unlabeled samples of size n from the two domains. Then for any $h \in \mathcal{H}$ and $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$:*

$$\epsilon^t(h) \leq \epsilon^s(h) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}^s, \mathcal{U}^t) + \lambda_{\mathcal{H}} + 4\sqrt{\frac{2d\log(2n) + \log(2/\delta)}{n}},$$

where $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}^s, \mathcal{U}^t)$ is the estimate of $d_{\mathcal{H}\Delta\mathcal{H}}(P_X^s, P_X^t)$ on the two sets of finite data samples.

Theory

- Domain adaptation error bound

Theorem 2 (Domain adaptation error bound on the representation space (non-asymptotic) [23]). *Let $g : \mathcal{X} \to \mathcal{Z}$ be a representation function towards some representation space $\mathcal{Z}$, and let $\mathcal{F}$ denote a hypothesis space with VC dimension d of classifiers on top of $\mathcal{Z}$. For unlabeled samples $\mathcal{U}^s$, $\mathcal{U}^t$ of $\mathbf{x}$ of size n from the two domains, denote the samples of representations as $\tilde{\mathcal{U}}^s := \{g(\mathbf{x}) \mid \mathbf{x} \in \mathcal{U}^s\}$ and $\tilde{\mathcal{U}}^t := \{g(\mathbf{x}) \mid \mathbf{x} \in \mathcal{U}^t\}$. Then for any $f \in \mathcal{F}$ and $\delta \in (0, 1)$, the following inequality holds with probability at least $1 - \delta$:*

$$\epsilon^t(f \circ g) \leq \hat{\epsilon}^s(f \circ g) + \hat{d}_{\mathcal{F}}(\tilde{\mathcal{U}}^s, \tilde{\mathcal{U}}^t) + \lambda_{\mathcal{F}} + \sqrt{\frac{4}{n}\left(d \log \frac{2n}{d} + d + \log \frac{4}{\delta}\right)},$$

where $\hat{\epsilon}^s(f \circ g) := \frac{1}{n}\sum_{\mathbf{x}_j \in \mathcal{U}^s} |f(g(\mathbf{x}_j)) - h^{*s}(\mathbf{x}_j)|$ is the empirical source risk on the finite data samples.

Theory

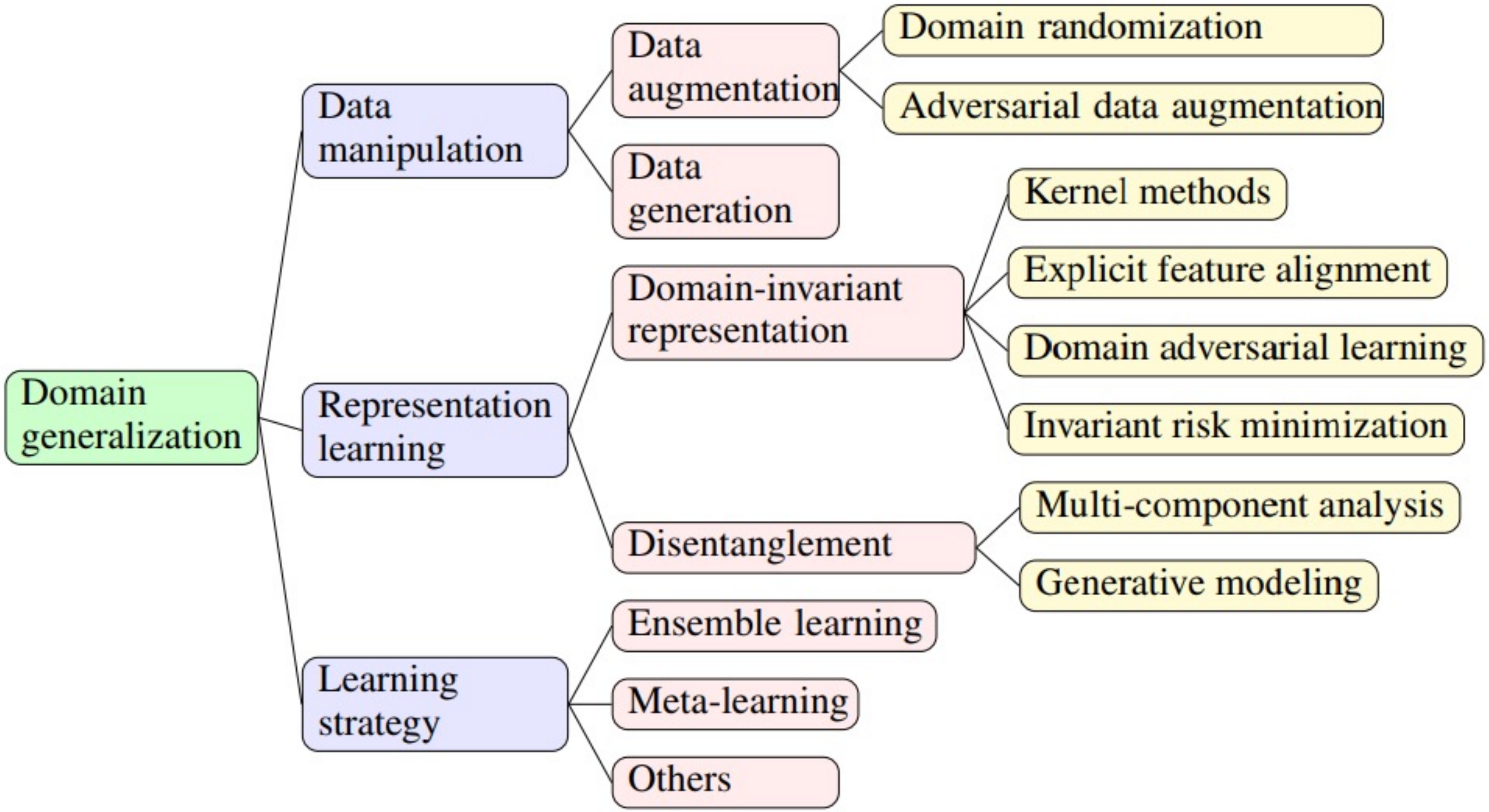
- Domain generalization error bound

Theorem 4 (Domain generalization error bound [27]). *Let* $\gamma := \min_{\pi \in \Delta_M} d_{\mathcal{H}}(P_X^t, \sum_{i=1}^{M} \pi_i P_X^i)$ *with minimizer* π^* [3] *be the distance of* P_X^t *from the convex hull* Λ, *and* $P_X^* := \sum_{i=1}^{M} \pi_i^* P_X^i$ *be the best approximator within* Λ. *Let* $\rho := \sup_{P'_X, P''_X \in \Lambda} d_{\mathcal{H}}(P'_X, P''_X)$ *be the diameter of* Λ. *Then it holds that*

$$\epsilon^t(h) \leq \sum_{i=1}^{M} \pi_i^* \epsilon^i(h) + \frac{\gamma + \rho}{2} + \lambda_{\mathcal{H},(P_X^t, P_X^*)},$$

where $\lambda_{\mathcal{H},(P_X^t, P_X^*)}$ *is the ideal joint risk across the target domain and the domain with the best approximator distribution* P_X^*.

Methodology



Data manipulation

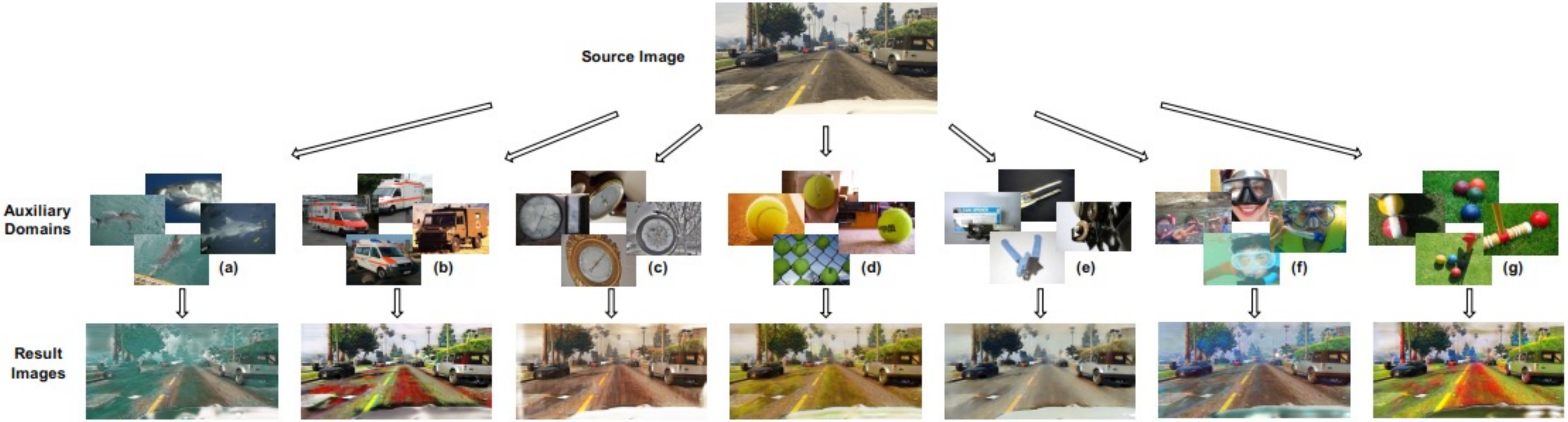
- Data quantity and quality are key factors of generalization
  - Increase quality and quantity

$$\min_h \mathbb{E}_{\mathbf{x},y}[\ell(h(\mathbf{x}), y)] + \mathbb{E}_{\mathbf{x}',y}[\ell(h(\mathbf{x}'), y)]$$

$$\mathbf{x}' = \mathrm{mani}(\mathbf{x}) \begin{cases} \text{Data augmentation} \\ \text{Data generation} \end{cases}$$

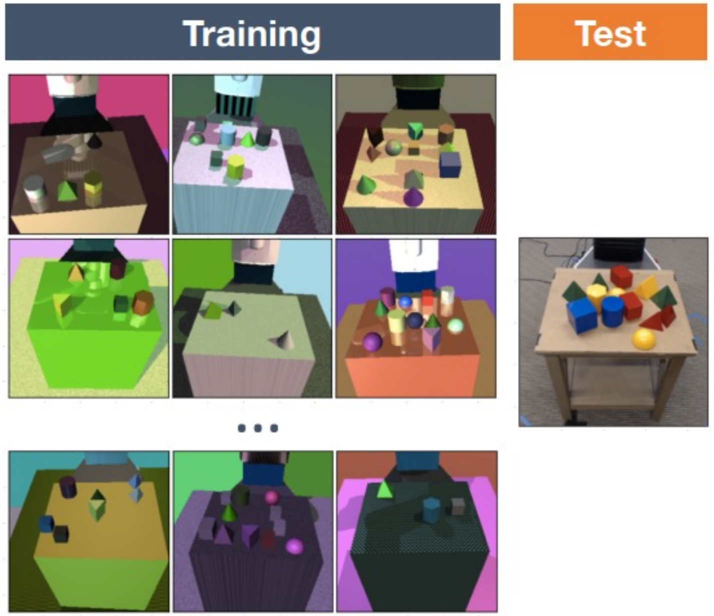
Data augmentation

- Typical augmentation
  - Rotation, noise, color…
- Domain randomization (DR)
  - Shape, position, texture, viewpoint, lighting condition, noise…

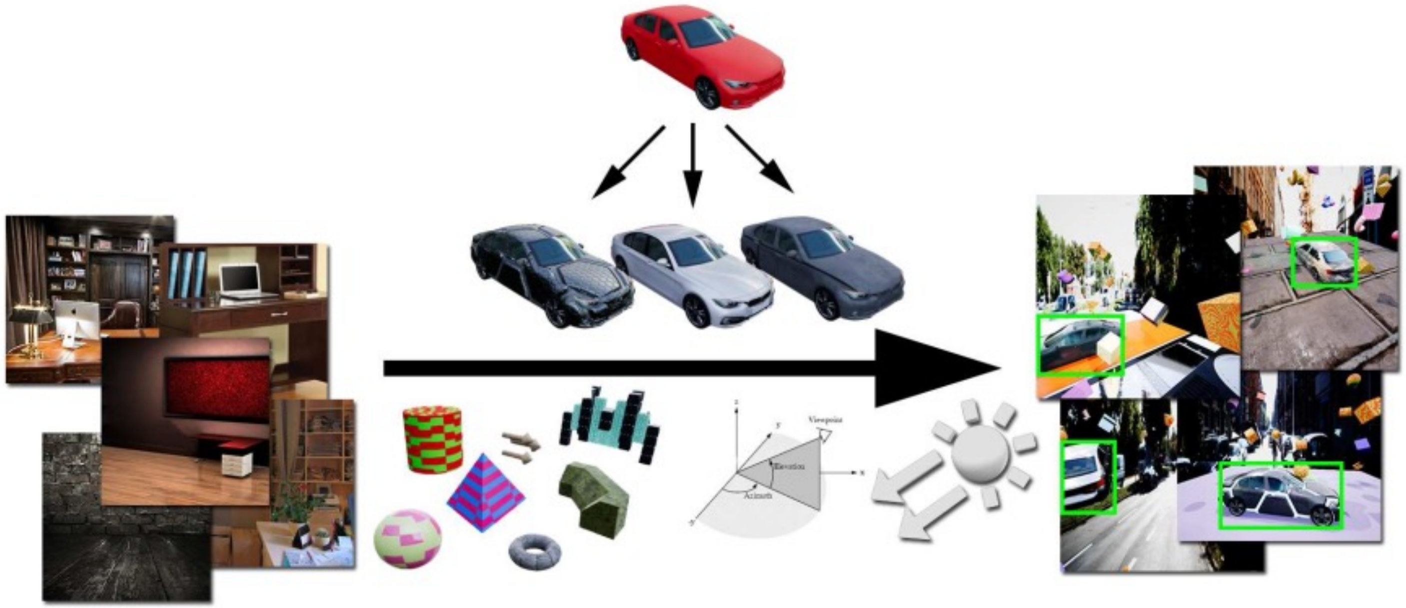


Yue et al. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data. 2019.

Domain randomization



Sim->Real robot control



Synthetic images -> Real images

- Tobin, et al. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. IROS 2017.
- Tremblay et al. Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. CVPR workshop 2018.

Context-aware randomization

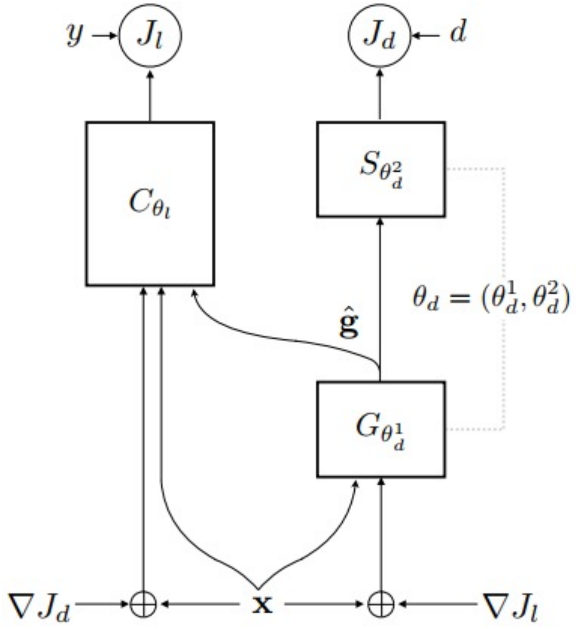
Urban Suburban Rural

Prakash et al. Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data. 2018.

Adversarial data augmentation

- CrossGrad: Adversarially augment data via gradient training
  - Generate data that are with *same* label y, but *different* domain label d

$$\mathbf{x}_i' = \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}_i} J_d(\mathbf{x}_i, d_i)$$



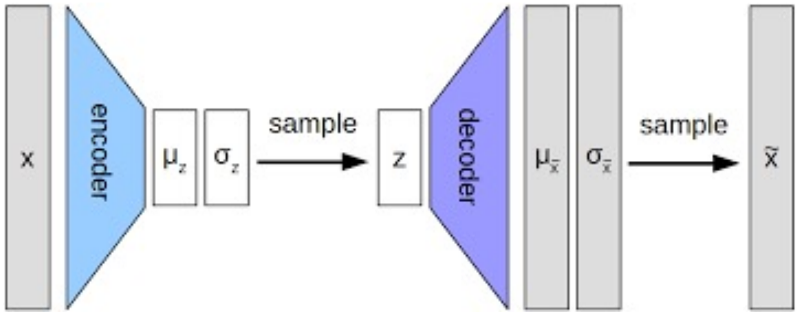
- ADV augmentation
  - Learning the *worse-case* distribution to enable generalization

$$\underset{\theta \in \Theta}{\text{minimize}} \ \sup_{P} \{\mathbb{E}_P[\ell(\theta; (X, Y))] : D_\theta(P, P_0) \leq \rho\}$$

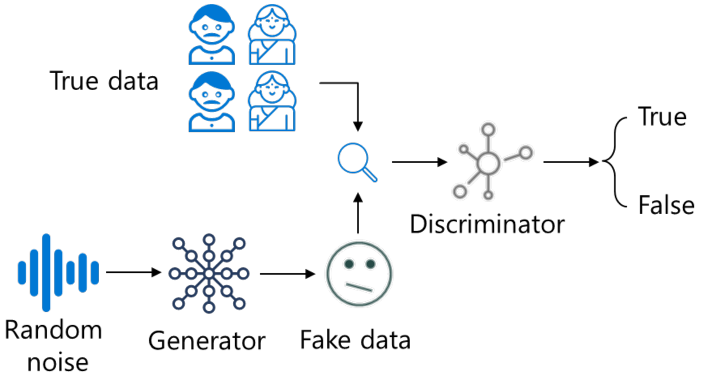
- Shankar et al. Generalizing across Domains via Cross-Gradient Training. ICLR 2018.
- Volpi, et al. Generalizing to Unseen Domains via Adversarial Data Augmentation. NeurIPS 2018.

Data generation

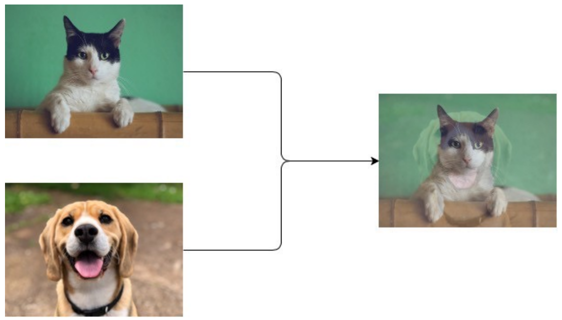
- Directly generate data
  - *Learning* to generate, instead of randomization / adversarial augmentation (Fixed scheme)



Variational auto-encoder (VAE)



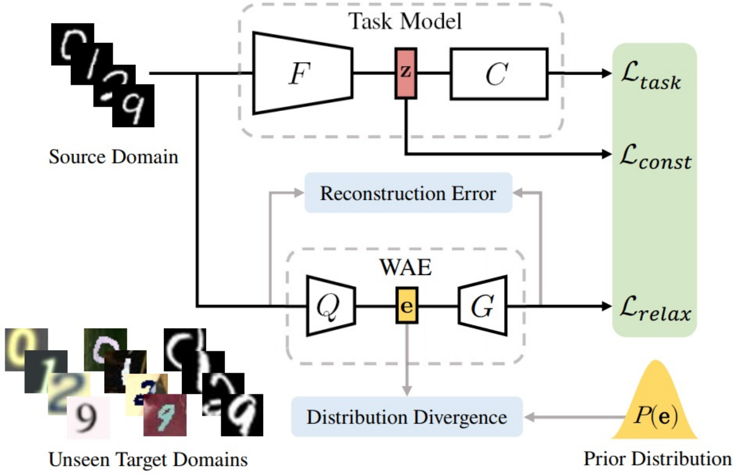
Generative adversarial net (GAN)



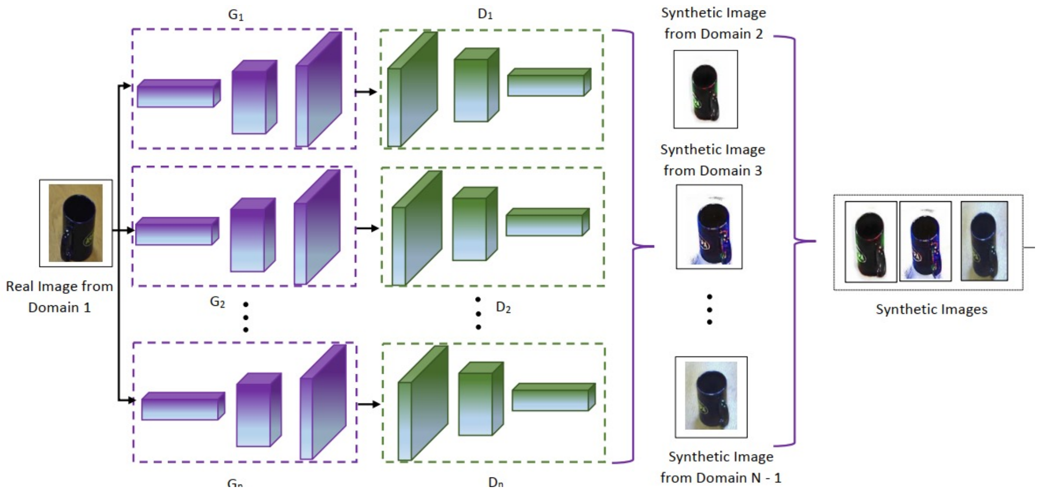
Mixup

- Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
- Zhang H, Cisse M, Dauphin Y N, et al. Mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

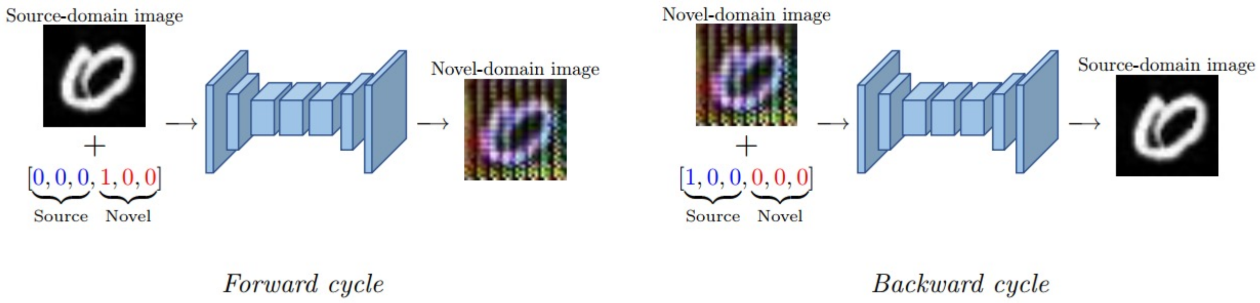
Data generation



VAE for generation



Multi-component generation



Conditional GAN for generation

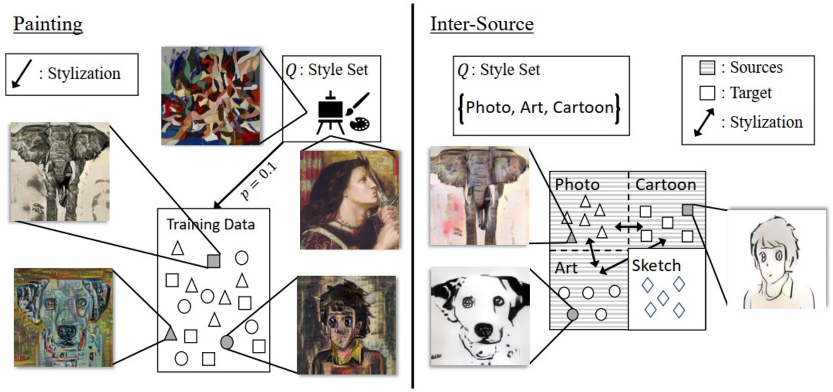
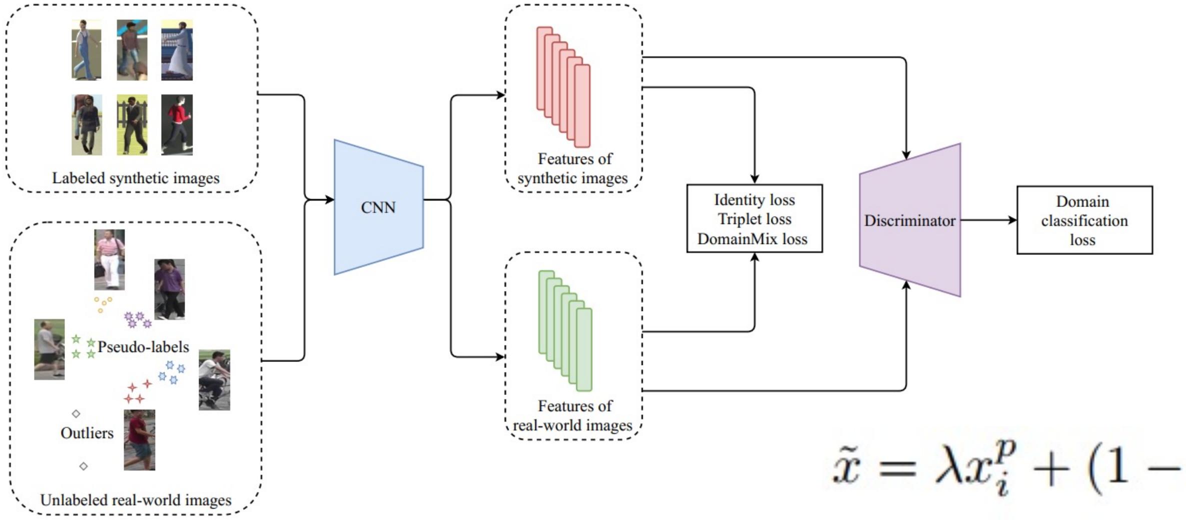


Image stylization

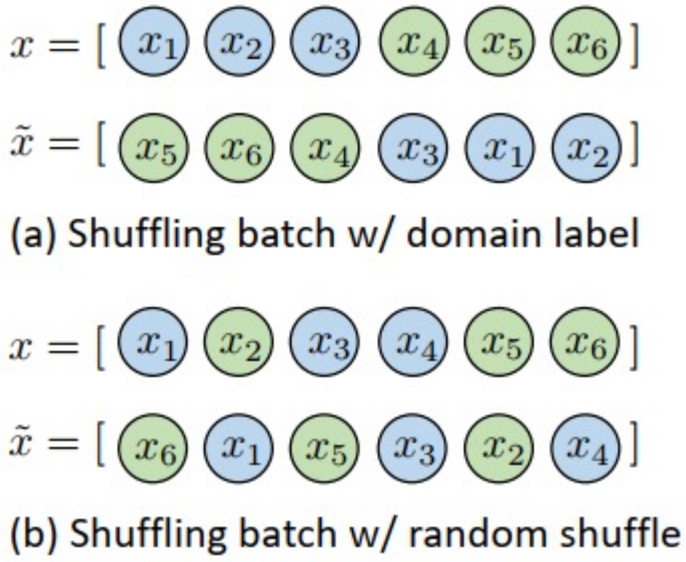
- Qiao et al. Learning to Learn Single Domain Generalization. CVPR 2020.
- Rahman et al. Multi-component Image Translation for Deep Domain Generalization. 2020.
- Zhou et al. Learning to Generate Novel Domains for Domain Generalization. ECCV 2020.
- Somavarapu et al. Frustratingly Simple Domain Generalization via Image Stylization. 2020.

Mixup



DomainMix

$$\tilde{x} = \lambda x_i^p + (1-\lambda)x_j^q$$
$$\tilde{y} = \lambda \hat{y}_i^p + (1-\lambda)\hat{y}_j^q$$



Style mixup

- Wang et al. DomainMix: Learning Generalizable Person Re-Identification Without Human Annotations. 2020.
- Wang et al. Heterogeneous domain generalization via domain mixup. ICASSP 2021.
- Zhou et al. Domain generalization with mixstyle. ICLR 2021.

Representation Learning

- Learning domain-invariant representations

$$\min_{f,g} \mathbb{E}_{\mathbf{x},y}\ell(f(g(\mathbf{x})),y) + \lambda\ell_{\text{reg}}$$

Classifier Feature Regularization

- How to learn representations?
    - Kernel-based methods
    - Domain adversarial learning
    - Explicit feature alignment
    - Invariant risk minimization

Kernel-based methods

- Using kernel methods to learn domain-invariant features
  - DICA: domain-invariant component analysis

$$\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{BS}) = \operatorname{tr}(\widetilde{K}Q) = \operatorname{tr}(B^\top KQKB)$$

  - TCA: Transfer Component Analysis

$$\min_{W} tr(W^T KLKW) + \mu \, tr(W^T W), \text{ s.t. } W^T KHKW = I.$$

  - SCA: Scatter Component Analysis

$$\Psi_\phi(\mathbb{P}) := \mathop{\mathbb{E}}_{x\sim\mathbb{P}}\left[\left\|\mu_{\mathbb{P}} - \phi(x)\right\|_{\mathcal{H}}^2\right] \qquad \sup \frac{\{\text{total scatter}\} + \{\text{between-class scatter}\}}{\{\text{domain scatter}\} + \{\text{within-class scatter}\}}$$

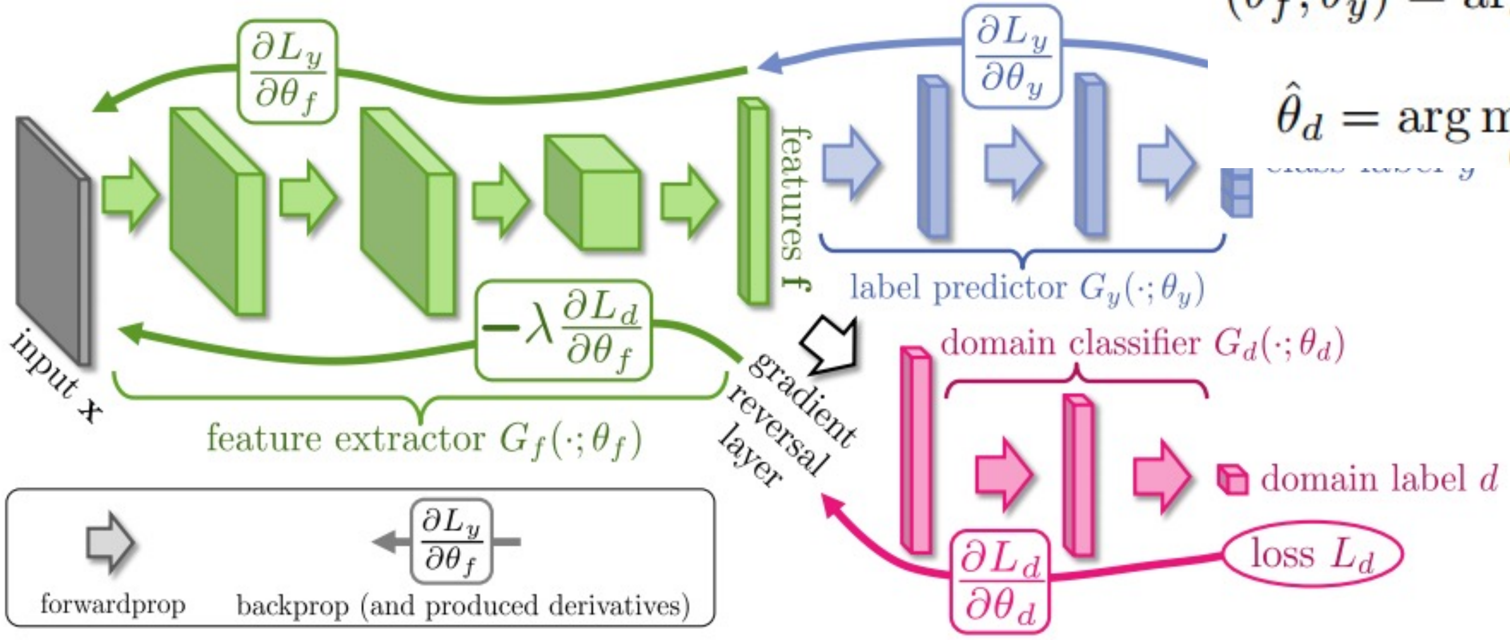
- Blanchard et al. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. NeurIPS 2011.
- Muandet et al. Domain Generalization via Invariant Feature Representation. ICML 2013.
- Grubinger et al. Domain Generalization Based on Transfer Component Analysis. IWANN 2015.
- Ghifary et al. Scatter Component Analysis: A Unified Framework for Domain Adaptation and Domain Generalization. TPAMI 2017.

Domain adversarial learning

- Adversarial training

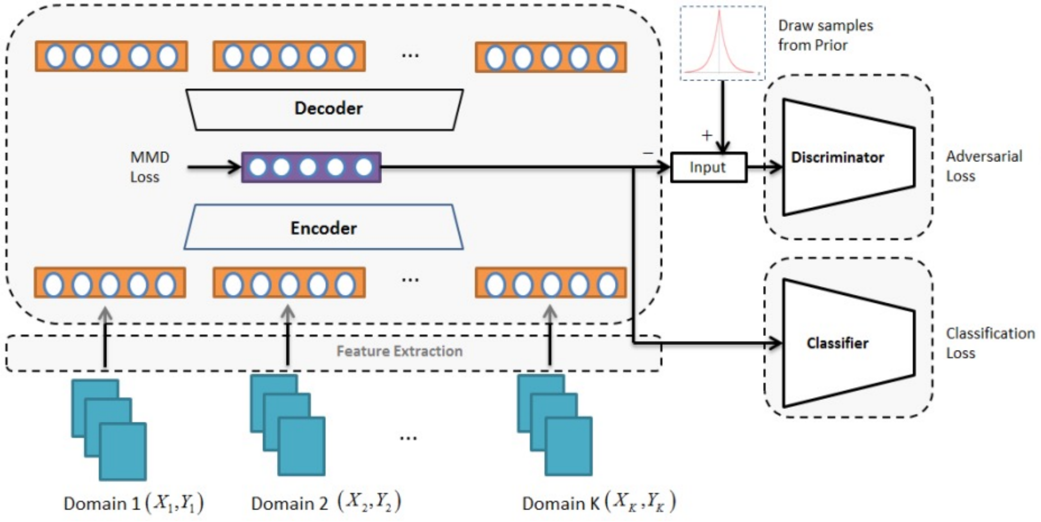
$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d)$$

$$\hat{\theta}_d = \arg\max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d) .$$

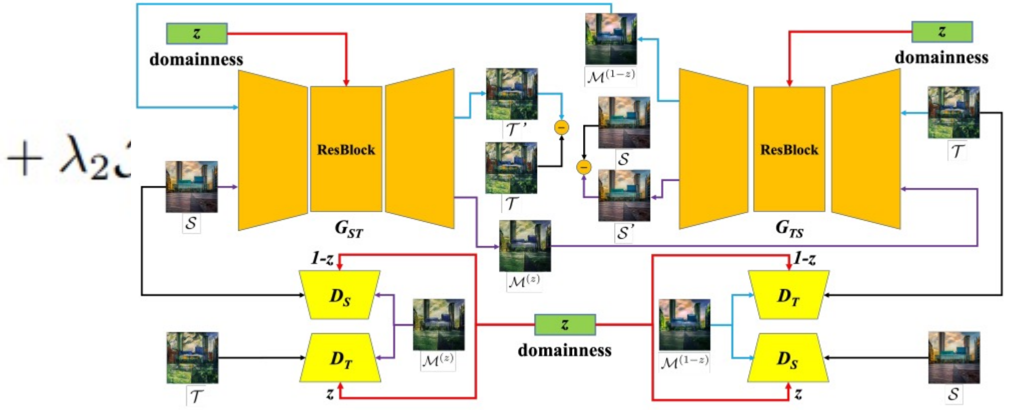


Ganin et al. Unsupervised Domain Adaptation by Backpropagation. ICML 2015.

Domain adversarial learning

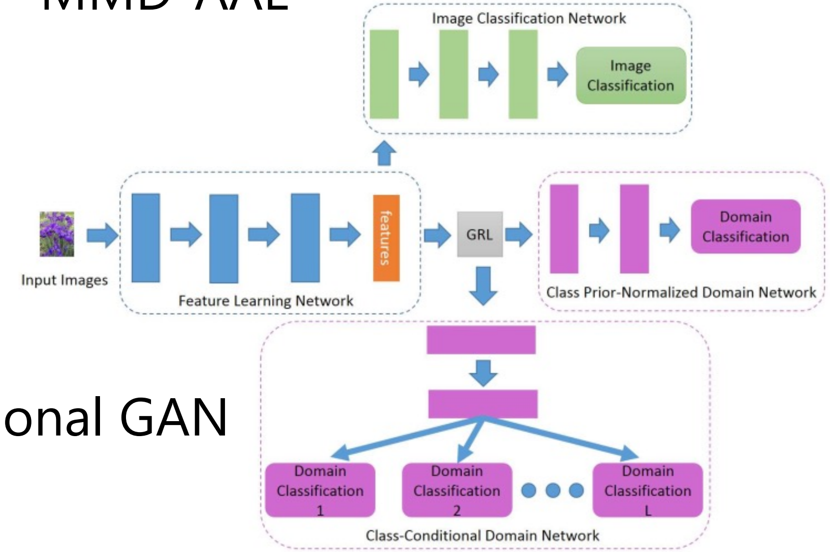


MMD-AAE



$$\mathcal{L} = (1-z)\cdot dist\left(P_S, P_M^{(z)}\right) + z\cdot dist\left(P_T, P_M^{(z)}\right)$$

DLOW

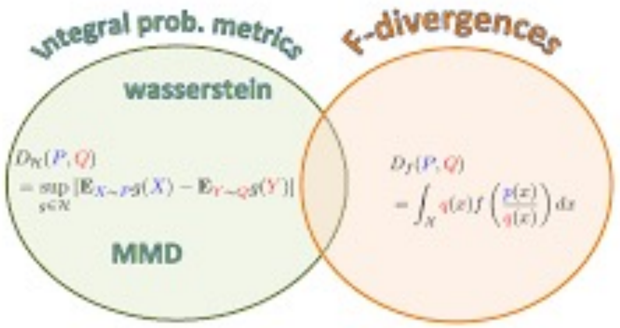


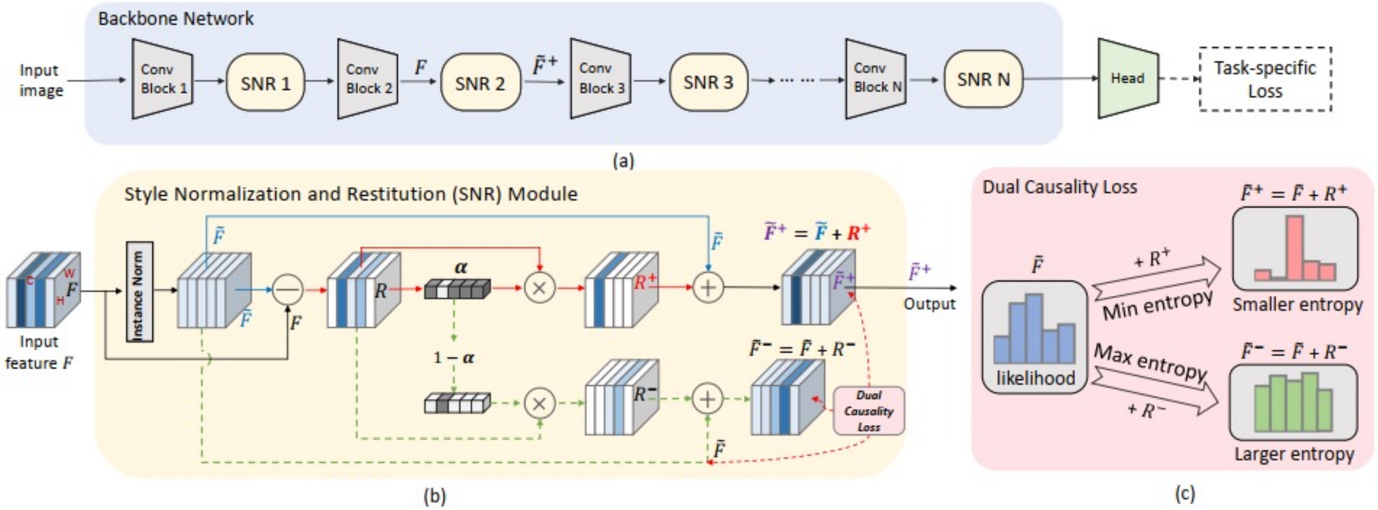
Conditional GAN

- Li et al. Domain Generalization with Adversarial Feature Learning. CVPR 2018.
- Gong et al. DLOW: Domain Flow for Adaptation and Generalization. CVPR 2019.
- Li et al. Deep Domain Generalization via Conditional Invariant Adversarial Networks. ECCV 2018.

Explicit feature alignment

- Distance
  - Maximum mean discrepancy: $\mathrm{MMD}(\mathcal{F}, P_X, P_Y) = \sup_{\|f\|_{\mathcal{H}\leq 1}} (\mathbb{E}_p(f(x)) - \mathbb{E}_p(f(y)))$
  - Correlation alignment: $\ell_{CORAL} = \frac{1}{4d^2}\|C_S - C_T\|_F^2$





Jin X, Lan C, Zeng W, et al. Style Normalization and Restitution for Domain Generalization and Adaptation[J]. arXiv preprint arXiv:2101.00588, 2021.

Invariant risk minimization

- IRM
  - Do not match distributions; enforce optimal *classifier* on top of the representation space to be the same across all domains

$$\min_{\substack{g \in \mathcal{G}, \\ f \in \bigcap_{i=1}^{M} \arg\min_{f' \in \mathcal{F}} \epsilon^i(f' \circ g)}} \sum_{i=1}^{M} \epsilon^i(f \circ g)$$

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{M} \epsilon^i(g) + \lambda \left\| \nabla_f \epsilon^i(f \circ g)\big|_{f=1} \right\|^2$$

Arjovsky et al. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019

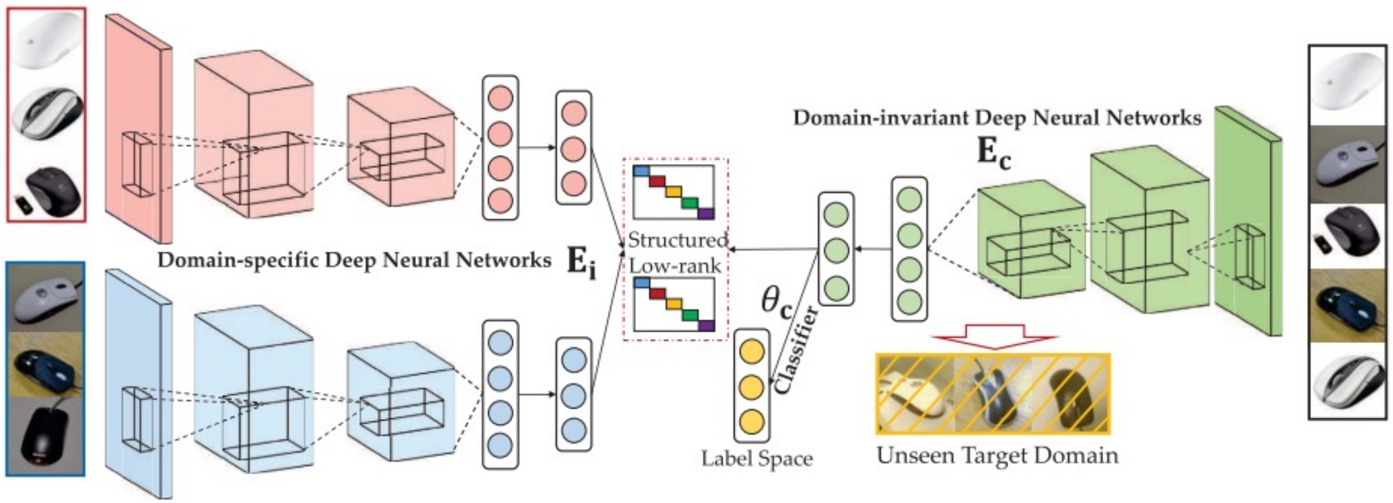
Feature disentanglement

- Formulation

$$\min_{g_c, g_s, f} \mathbb{E}_{\mathbf{x},y} \ell(f(g_c(\mathbf{x})), y) + \lambda \ell_{\text{reg}} + \mu \ell_{\text{recon}}([g_c(\mathbf{x}), g_s(\mathbf{x})], \mathbf{x})$$

- Multi-component analysis
- Generative modeling

- UndoBias $\mathbf{w}_i = \mathbf{w}_0 + \Delta_i$
- Structure low-rank DG

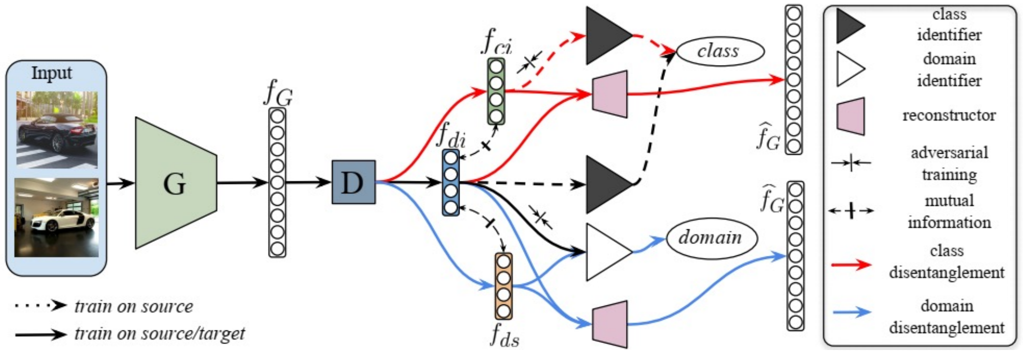


- Khosla A, Zhou T, Malisiewicz T, et al. Undoing the damage of dataset bias[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 158-171.
- Ding Z, Fu Y. Deep domain generalization with structured low-rank constraint[J]. IEEE Transactions on Image Processing, 2017, 27(1): 304-313.

Generative modeling

- DIVA: domain-invariant variational-autoencoder
- DAL: domain-agnostic learning

$$\mathcal{F}_{\mathrm{DIVA}}(d,\mathbf{x},y) := \mathcal{L}_s(d,\mathbf{x},y) + \alpha_d \mathbb{E}_{q_{\phi_d}(\mathbf{z}_d|\mathbf{x})}\left[\log q_{\omega_d}(d|\mathbf{z}_d)\right] + \alpha_y \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{x})}\left[\log q_{\omega_y}(y|\mathbf{z}_y)\right] ,$$

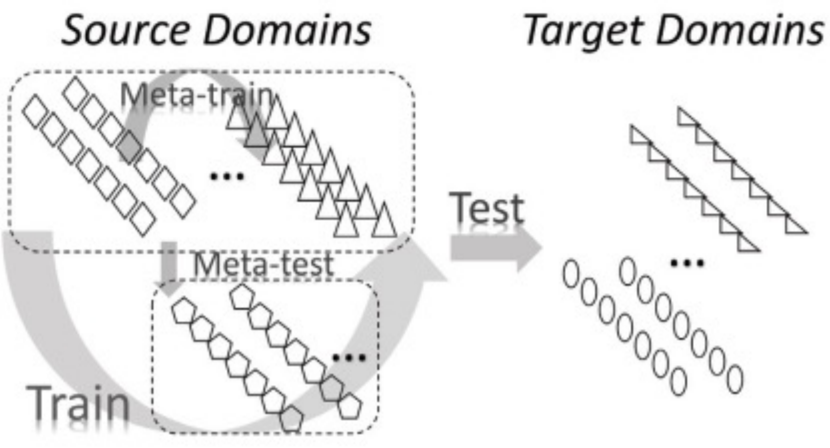


- X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *ICML*, 2019
  Ilse M, Tomczak J M, Louizos C, et al. Diva: Domain invariant variational autoencoders[C]//Medical Imaging with Deep Learning. PMLR, 2020: 322-348.

Learning strategy

- Meta-learning
  - Divide domains into several tasks, then use meta-learning to learn general knowledge
- Ensemble learning
  - Design ensemble models

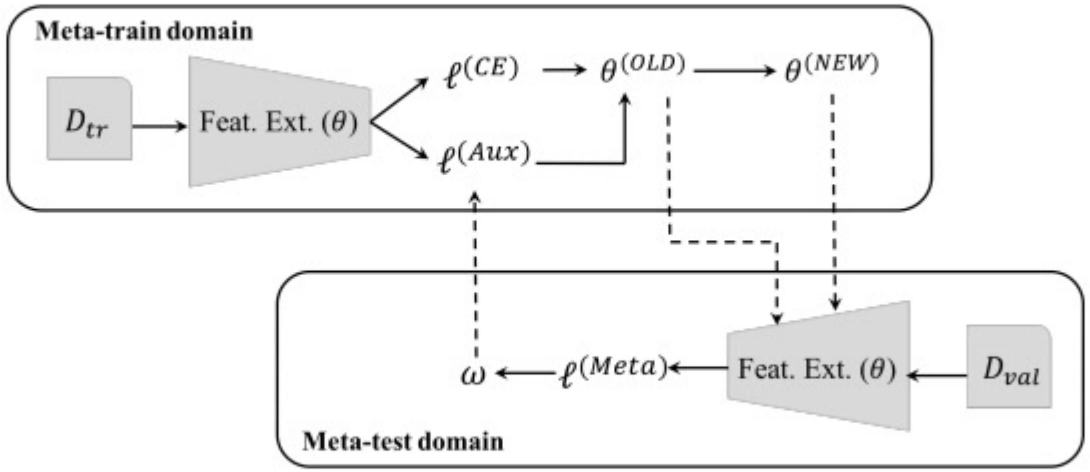
Meta-learning



MLDG

$$\theta^* = \text{Learn}(\mathcal{S}_{mte}; \phi^*)$$
$$= \text{Learn}(\mathcal{S}_{mte}; \text{MetaLearn}(\mathcal{S}_{mtrn})),$$

$$\theta = \theta - \alpha \frac{\partial(\ell(\mathcal{S}_{mte}; \theta) + \beta \ell(\mathcal{S}_{mtrn}; \phi))}{\partial \theta}$$

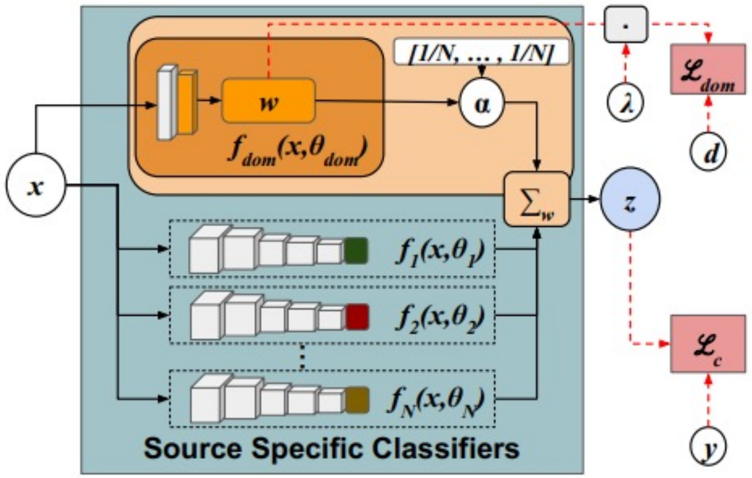


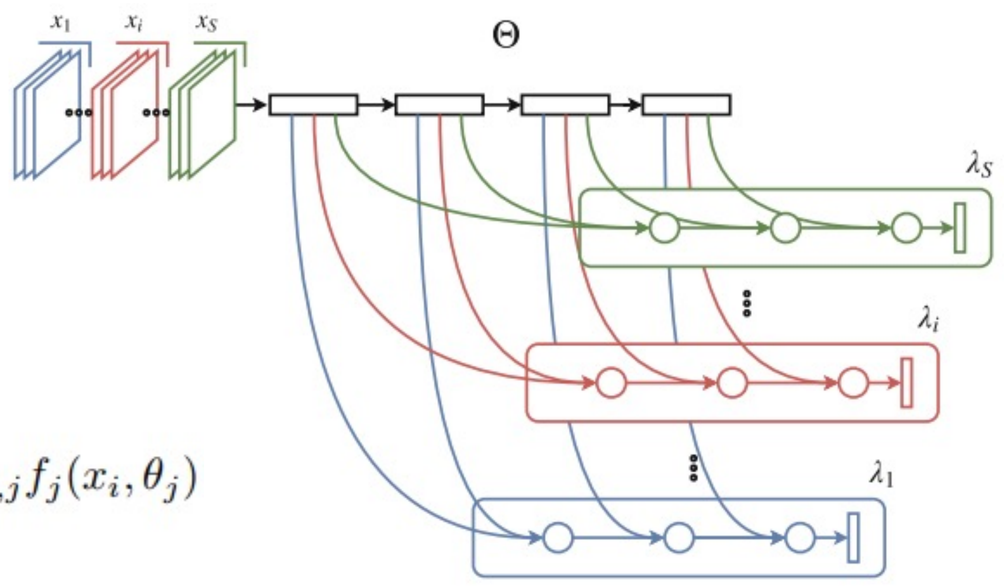
Feature Critic training

$$\min_{\theta, \phi_j s} \sum_{D_j \in \mathcal{D}_{\text{trn}}} \sum_{d_j \in D_j} \ell^{(\text{CE})}(g_{\phi_j}(f_\theta(x^{(j)})), y^{(j)}) + \ell^{(\text{Aux})}$$

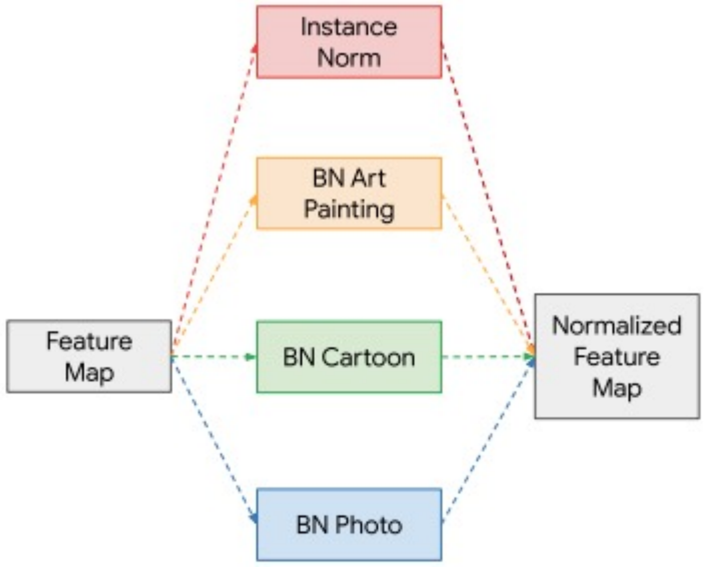
- Li D, Yang Y, Song Y Z, et al. Learning to generalize: Meta-learning for domain generalization. AAAI 2018.
- Li Y, Yang Y, Zhou W, et al. Feature-critic networks for heterogeneous domain generalization. ICML 2019.

Ensemble learning





$$z_i = f(x_i, \Theta) = \sum_{j=1}^{N} w_{i,j} f_j(x_i, \theta_j)$$



- Mancini M, Bulo S R, Caputo B, et al. Best sources forward: domain generalization through source-specific nets. ICIP 2018.
- Segu M, Tonioni A, Tombari F. Batch normalization embeddings for deep domain generalization[J]. arXiv preprint arXiv:2011.12672, 2020.
- D’Innocente A, Caputo B. Domain generalization with domain-specific aggregation modules[C]//German Conference on Pattern Recognition. Springer, Cham, 2018: 187-198.

Datasets and applications

- Datasets

Dataset	#Domain	#Class	#Sample	Description
Office-Caltech	4	10	2,533	Caltech, Amazon, Webcam, DSLR
Office-31	3	31	4,110	Amazon, Webcam, DSLR
PACS	4	7	9,991	Art, Cartoon, Photos, Sketches
VLCS	4	5	10,729	Caltech101, LabelMe, SUN09, VOC2007
Office-Home	4	65	15,588	Art, Clipart, Product, Real
Terra Incognita	4	10	24,788	Wild animal images taken at locations L100, L38, L43, L46
Rotated MNIST	6	10	70,000	Digits rotated from 0° to 90° with an interval of 15°
DomainNet	6	345	586,575	Clipart, Infograph, Painting, Quickdraw, Real, Sketch

- Application
  - Image classification / segmentation / detection / ReID
  - Reinforcement learning
  - Parkinson's disease
  - Activity recognition
  - Fault diagnosis

Challenges

- Continuous domain generalization
  - Continuous / online learning
- Generalize to novel categories
  - New categories instead of closed set
- Interpretable domain generalization
  - Learning to interpret: why it can generalize?
- Large-scale pre-training / self-learning and DG
  - The role of pre-training and self-learning with DG
- Performance evaluation
  - Develop more fair and application-driven evaluation standards

Microsoft

Thanks

jindong.wang@microsoft.com

https://arxiv.org/abs/2103.03097 https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG