IJCAI-ECAI 2022 Tutorial on

Domain Generalization


Jindong Wang
Microsoft Research Asia


Haoliang Li
City University of Hong Kong


Sinno Jialin Pan
Nanyang Technological University

Tutorial website: https://dgresearch.github.io/

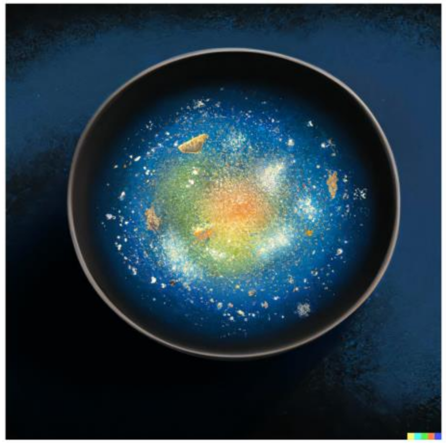
The breakthrough of AI

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

that is a portal to another dimension that looks like a monster as a planet in the universe

as a 1960s poster as mixed media with needlework as digital art



TEXT DESCRIPTION

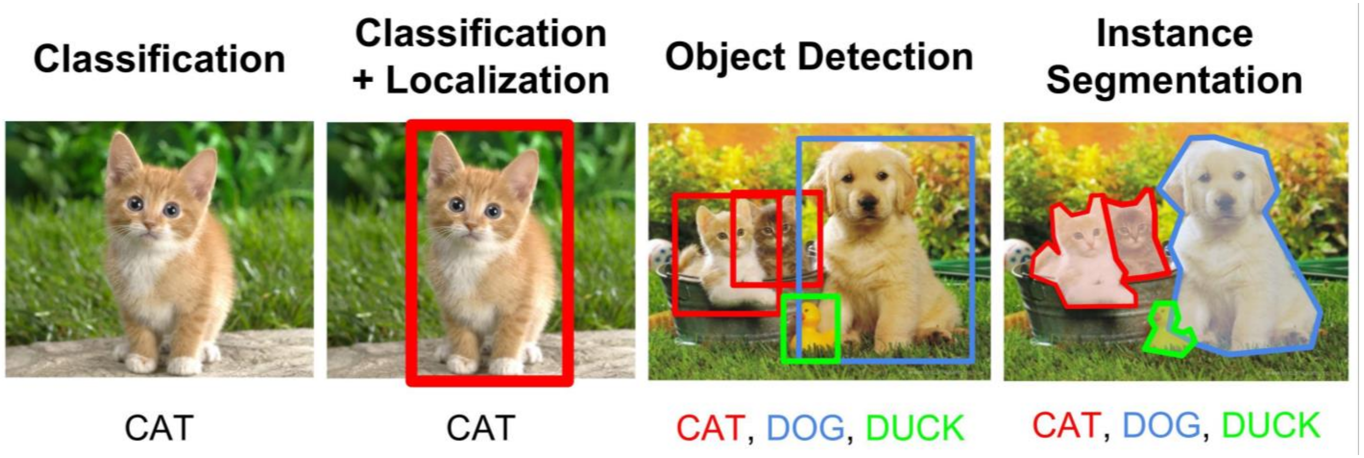
An astronaut Teddy bears A bowl of soup

mixing sparkling chemicals as mad scientists shopping for groceries working on new AI research

as kids' crayon art on the moon in the 1980s underwater with 1990s technology

https://openai.com/dall-e-2/#demos

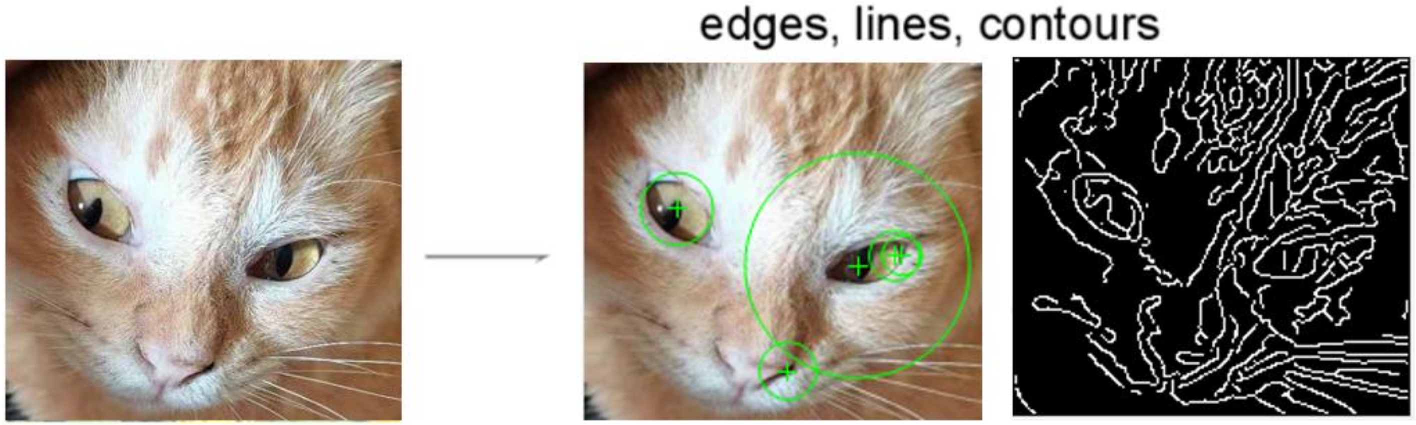
Artificial intelligence



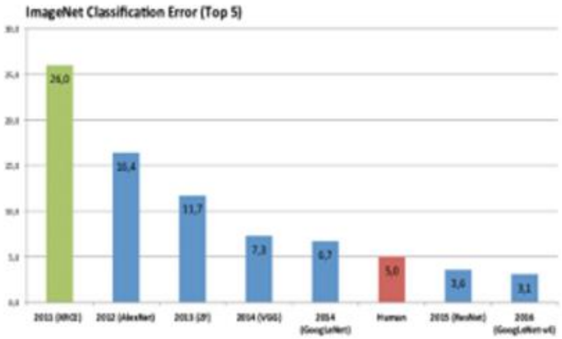
Background

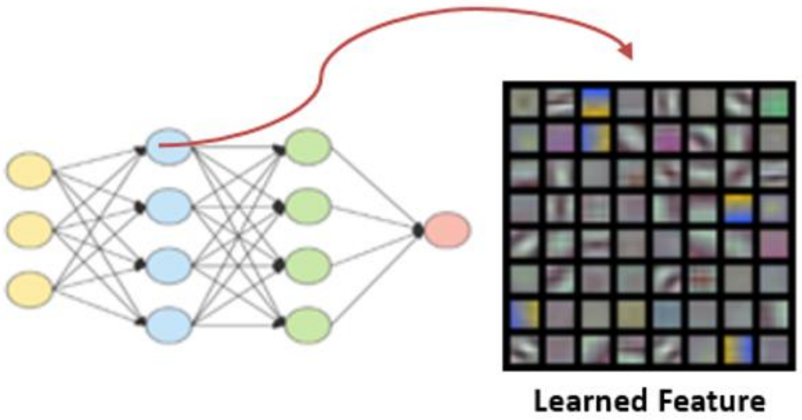
- Computer vision: How do we represent an image?

Past:



Now:





Artificial intelligence?

- More artificial, more intelligence…

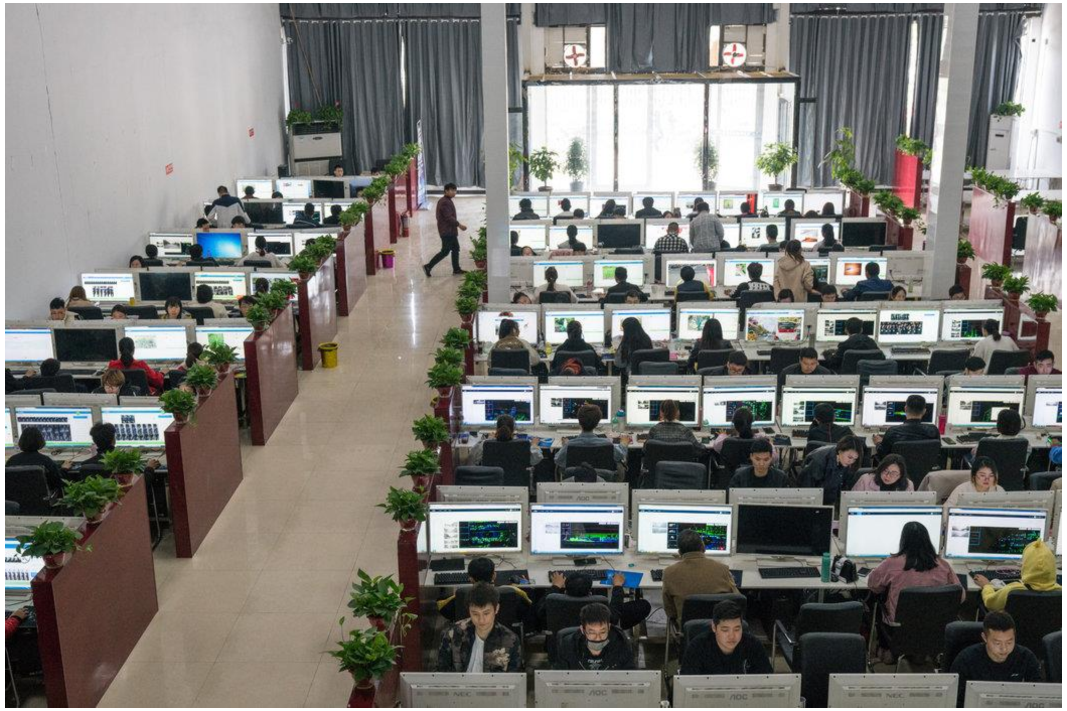


Figure credit: https://cn.nytimes.com/technology/20181126/china-artificial-intelligence-labeling/

Background

- Models **do not generalize well** to new domains; not like humans!
- Are big data always available?
  - It is impossible to consider data in **all scenarios**.
  - Data can be protected under **privacy regulation**.

- Pan et al. A Survey on Transfer Learning. IEEE TKDE 2010.
- Wang et al. Generalizing to unseen domains: a survey on domain generalization. IEEE TKDE 2022.

Domain adaptation

- DA: Train on source and adapt to target

ImageNet

CIFAR-100

Source Domain $\sim P_S(X, Y)$ $\neq$ Target Domain $\sim P_T(Z, H)$

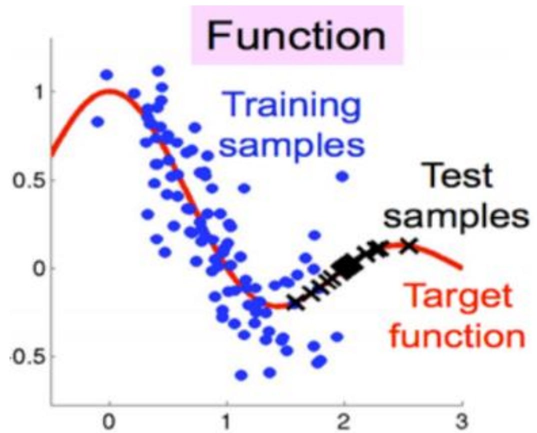
lots of **labeled** data

unlabeled or limited labels

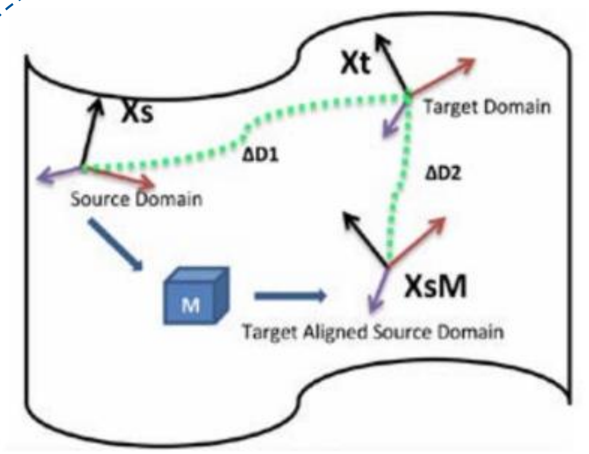
$$D_S = \{(\mathbf{x}_i, y_i), \forall i \in \{1, \ldots, N\}\}$$

$$D_T = \{(\mathbf{z}_j, ?), \forall j \in \{1, \ldots, M\}\}$$

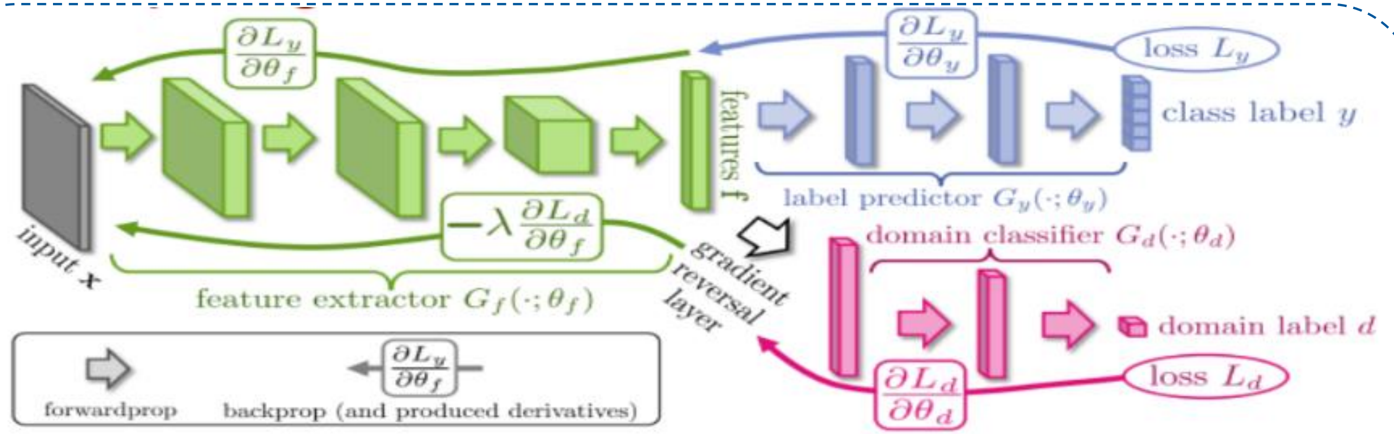
Domain adaptation



Sample Re-weighting [Huang,NIPS'06]



Subspace Learning [Fernando, ICCV'13]



Adversarial Learning [Ganin, ICML'15, Tzeng, CVPR'17]

Domain adaptation: Train on Source and Adapt to Target

Image

CiFAR100

Are we able to obtain unlabeled testing data?

Source Domai

lots of **label**

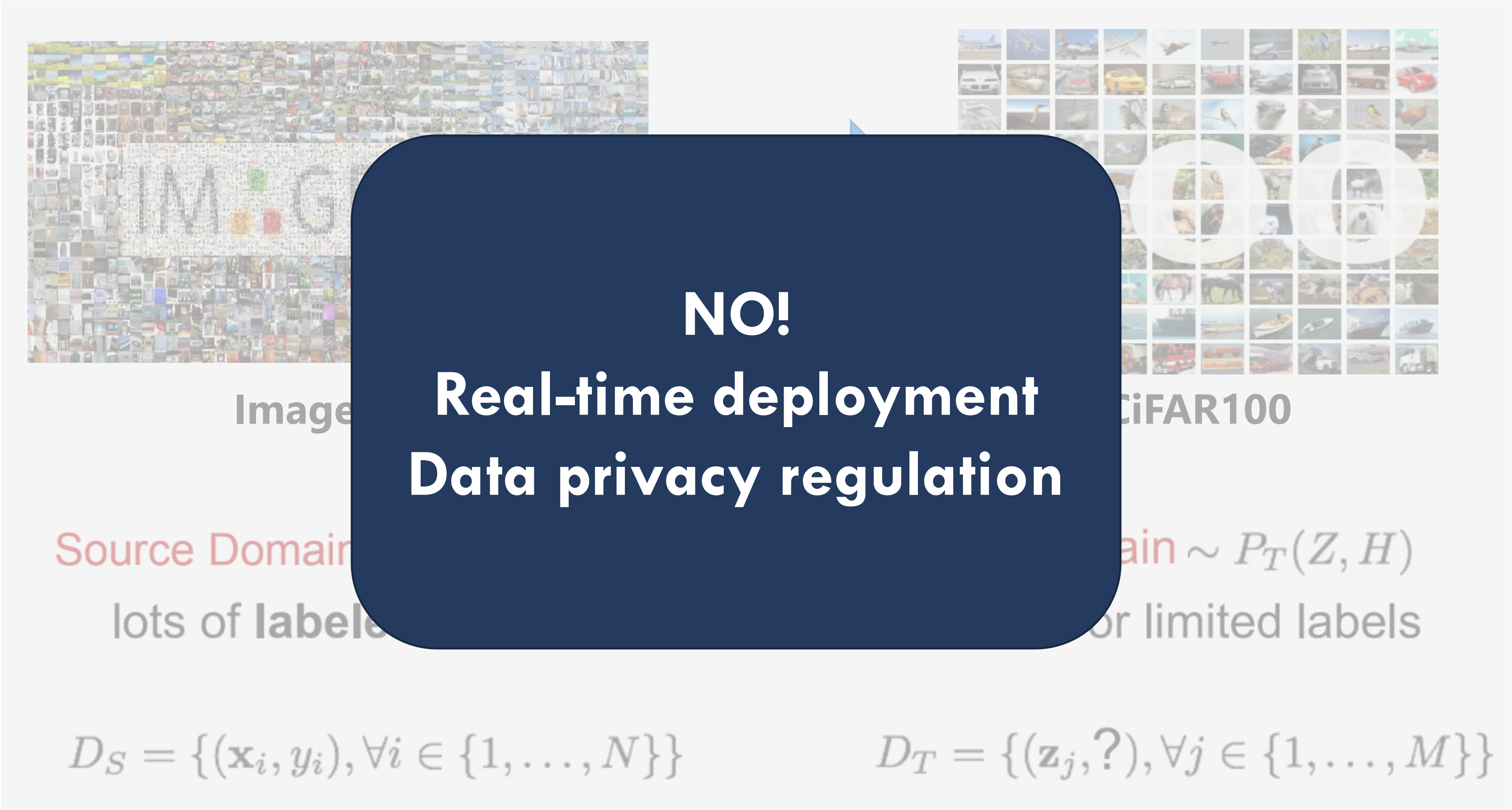
ain $\sim P_T(Z, H)$

or limited labels

$D_S = \{(\mathbf{x}_i, y_i), \forall i \in \{1, \ldots, N\}\}$

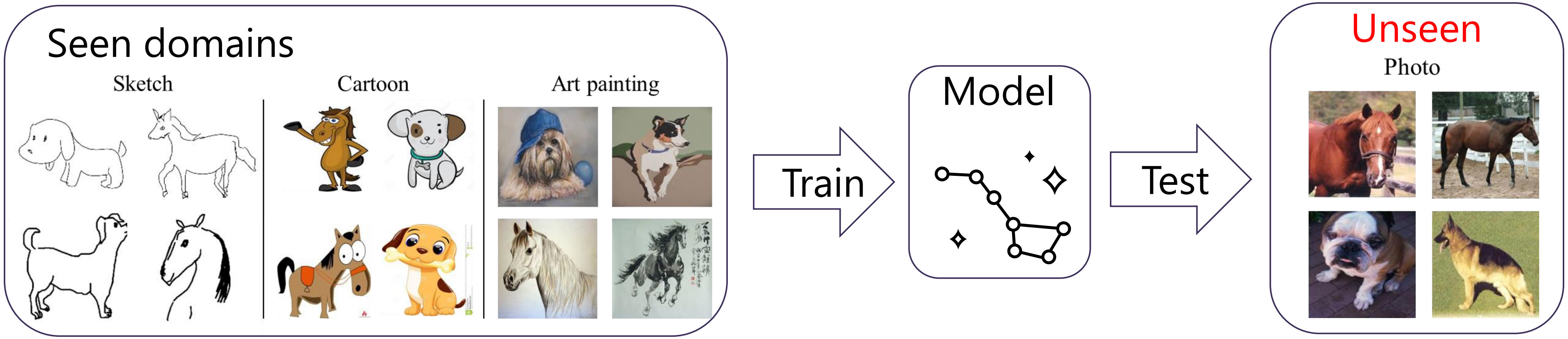
$D_T = \{(\mathbf{z}_j, ?), \forall j \in \{1, \ldots, M\}\}$

Domain adaptation: Train on Source and Adapt to Target



Domain Generalization

- DG: Build a system for previously ***unseen*** datasets given one or multiple training datasets.

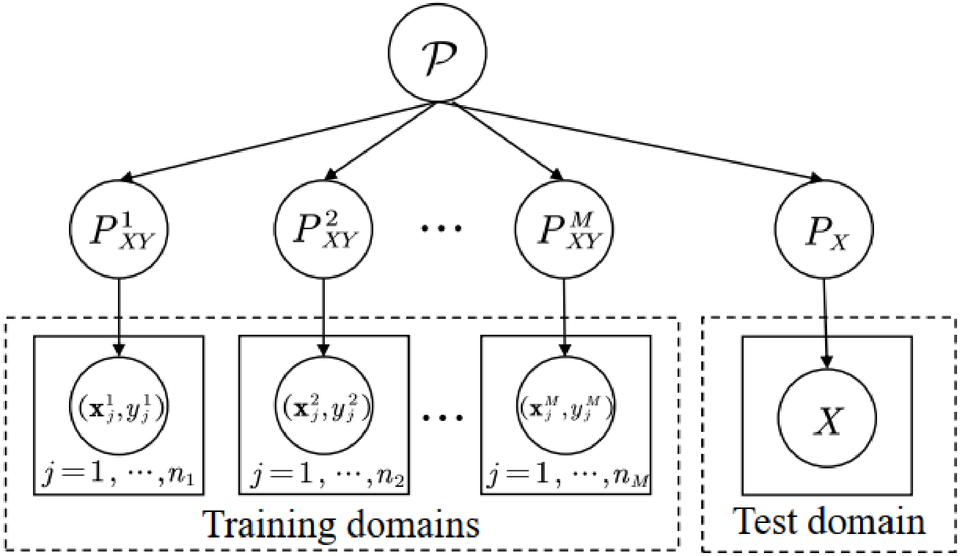


Wang et al. Generalizing to unseen domains: a survey on domain generalization. IEEE TKDE 2022.

Formal definition of 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
    - $(P_{XY}^i \neq P_{XY}^{test})$

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



Wang et al. Generalizing to unseen domains: a survey on domain generalization. IEEE TKDE 2022.

Different DG settings

- Different DG settings

The general setting;
Focus of this tutorial

Setting	Situation
Traditional domain generalization	The traditional setting
Single-source domain generalization	Only 1 source domain available for training
Semi-supervised domain generalization	Training domains are partially labeled
Federated domain generalization	Training data cannot be accessed by central server
Open domain generalization	Training and test domains have different label spaces
Unsupervised domain generalization	Training domains are totally unlabeled

- Peng X, Qiao F, Zhao L. Out-of-domain Generalization from a Single Source: A Uncertainty Quantification Approach[J]. arXiv preprint arXiv:2108.02888, 2021.
- Lin L, Xie H, Yang Z, et al. Semi-Supervised Domain Generalization in Real World: New Benchmark and Strong Baseline[J]. arXiv preprint arXiv:2111.10221, 2021.
- Zhang L, Lei X, Shi Y, et al. Federated Learning with Domain Generalization[J]. arXiv preprint arXiv:2111.10487, 2021.
- Shu Y, Cao Z, Wang C, et al. Open domain generalization with domain-augmented meta-learning. CVPR 2021.
- Qi L, Wang L, Shi Y, et al. Unsupervised Domain Generalization for Person Re-identification: A Domain-specific Adaptive Framework[J]. arXiv preprint arXiv:2111.15077, 2021.

Relation with existing paradigms

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	$\mathcal{S}^1, \cdots, \mathcal{S}^n$	$\mathcal{S}^{n+1}$	$P(\mathcal{S}^i) \neq P(\mathcal{S}^j), 1 \leq i \neq j \leq n+1$	×

DG has close relationship with other paradigms, but also different from them

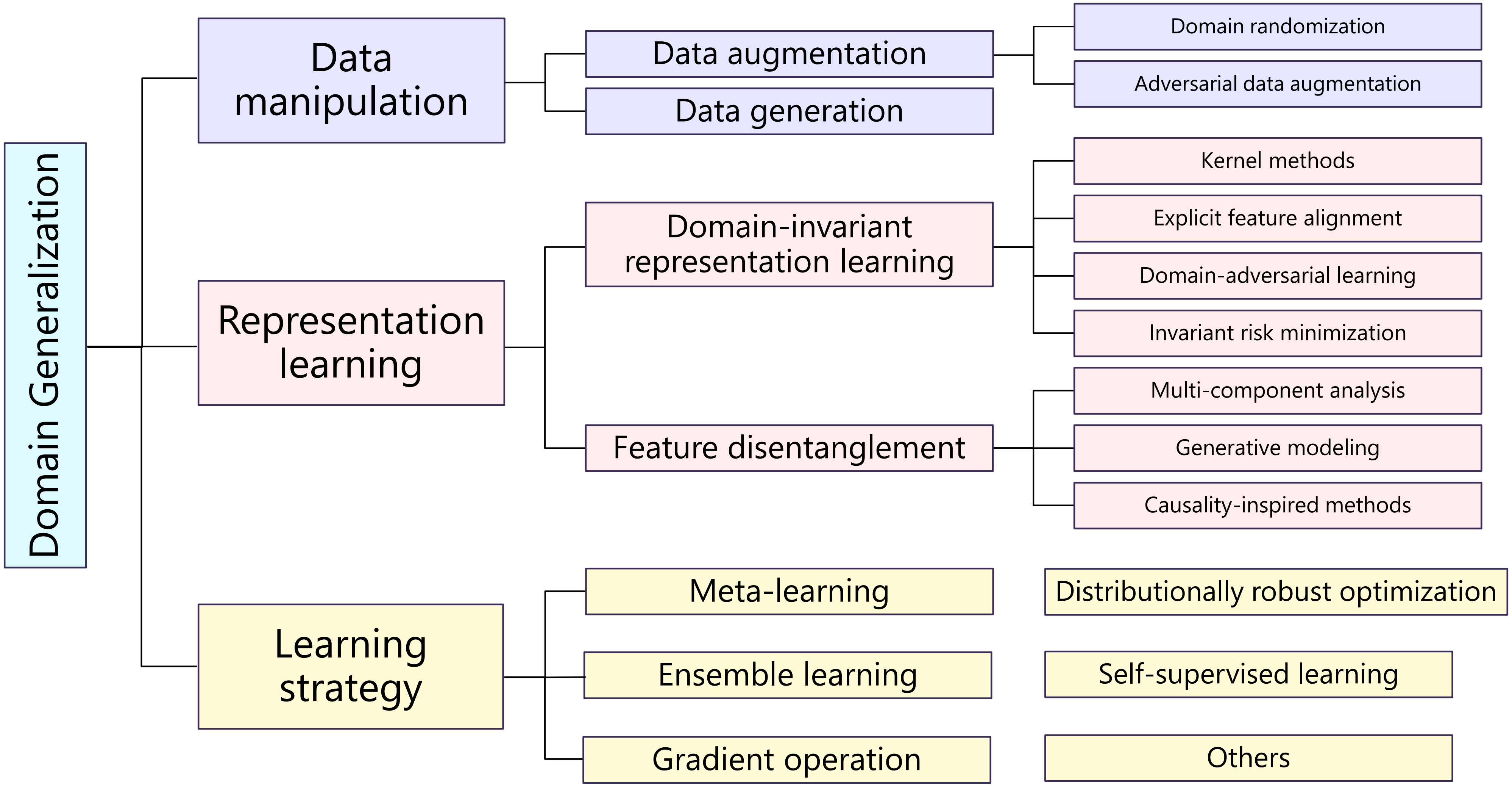
Wang et al. Generalizing to unseen domains: a survey on domain generalization. IEEE TKDE 2022.

Overview of this tutorial

- Introduction of DG & related areas
- Methodology
- Applications
- Datasets, benchmark & evaluation
- Theory & challenges

- Introduction of DG & related areas
- **Methodology**
- Applications
- Datasets, benchmark & evaluation
- Theory & challenges

Overview of DG methodology



Wang et al. Generalizing to unseen domains: a survey on domain generalization. IEEE TKDE 2022.

Data manipulation for DG

Data manipulation

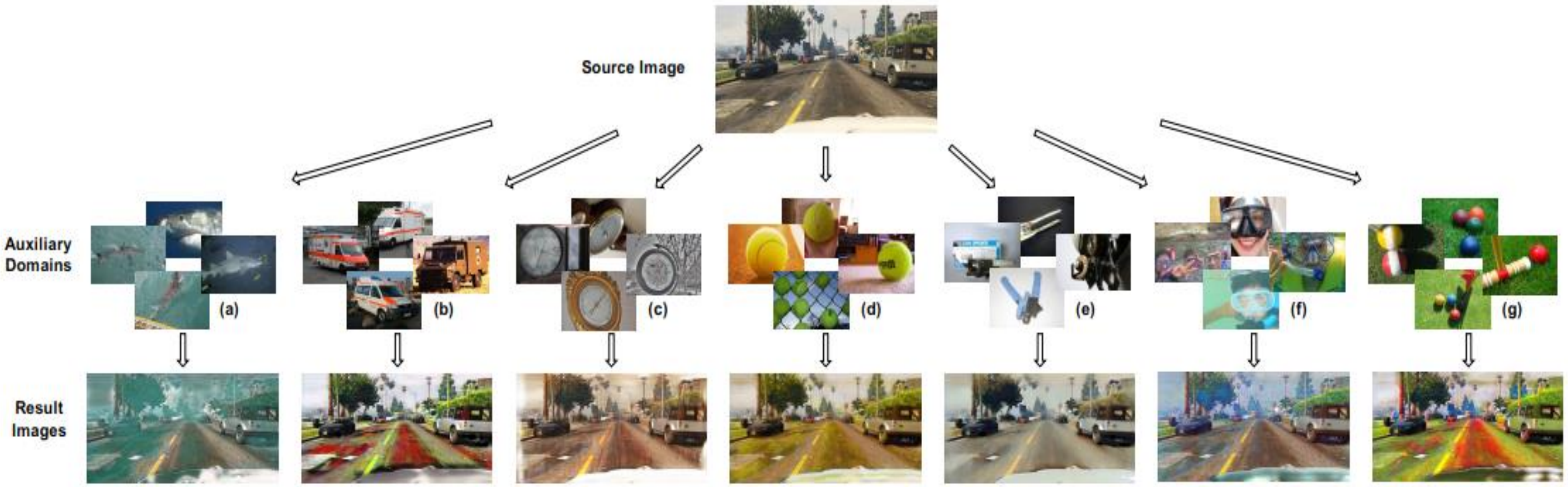
- Data quantity and quality are key factors of generalization
  - Increase *quality* and *quantity*

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

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

Data augmentation

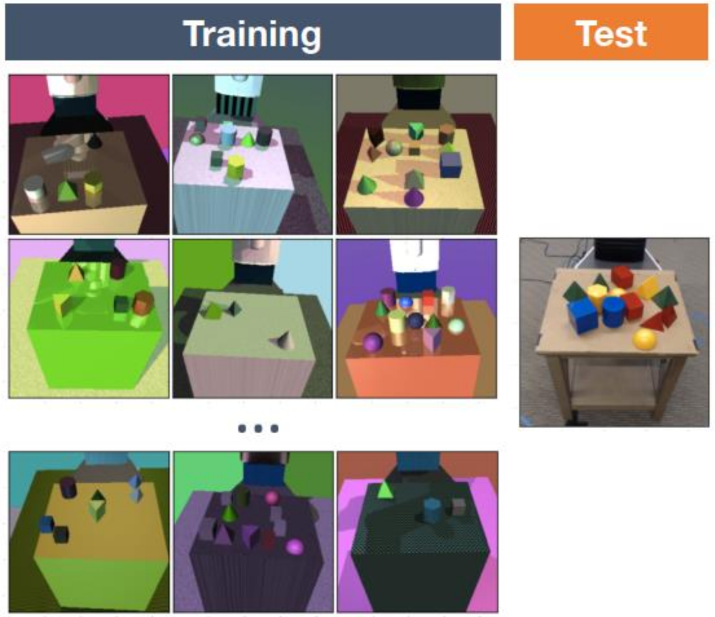
- Typical augmentation
  - Rotation, noise, color…
- Domain randomization (DR)
  - Randomly draw K real-life categories from ImageNet for stylizing the synthetic images.

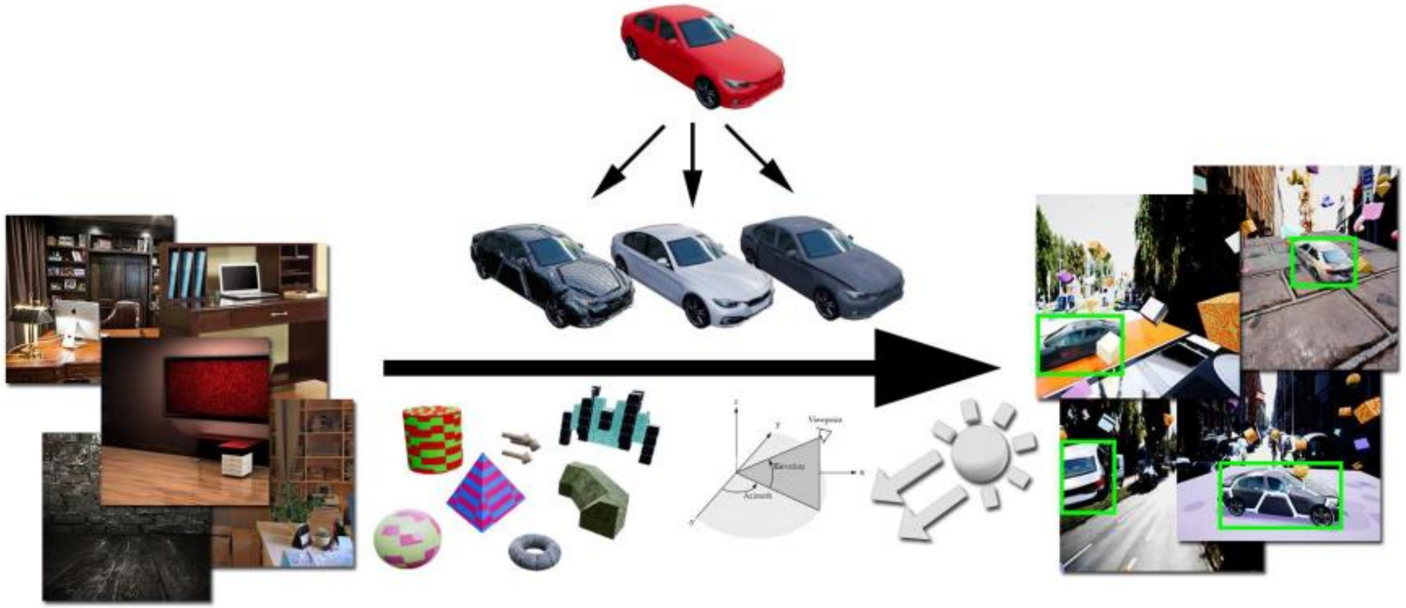


Yue et al. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data. ICCV,2019.

Domain randomization

Domain randomization through graphics software.



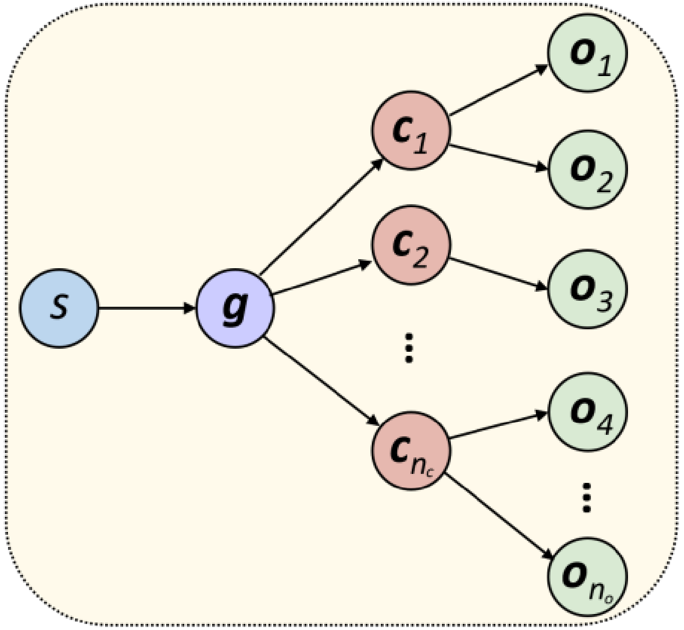


Sim->Real robot control

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



$$p(I, s, \mathbf{g}, \mathbf{o}_{1..n_o}, \mathbf{c}_{1..n_c}) = p(I|s, \mathbf{g}, \mathbf{o}_{1..n_o}, \mathbf{c}_{1..n_c}) \cdot \prod_{j=1}^{n_o} p(\mathbf{o}_j|\mathbf{c}_i) \prod_{i=1}^{n_c} p(\mathbf{c}_i|\mathbf{g}) p(\mathbf{g}|s) p(s)$$

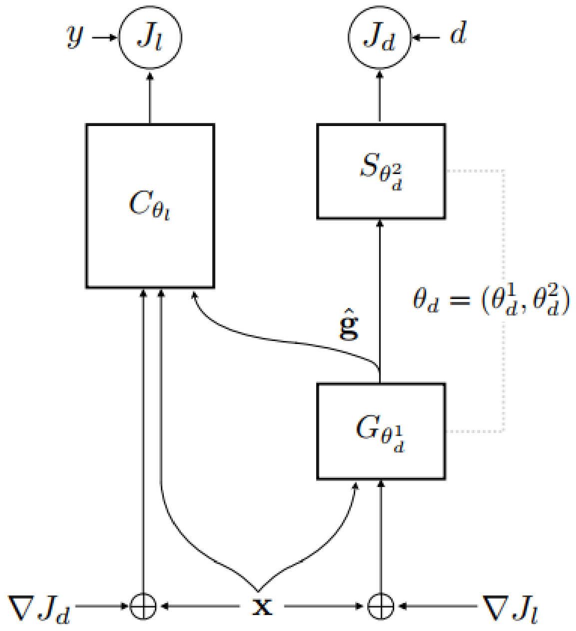
Urban Suburban Rural

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

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



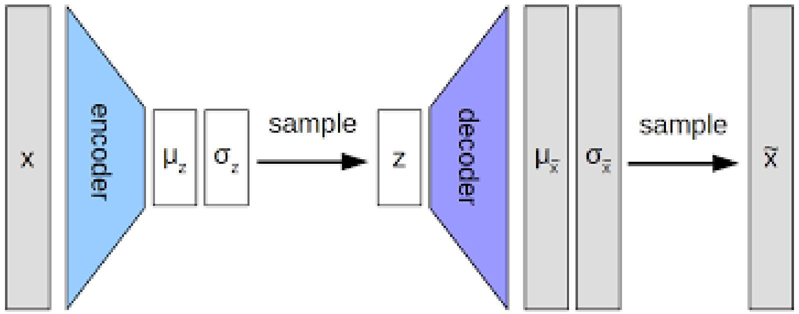
- ADV augmentation
  - Learning the *worse-case* distribution to enable generalization

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

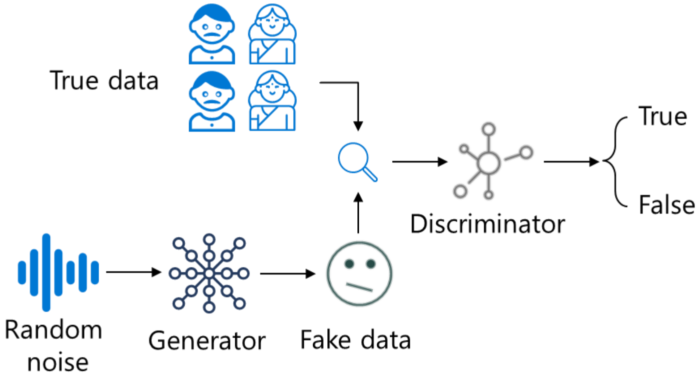
- Shankar et al. Generalizing across Domains via Cross-Gradient Training. ICLR 2018.
- Volpi, et al. Generalizing to Unseen Domains via Adversarial Data Augmentation. NeurIPS 2018.

Data generation

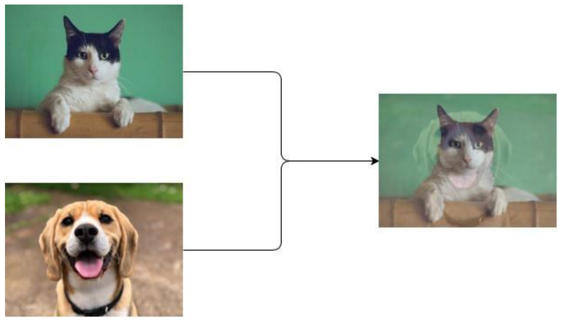
- Directly generate data
  - *Learning* to generate, instead of randomization / adversarial augmentation (Fixed scheme)



Variational auto-encoder (VAE)



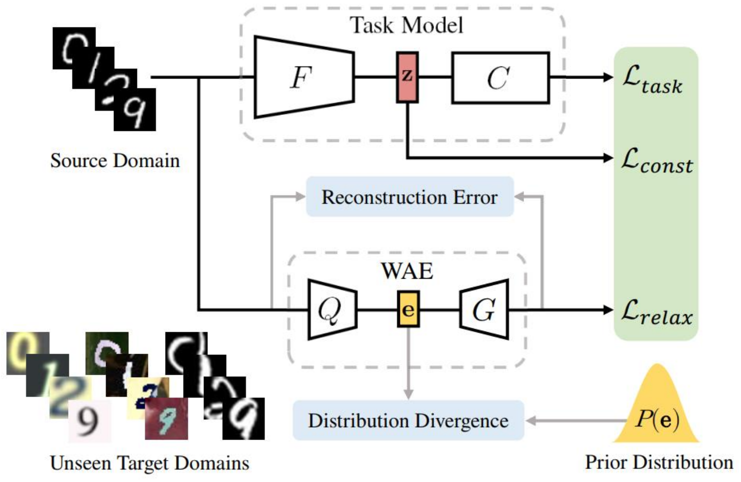
Generative adversarial net (GAN)



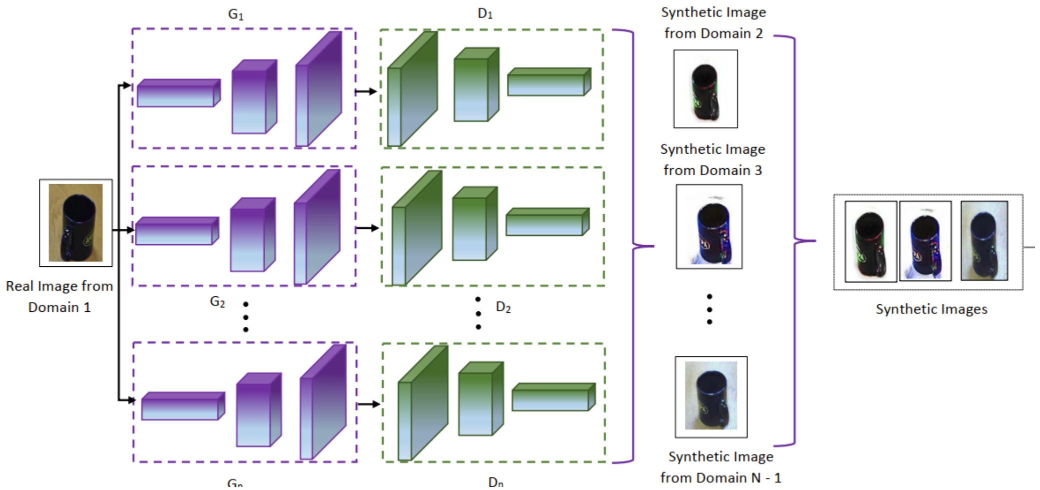
Mixup

- 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.

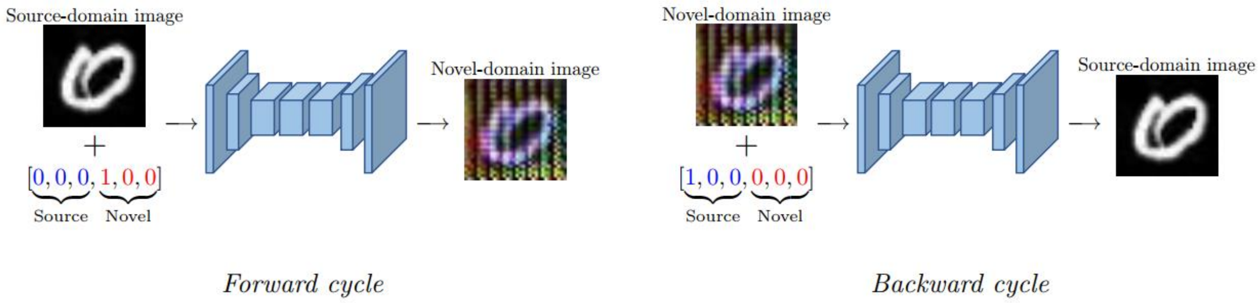
Data generation



VAE for generation



Multi-component generation



Conditional GAN for generation

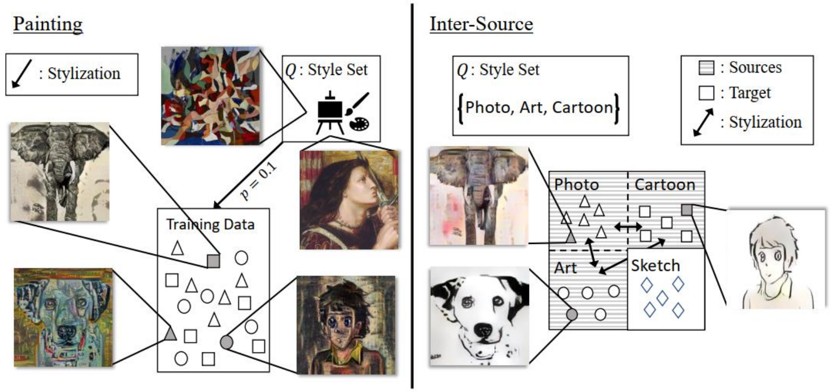
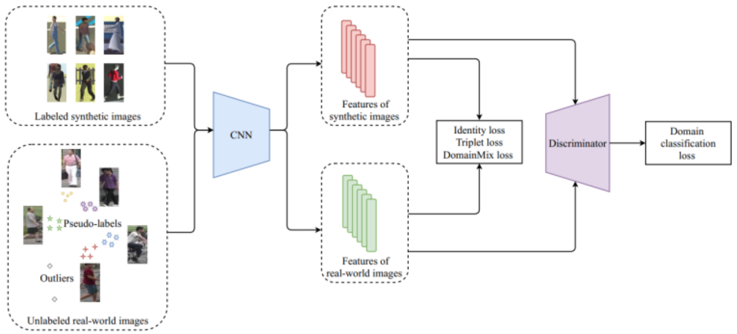


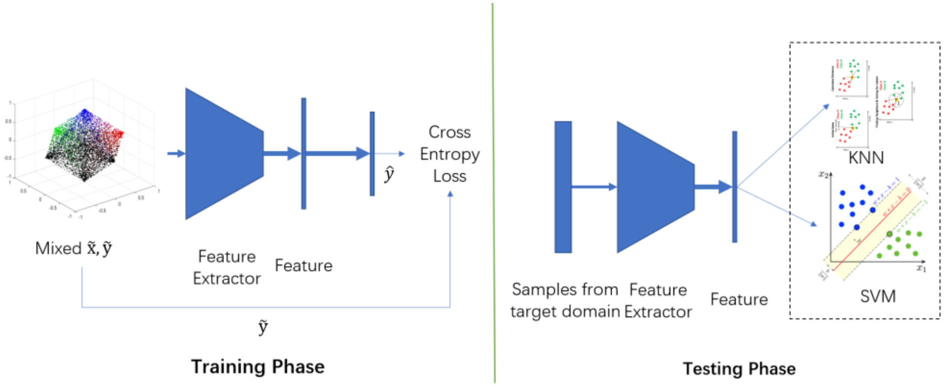
Image stylization

- 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.

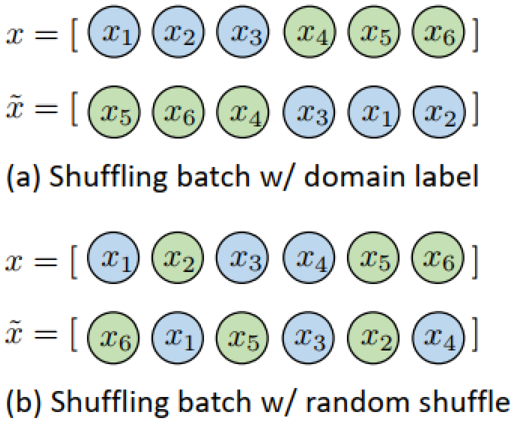
Mixup



DomainMix



MixAll

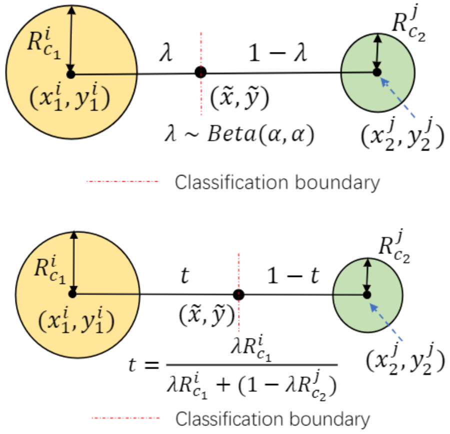


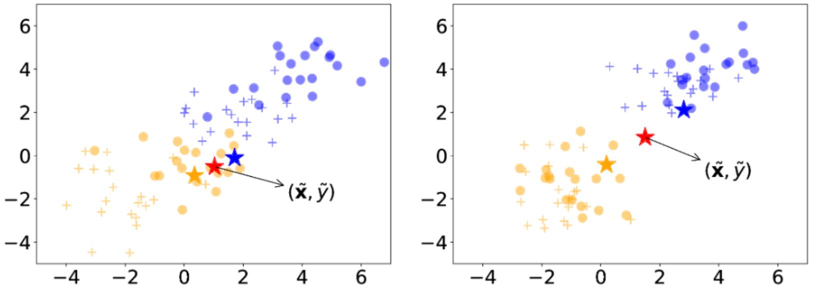
Style Mixup

- 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.

Data generation for DG

- Is vanilla Mixup enough for DG?
  - No.
    - Consider semantic range.
    - We also need a large margin.
  - **SDMix**: Semantic-Discriminative Mixup.



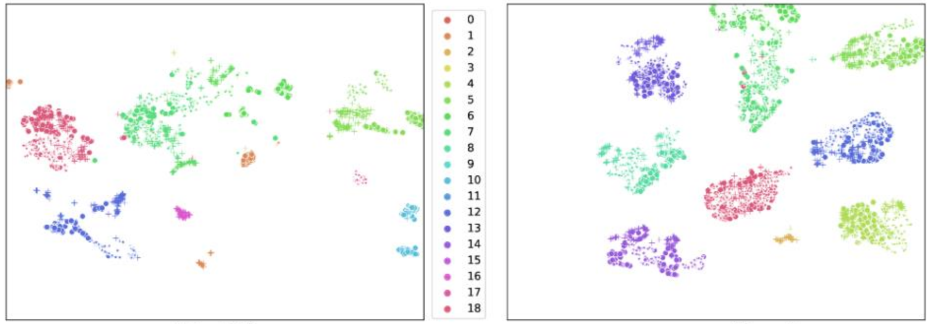


(a) Discrimination (vanilla Mixup)

(b) Discrimination (ours)

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}^i_1 + (1-\lambda)\mathbf{x}^j_2,$$
$$\tilde{y} = t y^i_1 + (1-t) y^j_2, \quad t = \frac{\lambda \times R^i_{c_1}}{\lambda \times R^i_{c_1} + (1-\lambda) \times R^j_{c_2}}$$
$$\lambda \sim Beta(\alpha, \alpha),$$

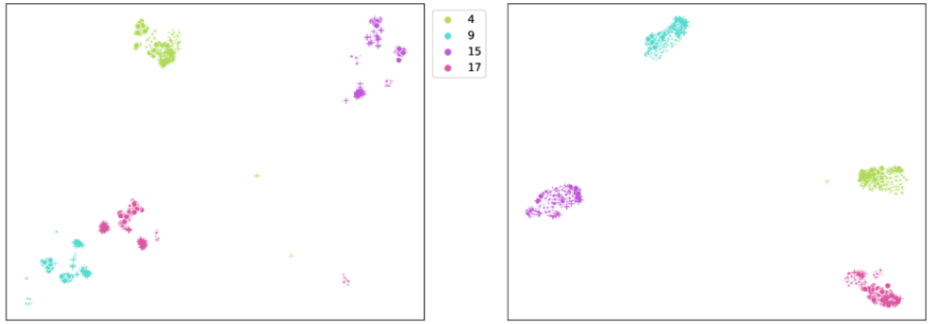
Semantic range



vanilla Mixup

our method

Discrimination



vanilla Mixup

our method

$$\ell_y(\mathbf{x}_k, y_k) = \mathscr{A}_{c \neq y_k} \max\{0, \gamma + u_{h,\mathbf{x}_k,\{c,y_k\}} \operatorname{sign}(h_c(\mathbf{x}_k) - h_{y_k}(\mathbf{x}_k))\}$$

Lu et al. Semantic-Discriminative Mixup for Generalizable Sensor-based Cross-domain Activity Recognition. ACM IMWUT, 2022.

Summary of data manipulation

- Advantages
  - Easy to understand and simple to implement
  - General to all kinds of data and networks
- Potential disadvantages
  - Lack of theoretical guarantee
  - Restricted by quality of training data

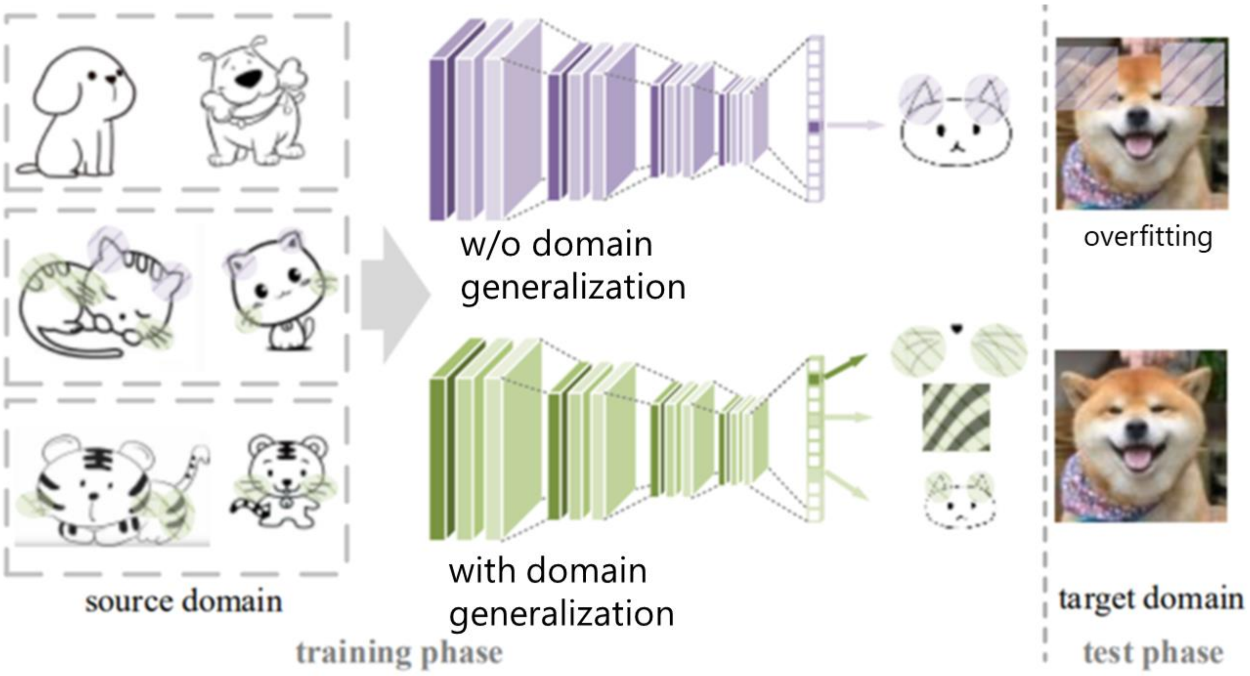
Representation learning for DG

Representation Learning

- Learning domain-invariant representations
  - Learning features which are expected to be better generalized to unseen target domain.

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

Classifier Feature Regularization



Representation learning

- How to learn generalized representations for DG?
  - 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

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

  - TCA: Transfer Component Analysis

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

- 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.

Kernel-based methods

- Marginal distribution adaptation
  - $Distance(D_s, D_t) \approx MMD(P_s(x), P_t(x), f)$

$$= \sup_{f \in \mathcal{F}} \mathbb{E}_P \left[\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{j=1}^{n} f(y_j)\right]$$

  - (raw version) $= \mathrm{tr}(\mathbf{A}^{\mathbf{T}}\mathbf{X}\mathbf{M}\mathbf{X}^{\mathbf{T}}\mathbf{A})$
  - (kernel version) $= \mathrm{tr}(\mathbf{KM})$ $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t] \in \mathbb{R}^{d\times(m+n)}, \mathbf{A} \in \mathbb{R}^{(m+n)\times(m+n)}$

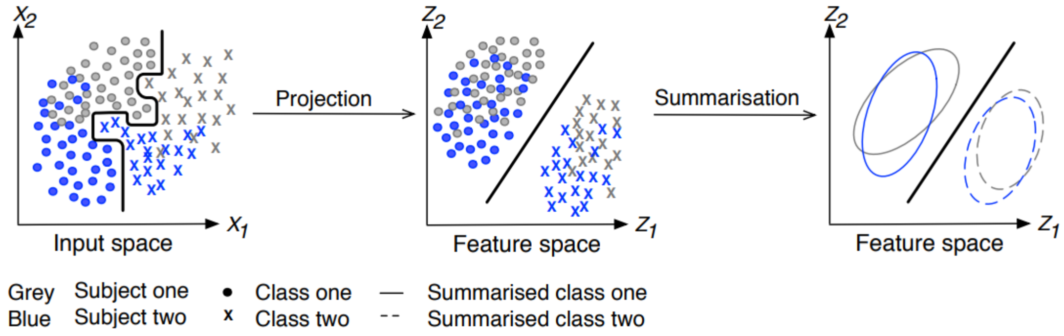
$$K = \begin{bmatrix} K_{s,s}, K_{s,t} \\ K_{t,s}, K_{t,t} \end{bmatrix} \quad M_{i,j} = \begin{cases} \frac{1}{m^2}, x_i, x_j \in D_s \\ \frac{1}{n^2}, x_i, x_j \in D_t \\ \frac{-1}{mn}, otherwise \end{cases}$$

$$\min \mathrm{tr}(\mathbf{KM}) - \lambda\, \mathrm{tr}(\mathbf{K})$$

[1] Pan et al. Domain adaptation via transfer component analysis. IEEE TNN 2011.

Kernel-based methods

- More than just distribution adaptation
  - ESRand: Elliptical Summary Randomisation (ESRand)
    - comprises of a randomised kernel and elliptical data summarization
    - projected each domain into an ellipse to represent the domain information and then used some similarity metric to compute the distance.



  - SCA: scatter component analysis
    - Adopted Fisher's discriminant analysis to minimize the discrepancy of representations from the same class and the same domain, and maximize the discrepancy of representations from the different classes and different domains

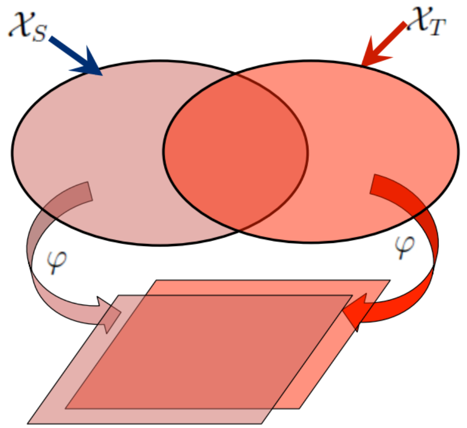
$$\Psi_\phi(\mathbb{P}) := \mathop{\mathbb{E}}_{x\sim\mathbb{P}}\left[\left\|\mu_{\mathbb{P}} - \phi(x)\right\|_{\mathcal{H}}^2\right]$$

$$\sup \frac{\{\text{total scatter}\} + \{\text{between-class scatter}\}}{\{\text{domain scatter}\} + \{\text{within-class scatter}\}}$$

- Erfani S, Baktashmotlagh M, Moshtaghi M, et al. Robust domain generalisation by enforcing distribution invariance. AAAI 2016.
- Ghifary et al. Scatter Component Analysis: A Unified Framework for Domain Adaptation and Domain Generalization. TPAMI 2017.

Explicit feature alignment

- Explicit distance: $R(\cdot,\cdot) \approx Distance(D_s, D_t)$
  - Goal: $f^* = \arg min_f \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) + \lambda \cdot Distance(D_s, D_t)$
  - Kernel-based distance
    - Maximum mean discrepancy (MMD) [1]
    - KL-divergence
    - Cosine similarity
  - Geometrical distance
    - Geodesic flow kernel (GFK) [2]
    - Correlation alignment (CORAL)[3]
    - Riemannian manifold [4]



[1] Pan et al. Domain adaptation via transfer component analysis. IEEE TNN 2011.
[2] Gong et al. Geodesic flow kernel for unsupervised domain adaptation. CVPR 2012.
[3] Sun et al. Return of frustratingly easy domain adaptation. AAAI 2016.
[4] Baktashmotlagh et al. Domain adaptation on statistical manifold. CVPR 2014.

Explicit distance

- Maximum mean discrepancy (MMD)
  - Given $x \sim P, y \sim Q$, f is a feature map: $x \to \mathcal{H}$, where $\mathcal{H}$ is reproducing kernel Hilbert space (RKHS), then

$$MMD(P,Q,\mathcal{F}) \coloneqq \sup_{f\in\mathcal{F}} \mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(y)]$$

  - Empirical estimate

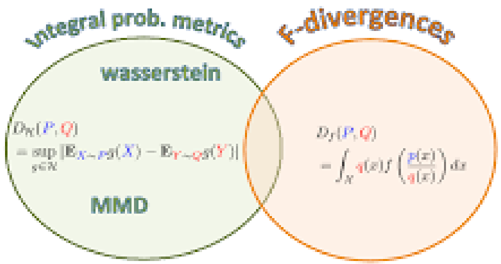
$$MMD(P,Q,\mathcal{F}) \coloneqq \sup_{f\in\mathcal{F}} \mathbb{E}_P\left[\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{j=1}^{n} f(y_j)\right]$$

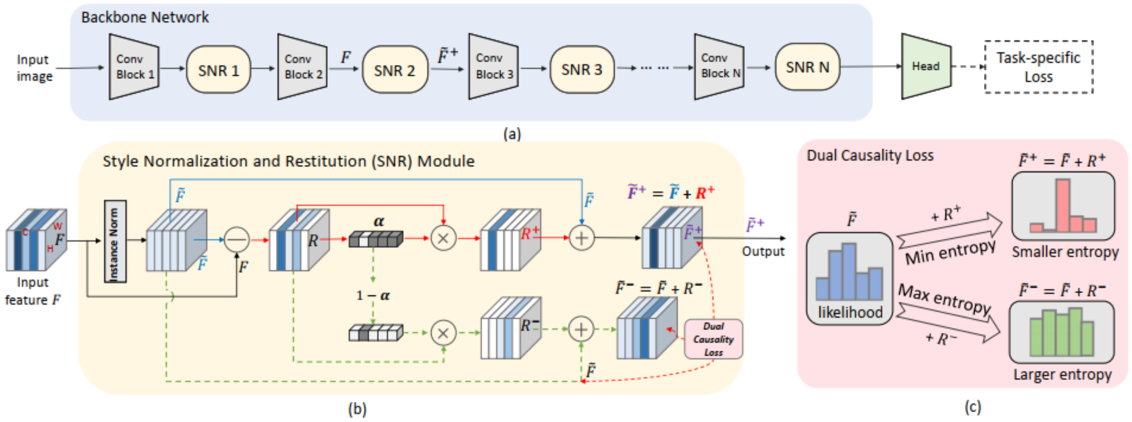
Theorem: $MMD(P,Q,F) = 0 \; iff \; P = Q$, when $\mathcal{F} = \{f| \; ||f||_{\mathcal{H}} \leq 1\}$ is a unit ball in a RKHS, provided that $\mathcal{H}$ is universal. [1]

[1] Alexander J. Smola. Maximum mean discrepancy. ICONIP 2016, Hong Kong. http://alex.smola.org/teaching/iconip2006/iconip_3.pdf

Explicit feature alignment

- Learning shareable information across domain
  - Maximum mean discrepancy: $\mathrm{MMD}(\mathcal{F}, P_X, P_Y) = \sup_{\|f\|_{\mathcal{H} \leq 1}} (\mathbb{E}_p(f(x)) - \mathbb{E}_p(f(y)))$
  - KL Divergence: $KL(q(\mathcal{Z}|\mathcal{X})||\mathcal{N} \sim (0,1))$
  - Correlation alignment: $\ell_{CORAL} = \frac{1}{4d^2} \|C_S - C_T\|_F^2$



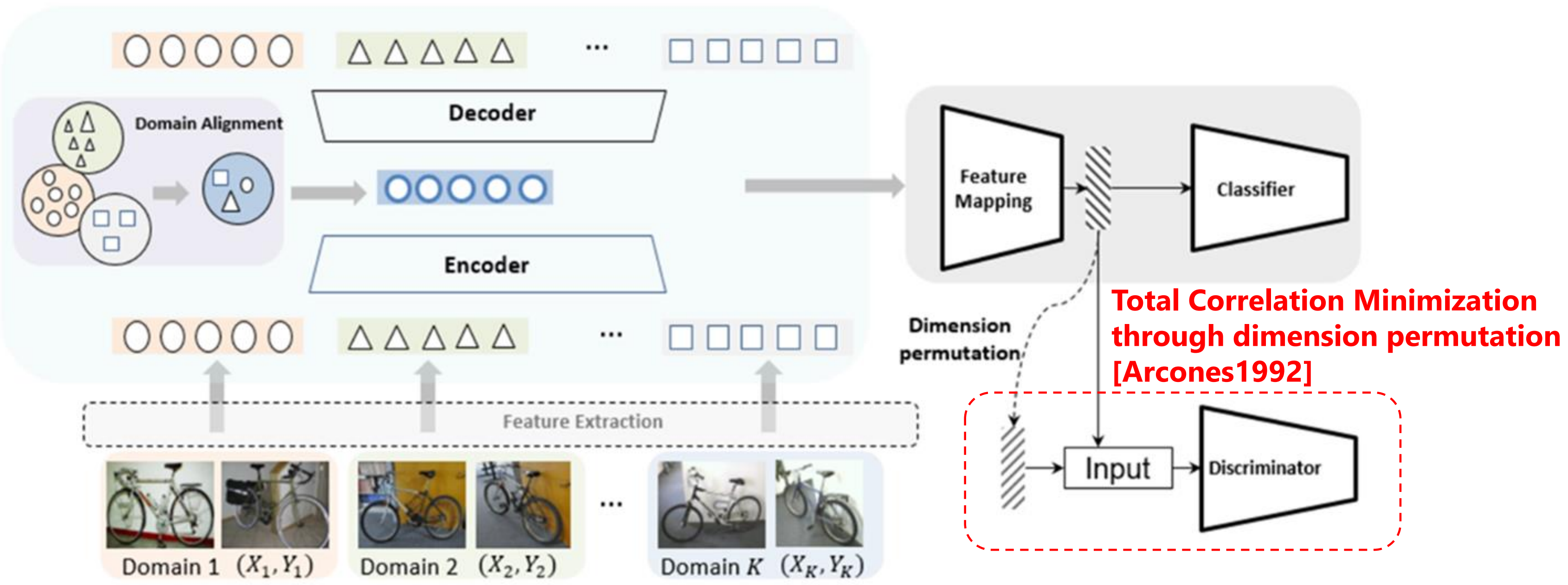


- Ya Li, et al., Deep domain generalization via conditional invariant adversarial networks, ECCV 2018
- Haoliang Li, et al., Domain Generalization for Medical Imaging Classification with Linear-Dependency, NeurIPS, 2020
- Jin X, Lan C, Zeng W, et al. Style Normalization and Restitution for Domain Generalization and Adaptation, Arxiv, 2021.

Multi-layer Feature Learning

- Feature disentanglement at deep layer.
  - Neuron independence regularization

$$P\left(H^1, H^2, \ldots, H^{d'}\right) = P(H^1)P(H^2)\ldots P(H^d)$$



[Arcones1992] M. A. Arcones and E. Gine, "On the bootstrap of u and v statistics," The Annals of Statistics, pp. 655–674, 1992.

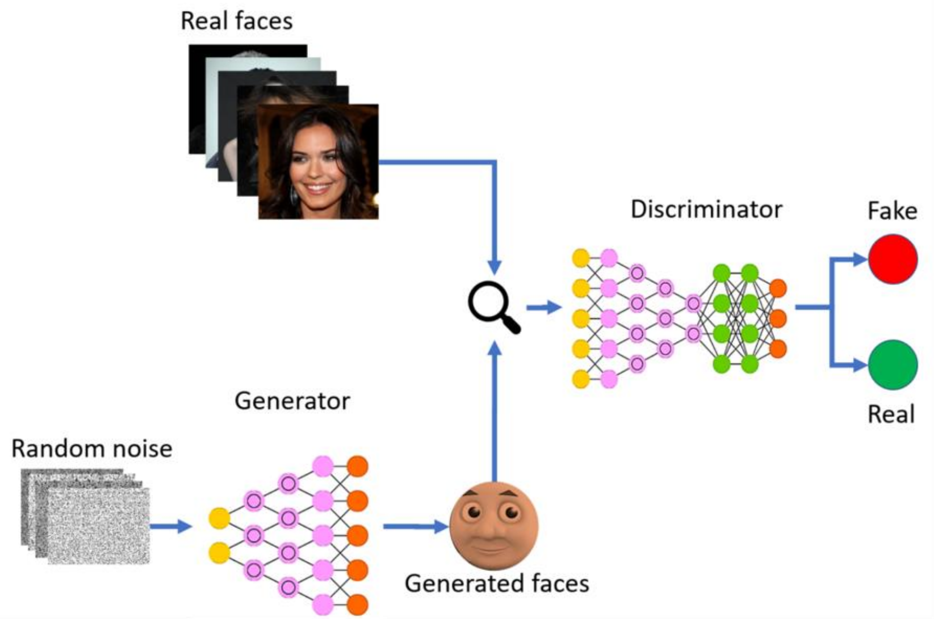
Domain-adversarial training

- Implicit distance: $R(\cdot,\cdot) \approx Separability(D_s, D_t)$
  - Goal: $f^* = \arg min_f \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) + \lambda \cdot Separability(D_s, D_t)$
  - How to measure $Separability(D_s, D_t)$?
    - Domain discriminator in generative adversarial nets (GAN) [1]
  - Objective: $\ell_{\text{adv}} = \mathbb{E}_{z \sim P(z)} \log\left(1 - D\big(G(z)\big)\right) + \mathbb{E}_{x \sim P_{\text{img}}(x)} \log D(x)$
    - Train: $\min_G \max_D \ell_{\text{adv}}$
  - How to use GAN for transfer?

[1] Ganin et al. Unsupervised domain adaptation by backpropagation. ICML 2015.
Figure: https://becominghuman.ai/generative-adversarial-networks-gans-human-creativity-2fc61283f3f6

GANs

- Generative adversarial nets
  - GAN -> transfer learning -> domain generalization



GAN	GAN-based DA	GAN-based DG
Real faces	Source domain	Domain 1
Random noise	Target domain	Domain 2...
Generator	Generator	Generator
Discriminator	Discriminator	Discriminator

Figure: https://medium.com/sigmoid/a-brief-introduction-to-gans-and-how-to-code-them-2620ee465c30

DANN

- Domain adversarial neural network (DANN)[1]
  - Feature extractor: $G_f(\cdot;\theta_f)$
  - Label predictor: $G_y(\cdot;\theta_y)$
  - Domain classifier: $G_d(\cdot;\theta_d)$

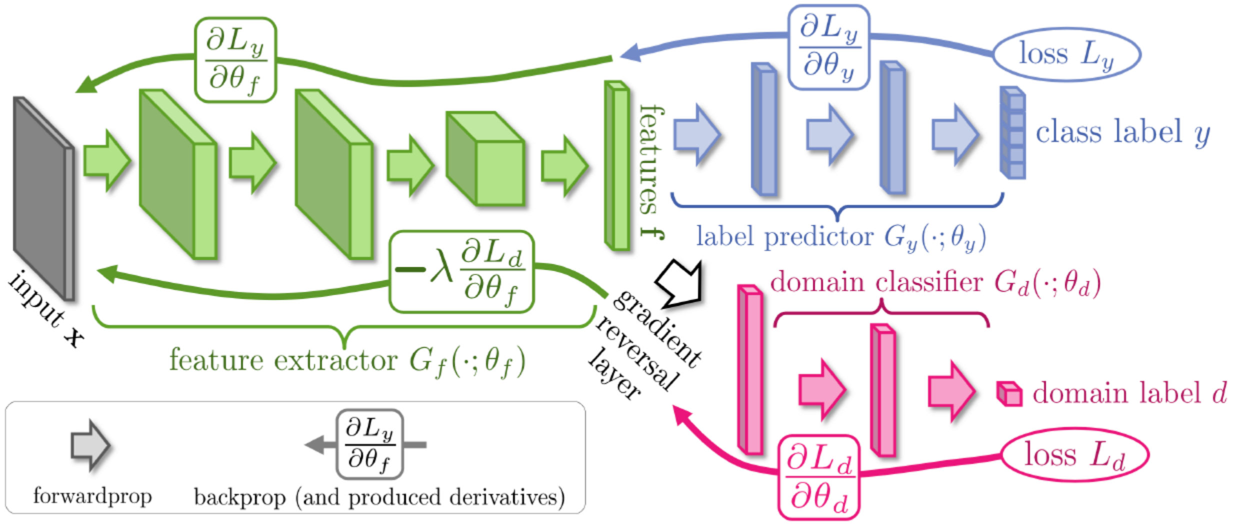


Figure: Ganin et al. Unsupervised domain adaptation by backpropagation. ICML 2015.

DANN

- Training of DANN
  - Objective:

$$E(\theta_f, \theta_y, \theta_d) = \underbrace{\sum_{\mathbf{x}_i \in \mathcal{D}_s} L_y\left(G_y\left(G_f(\mathbf{x}_i)\right), y_i\right)}_{\text{Classification loss}} - \lambda \underbrace{\sum_{\mathbf{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d\left(G_d\left(G_f(\mathbf{x}_i)\right), d_i\right)}_{\text{Separation loss}}$$

  - Learning:
    - Minimize feature extraction and classification loss

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

    - Maximize domain confusion

$$(\hat{\theta}_d) = \arg\max_{\theta_d} E(\theta_f, \theta_y, \theta_d)$$

  - Stochastic gradient descent
    - Problem: λ is hard to implement in SGD

$$\theta_f \longleftarrow \theta_f - \mu\left(\frac{\partial L_y^i}{\partial \theta_f} - \lambda \frac{\partial L_d^i}{\partial \theta_f}\right)$$

$$\theta_y \longleftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y}$$

$$\theta_d \longleftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_d}$$

DANN

- Train DANN in SGD
  - Gradient reversal layer (GRL)
    - Forward propagation: GRL is an identity map

$$R_\lambda(x) = x$$

    - Backward propagation: take gradient from subsequent level, and $\times\ (-\lambda)$

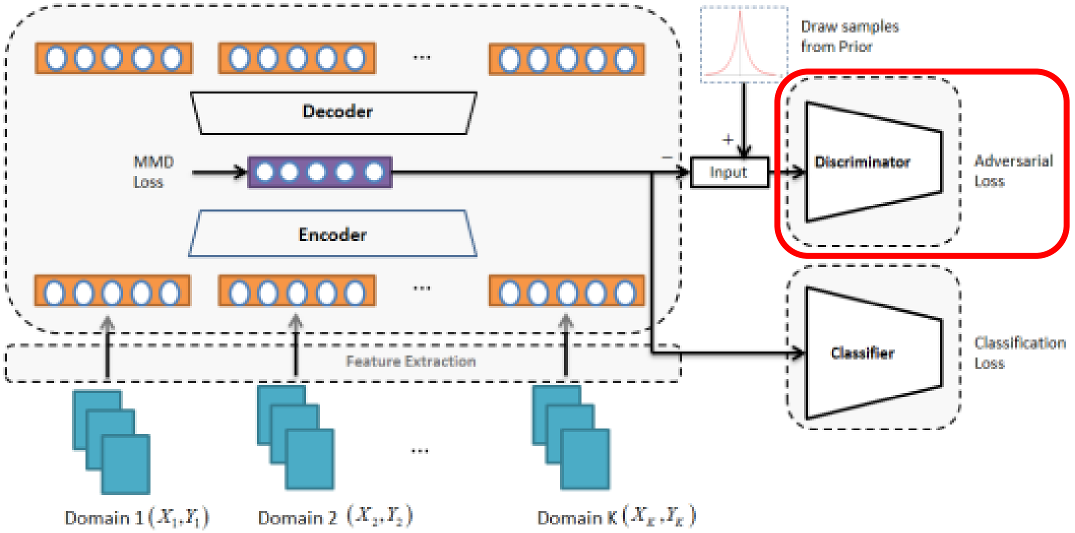
$$\frac{dR_\lambda}{dx} = -\lambda \mathbf{I}$$

    - Transformed objective function:
  - GRL can be easily implemented in Pytorch/Tensorflow/Caffe…

$$E(\theta_f, \theta_y, \theta_d) = \sum_{\mathbf{x}_i \in \mathcal{D}_s} L_y\left(G_y\left(G_f\left(\mathbf{x}_i\right)\right), y_i\right) + \lambda \sum_{\mathbf{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d\left(G_d\left(G_f\left(\mathbf{x}_i\right)\right), d_i\right)$$

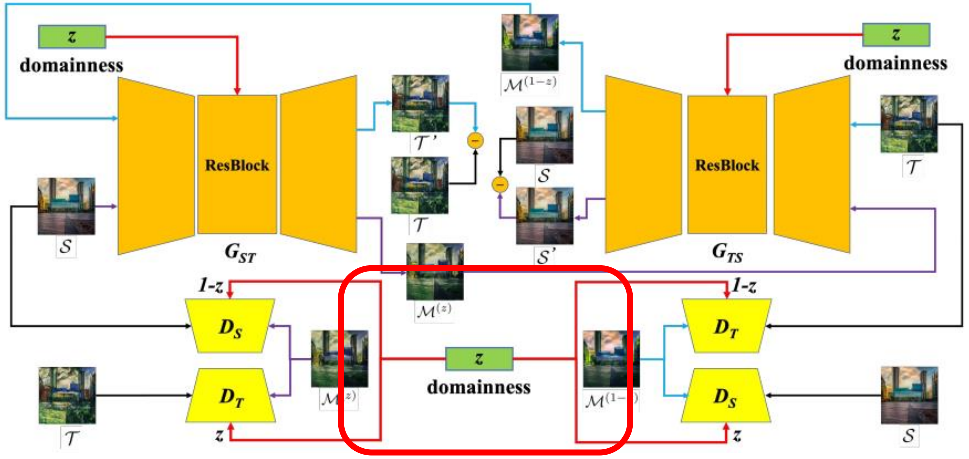
Code of DANN: http://github.com/jindongwang/transferlearning/code/deep/DANN

Domain-adversarial learning for DG



MMD-AAE

$$\min_{Q,P} \max_{D} \mathcal{L}_{\text{ae}} + \lambda_1 \mathcal{R}_{\text{mmd}} + \lambda_2 \mathcal{J}_{\text{gan}}$$



DLOW

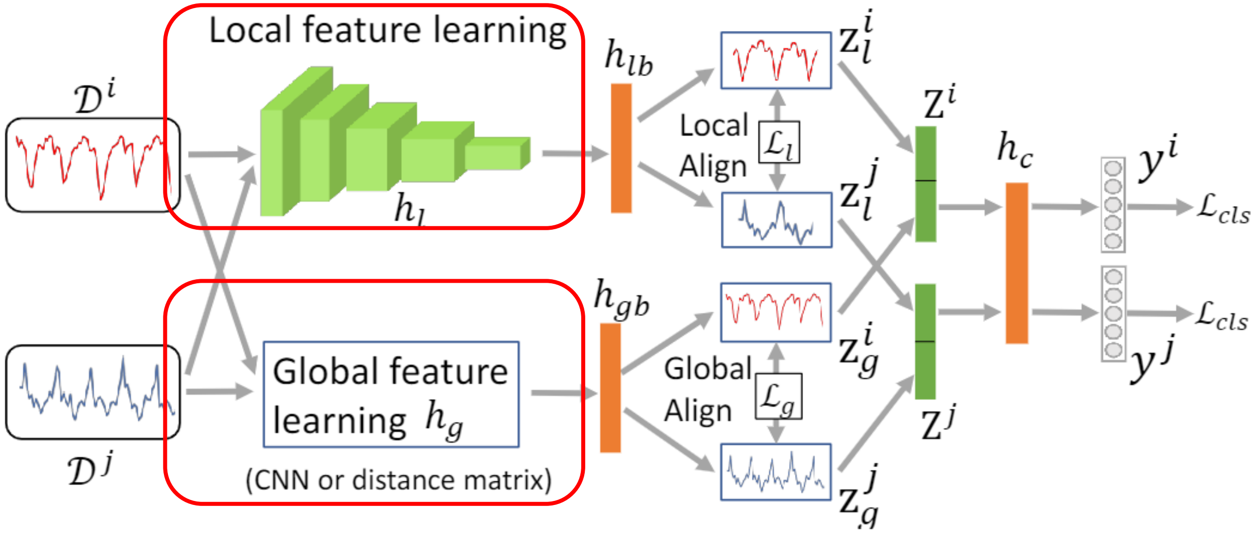
$$\mathcal{L}_{adv}(G_{ST}, D_S) = \mathbb{E}_{\mathbf{x}^s \sim P_S}[\log(D_S(\mathbf{x}^s))] + \mathbb{E}_{\mathbf{x}^s \sim P_S}[\log(1 - D_S(G_{ST}(\mathbf{x}^s, z)))]$$

$$\mathcal{L}_{adv}(G_{ST}, D_T) = \mathbb{E}_{\mathbf{x}^t \sim P_T}[\log(D_T(\mathbf{x}^t))] + \mathbb{E}_{\mathbf{x}^s \sim P_S}[\log(1 - D_T(G_{ST}(\mathbf{x}^s, z)))].$$

- Haoliang Li et al. Domain Generalization with Adversarial Feature Learning. CVPR 2018.
- Rui Gong et al. DLOW: Domain Flow for Adaptation and Generalization. CVPR 2019.

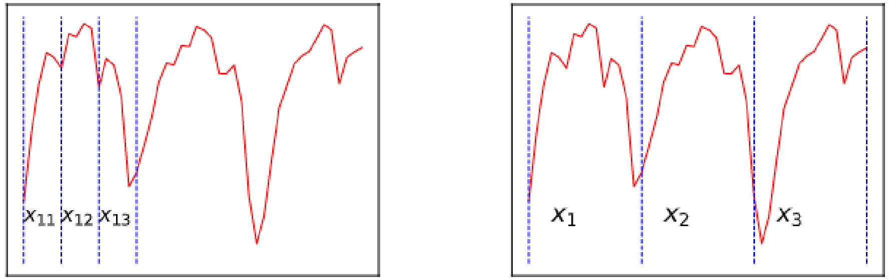
Domain-adversarial learning for DG

- Is local alignment enough for DG?
  - No. Ignore some big picture features.
  - We also need a global alignment.
  - **LAG**: Local and Global Alignment.



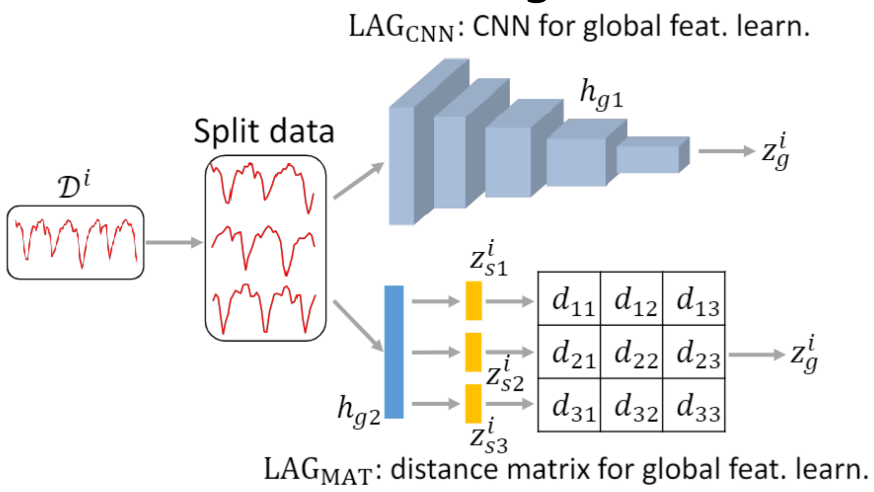
$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_l + \lambda_2 \mathcal{L}_g$$

Data of walking activity



(a) Local correlation (b) Global correlation

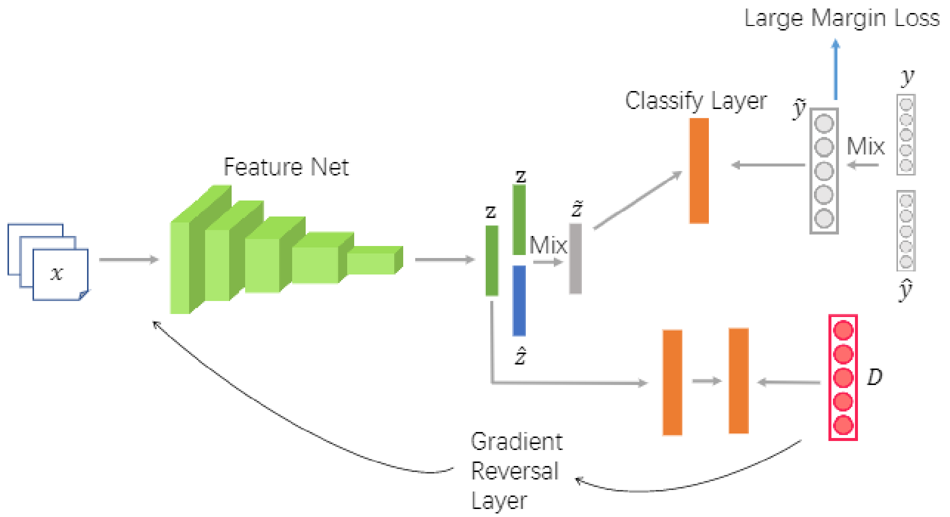
Global feature learning

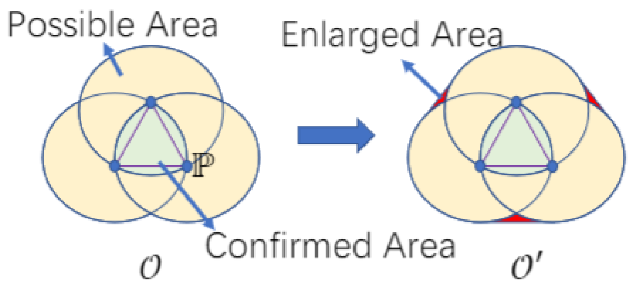


Lu et al. Local and global alignments for generalizable sensor-based human activity recognition. ICASSP 2022.

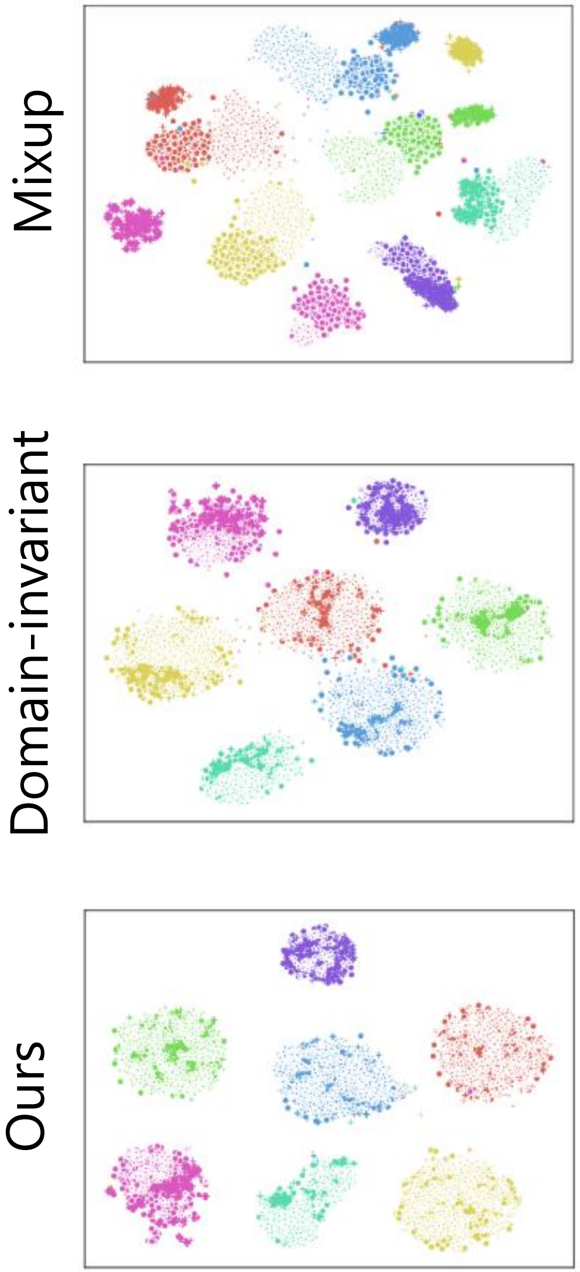
Representation augmentation for DG

- Is vanilla Mixup or simple alignment enough?
  - No. Domain-invariant feature Mixup.
  - FIXED: Domain-invariant Feature MIXup with Enhanced Discrimination.





Enlarged distribution cover range



$$\min \mathbb{E}_{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2)\sim\mathbb{P}}\mathbb{E}_{\lambda\sim Beta(\alpha,\alpha)}[\ell_{lm}(G_y(\mathrm{Mix}_\lambda(\mathbf{z}_1,\mathbf{z}_2)),\mathrm{Mix}_\lambda(y_1,y_2))+\ell_d(G_d(R_\eta(\mathbf{z}_1)),D)],$$

Lu et al. FIXED: Frustratingly Easy Domain Generalization with Mixup. Under Review, 2022.

Invariant risk minimization

- IRM
  - Do not seek to match the representation distribution of all domains, but to enforce the optimal ***classifier*** on top of the representation space to be the same across all domains
  - The intuition is that the ideal representation for prediction is the cause of y, and the causal mechanism should not be affected by other factors/mechanisms, thus is domain-invariant.

$$\min_{\substack{g \in \mathcal{G}, \\ f \in \bigcap_{i=1}^{M} \arg\min_{f' \in \mathcal{F}} \epsilon^i(f' \circ g)}} \sum_{i=1}^{M} \epsilon^i(f \circ g) \xrightarrow{\text{Learn } g} \min_{g \in \mathcal{G}} \sum_{i=1}^{M} \epsilon^i(g) + \lambda \left\| \nabla_f \epsilon^i(f \circ g)\big|_{f=1} \right\|^2$$

Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019.

Feature disentanglement

- What is disentanglement
  - Learn a function that maps a sample to a feature vector, which contains all the information about **different factors of variation** and each dimension (or a subset of dimensions) contains information about only some factor(s).
- Formulation

Common features Specific features

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

Reconstruction error

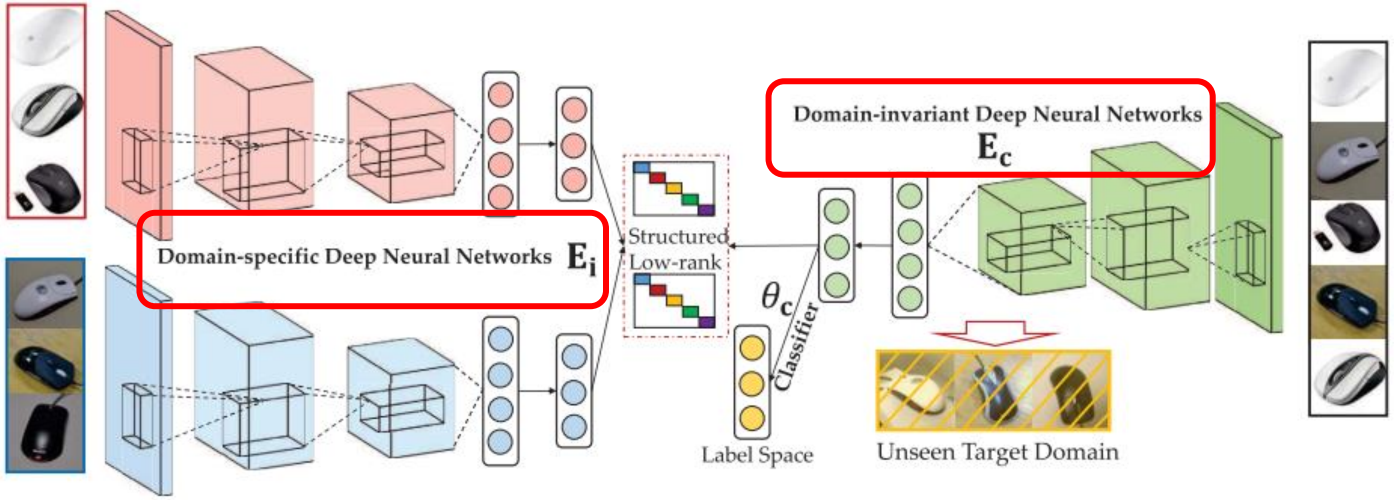
Feature disentanglement
- Multi-component analysis
- Generative modeling

Multi-component analysis

- UndoBias
  - Weights can be disentangled into: common and specific weights

$$\mathbf{w}_i = \mathbf{w}_0 + \Delta_i$$

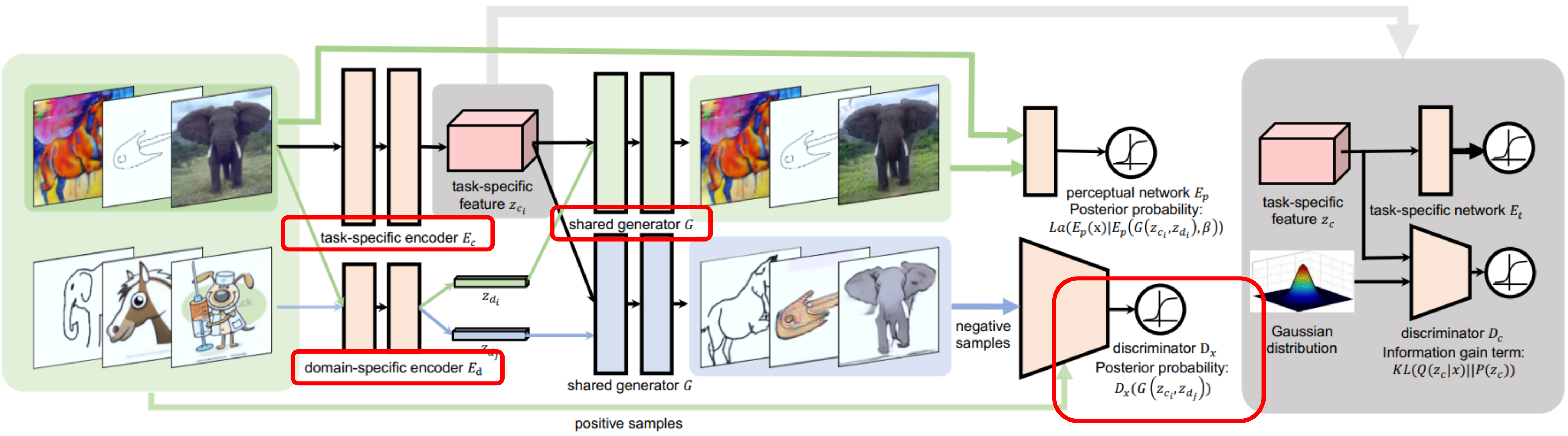
- Structure low-rank DG



- Khosla A, Zhou T, Malisiewicz T, et al. Undoing the damage of dataset bias. ECCV 2012.
- Ding Z, Fu Y. Deep domain generalization with structured low-rank constraint. IEEE TIP 2017.

Feature disentanglement

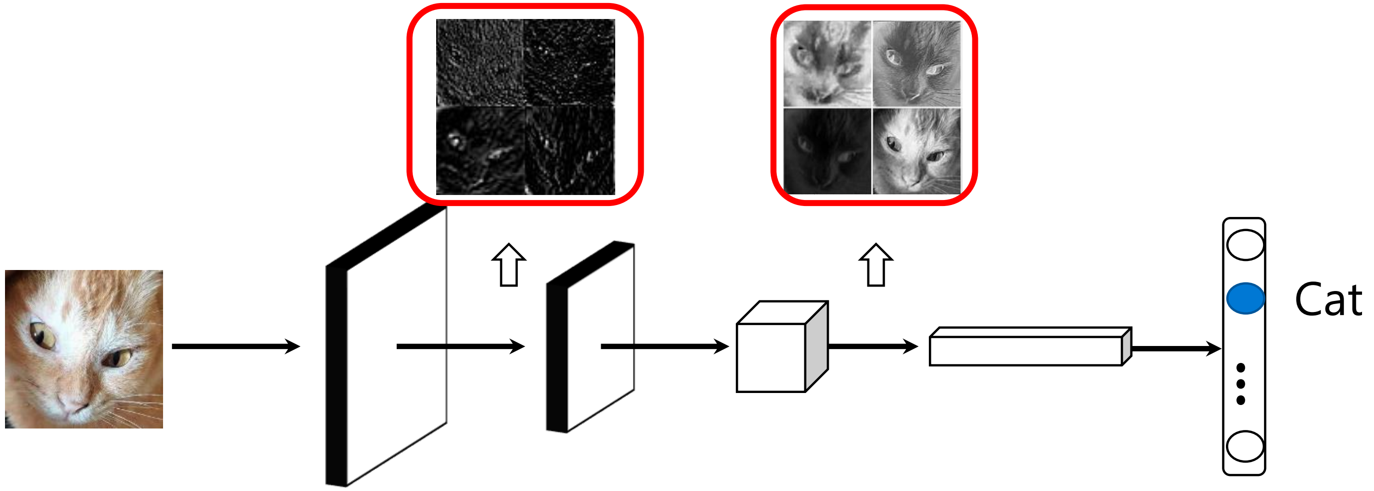
Invariant feature learning + style transfer



Yufei Wang, et al., Variational Disentanglement for Domain Generalization, Arxiv 2021

Multi-layer Feature Learning

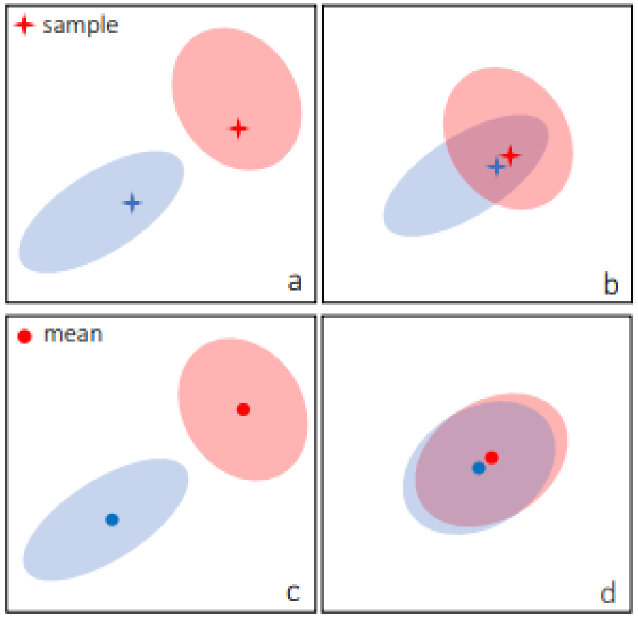
- Deep features eventually transit from general to specific along the network.
- Shallow Layer extracts shareable information while deep layer extracts category specific information (with regularization).



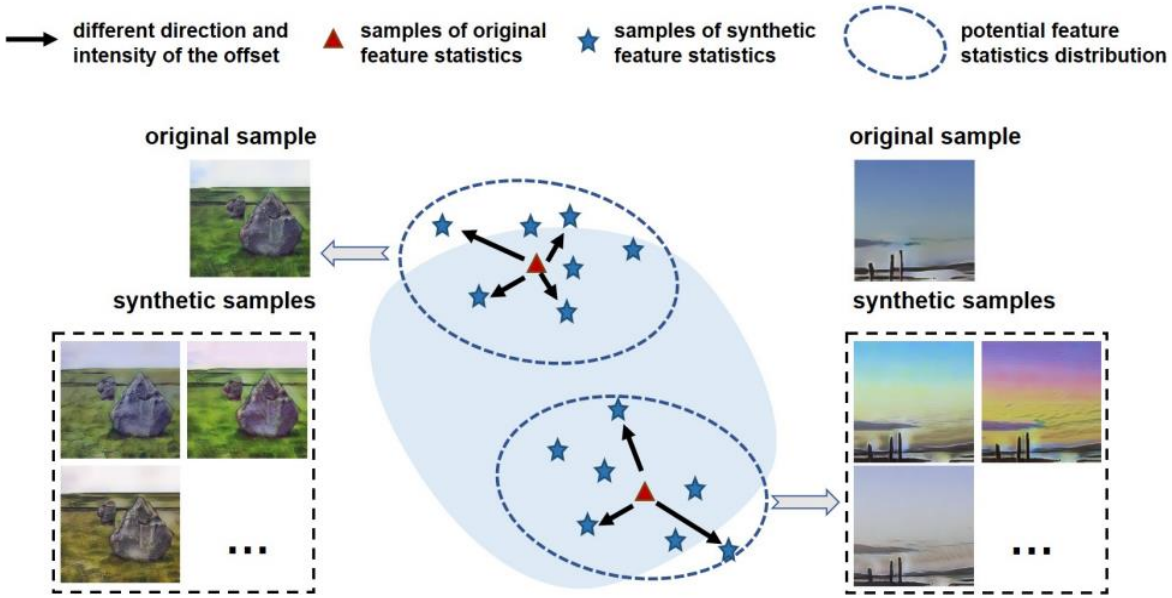
- Haoliang Li, *et.al.*, "GMFAD: Towards Generalized Visual Recognition via Multi-Layer Feature Alignment and Disentanglement", T-PAMI 2020

Domain-Invariant Learning with Uncertainty

- Uncertainty should be considered during domain-invariant learning.



Bayesian Neural Network



Uncertainty modeling through re-parameterization trick

- Zehan Xiao, et al., A Bit More Bayesian: Domain-Invariant Learning with Uncertainty , ICML'21
- Xiaotong Li, et al., Uncertainty Modeling for Out-of-Distribution Generalization." ICLR'22.

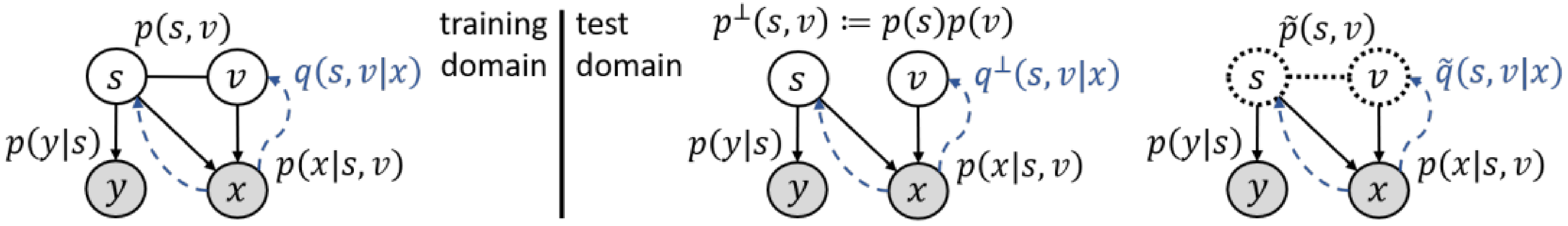
Generative modeling

- DIVA: domain-invariant variational-autoencoder

Domain-input-label

$$\mathcal{L}_s(d,\mathbf{x},y) = \mathbb{E}_{q_{\phi_d}(\mathbf{z}_d|\mathbf{x})q_{\phi_x}(\mathbf{z}_x|\mathbf{x}),q_{\phi_y}(\mathbf{z}_y|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\boxed{\mathbf{z}_d,\mathbf{z}_x,\mathbf{z}_y})\right] - \beta KL\left(q_{\phi_d}(\mathbf{z}_d|\mathbf{x})||p_{\theta_d}(\mathbf{z}_d|d)\right)$$
$$-\beta KL\left(q_{\phi_x}(\mathbf{z}_x|\mathbf{x})||p(\mathbf{z}_x)\right) - \beta KL\left(q_{\phi_y}(\mathbf{z}_y|\mathbf{x})||p_{\theta_y}(\mathbf{z}_y|y)\right).$$

- CSG: Causal semantic generative model



(a) CSG (b) CSG-ind (c) CSG-DA

S: semantic factor
V: variation factor

$$\mathbb{E}_{p^*(x)}\mathbb{E}_{p^*(y|x)}[\log q(y|x)] + \mathbb{E}_{p^*(x)}\mathbb{E}_{q(s,v,y|x)}\left[\frac{p^*(y|x)}{q(y|x)}\log\frac{p(s,v,x,y)}{q(s,v,y|x)}\right]$$

- Liu et al, Learning Causal Semantic Representation for Out-of-Distribution Prediction. NeurIPS 2021.
- Ilse M, Tomczak J M, Louizos C, et al. Diva: Domain invariant variational autoencoders[C]//Medical Imaging with Deep Learning. PMLR, 2020: 322-348.

Summary of representation learning

- Advantages
  - General and popular
  - Better performance
  - Some theoretical guarantee
- Potential disadvantages
  - Still difficult to remove spurious features
  - Data-driven

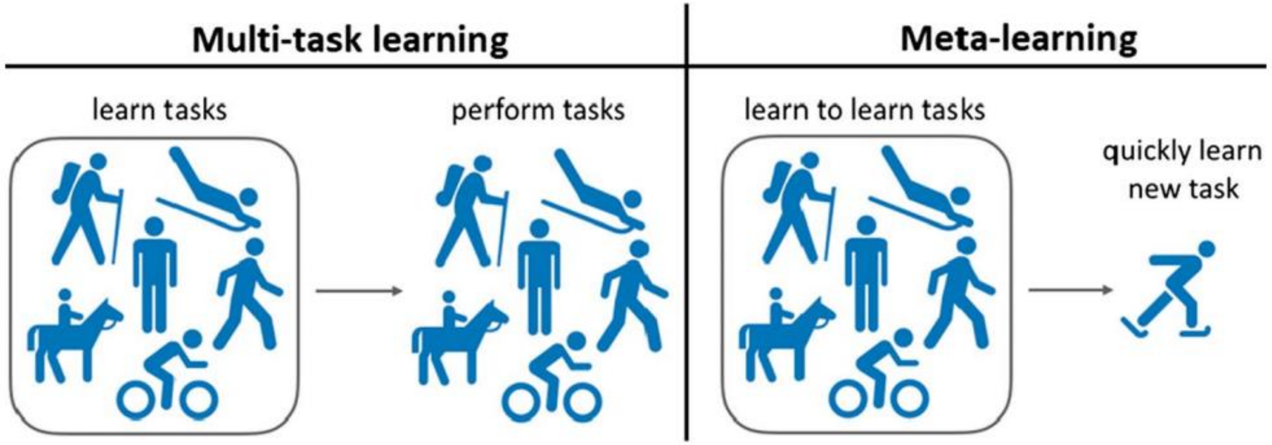
Learning strategy for DG

Different learning strategy for DG

- Meta-learning
  - Divide domains into several tasks, then use meta-learning to learn general knowledge
- Ensemble learning
  - Design ensemble models
- Gradient operation
  - Alter the gradient interaction between domains
- Distributionally robust optimization
  - Acquire models that are better for worst-case distribution scenario
- Self-supervised learning
- Others

Meta-learning

- Learning to learn, or meta-learn the general knowledge
  - Instead of the original tasks, meta-learning wants to acquire knowledge about **new tasks**



Meta-knowledge acquisition

$$\phi^* = \arg\max_{\phi} \log P\left(\phi \mid \mathcal{D}_{\text{src}}\right)$$

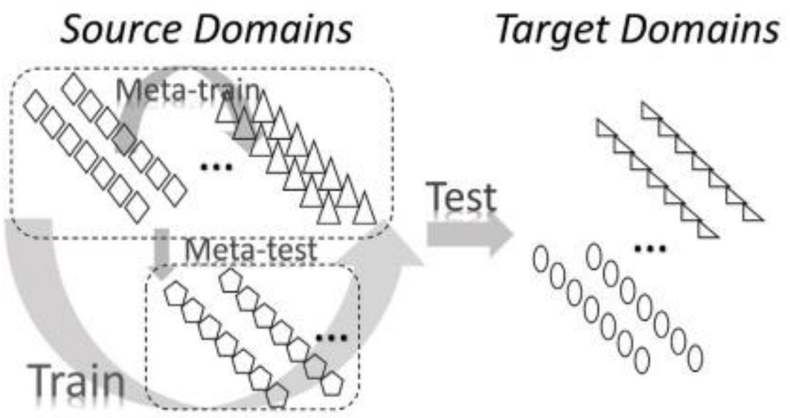
Meta-knowledge validation

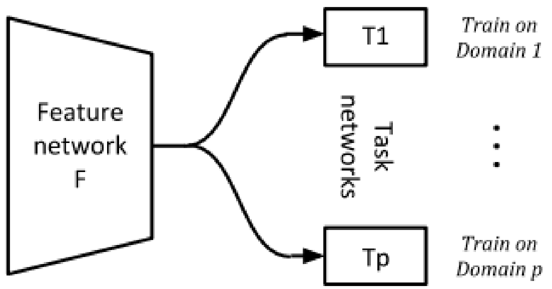
$$\theta^{*(i)} = \arg\max_{\theta} \log P\left(\theta \mid \phi^*, \mathcal{D}_{\text{tar}}^{\text{train }(i)}\right)$$

Huisman M, Van Rijn J N, Plaat A. A survey of deep meta-learning[J]. Artificial Intelligence Review, 2021, 54(6): 4483-4541.

Meta-learning for DG

- How to adopt meta-learning for DG?
  - Key: Old tasks to new tasks in meta-learning → Old domains to new domains
- MLDG: Meta-learning for DG
- MetaReg: meta-learning for regularization





Algorithm 1 Meta-Learning Domain Generalization

procedure MLDG
  Input: Domains $\mathcal{S}$
  Init: Model parameters Θ. Hyperparameters α, β, γ.
  for ite **in** iterations **do**
    Split: $\bar{\mathcal{S}}$ and $\breve{\mathcal{S}} \leftarrow \mathcal{S}$
    Meta-train: Gradients $\nabla_\Theta = \mathcal{F}'_\Theta(\bar{\mathcal{S}}; \Theta)$
    Updated parameters $\Theta' = \Theta - \alpha \nabla_\Theta$
    Meta-test: Loss is $\mathcal{G}(\breve{\mathcal{S}}; \Theta')$.
    Meta-optimization: Update Θ

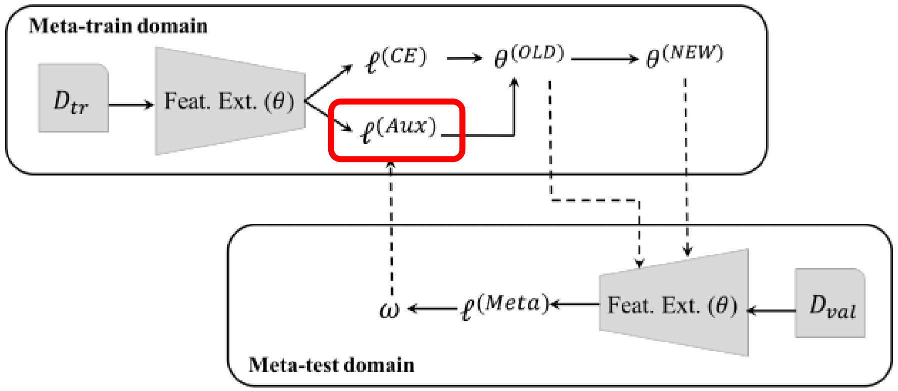
$$\Theta = \Theta - \gamma \frac{\partial(\mathcal{F}(\bar{\mathcal{S}}; \Theta) + \beta \mathcal{G}(\breve{\mathcal{S}}; \Theta - \alpha \nabla_\Theta))}{\partial \Theta}$$

  end for
end procedure

- Li D, Yang Y, Song Y Z, et al. Learning to generalize: Meta-learning for domain generalization. AAAI 2018.
- Balaji Y, Sankaranarayanan S, Chellappa R. Metareg: Towards domain generalization using meta-regularization. NeurIPS 2018.

Meta-learning for DG

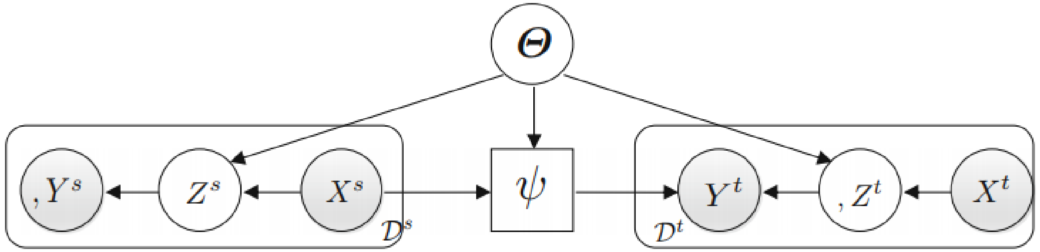
- Feature-critic training
  - Learning the regularization terms using meta-learning



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

$$\max_{\omega} \sum_{D_j \in \mathcal{D}_{\text{val}}} \sum_{d_j \in D_j} \tanh(\gamma(\theta^{(\text{NEW})}, \phi_j, x^{(j)}, y^{(j)}) \\ -\gamma(\theta^{(\text{OLD})}, \phi_j, x^{(j)}, y^{(j)}))$$

- Meta-VIB
  - Meta variational information bottleneck to model uncertainty between domain shifts

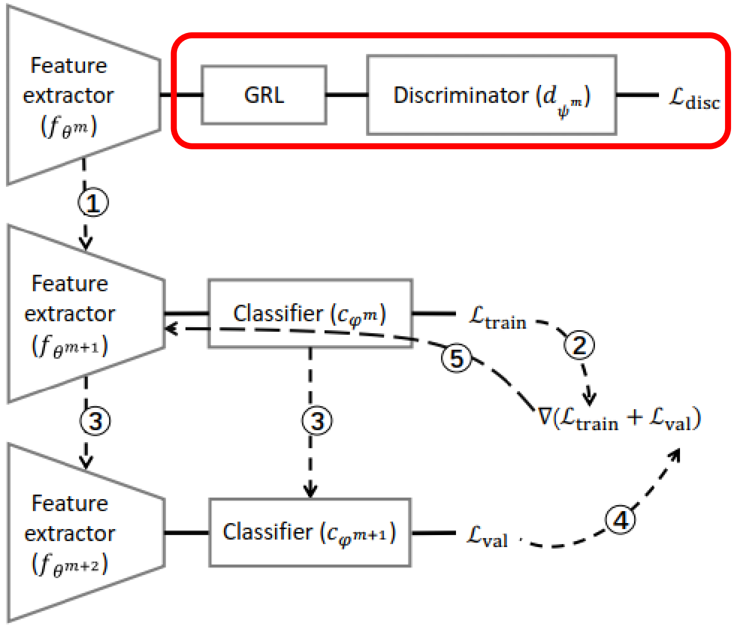


$$\tilde{\mathcal{L}}_{\text{MetaVIB}} = \frac{1}{N} \sum_{n=1}^{N} \int [p(\mathbf{z}_n|\mathbf{x}_n) p(\psi|D^s) \log q(\mathbf{y}_n|\mathbf{z}_n, \psi) \\ - \beta \boxed{p(\mathbf{z}_n|\mathbf{x}_n) \log \frac{p(\mathbf{z}_n|\mathbf{x}_n)}{q(\mathbf{z}_n|D^s)}] \mathrm{d}\mathbf{z}_n \mathrm{d}\psi.}$$

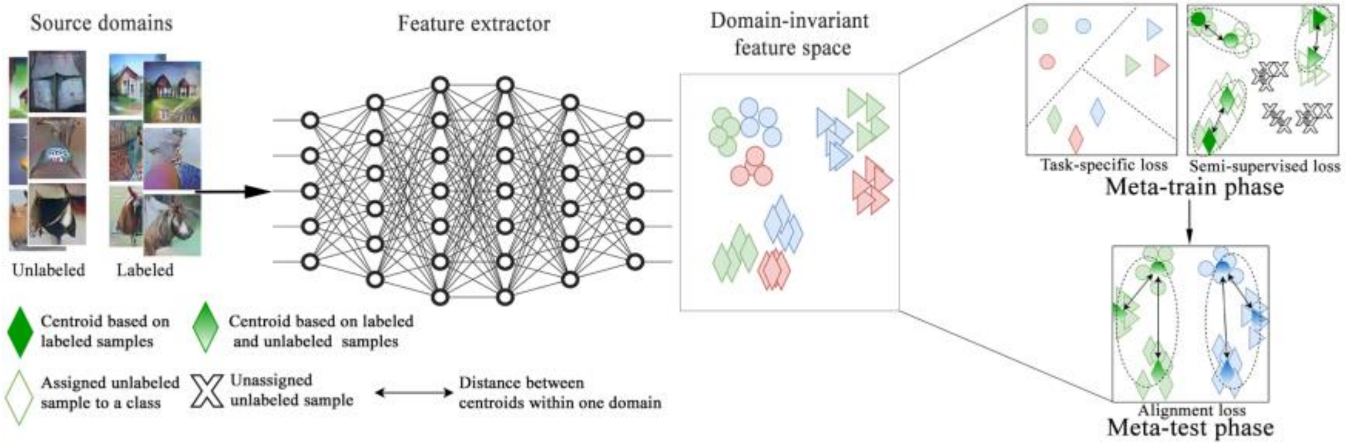
Li Y, Yang Y, Zhou W, et al. Feature-critic networks for heterogeneous domain generalization. ICML 2019.
Du Y, Xu J, Xiong H, et al. Learning to learn with variational information bottleneck for domain generalization. ECCV 2020.

Meta-learning for DG

- DADG: MLDG with adversarial training



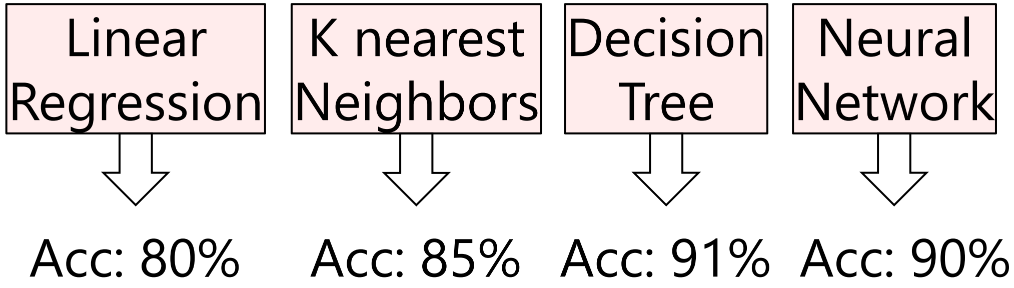
- DGSML: MLDG with semi-supervised learning

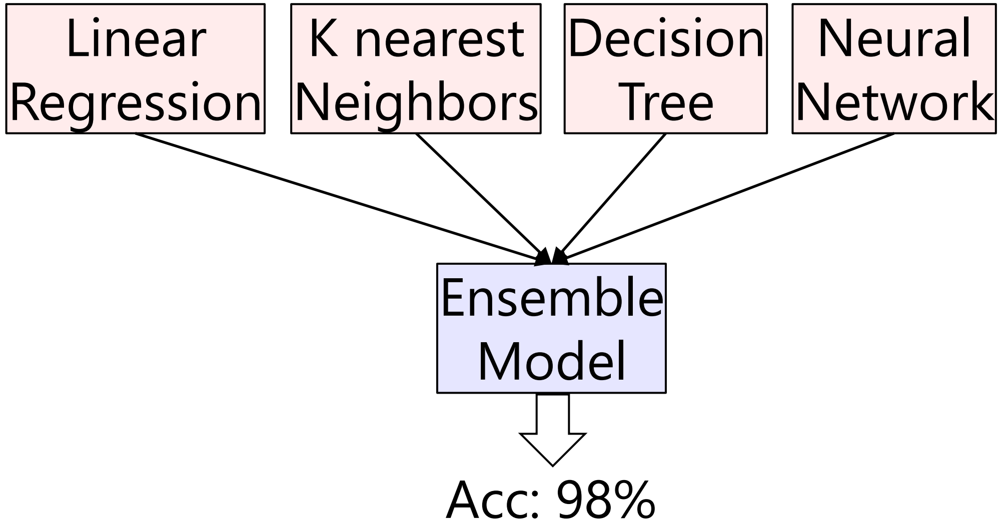


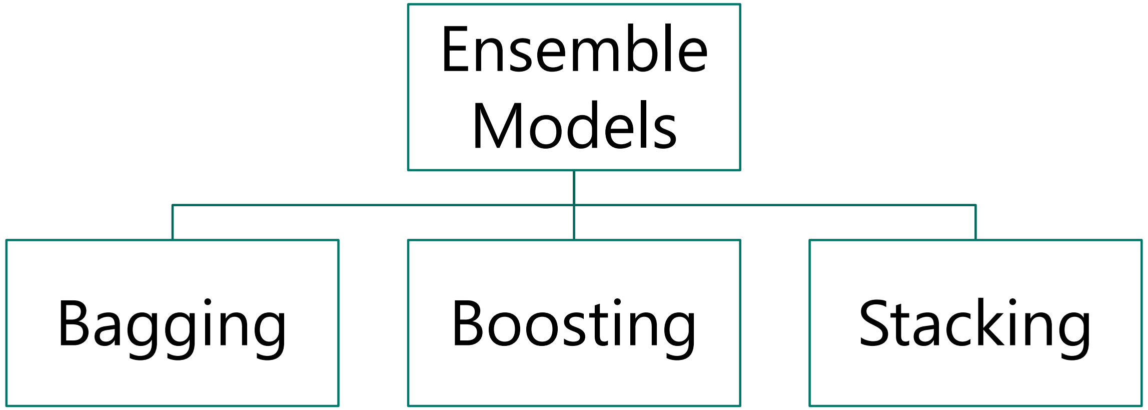
- Chen K, Zhuang D, Chang J M. Discriminative adversarial domain generalization with meta-learning based cross-domain validation. Neurocomputing 2022.
- Sharifi-Noghabi H, Asghari H, Mehrasa N, et al. Domain generalization via semi-supervised meta learning[J]. arXiv preprint arXiv:2009.12658, 2020.

Ensemble learning

- Is a single model or representation enough for generalization?







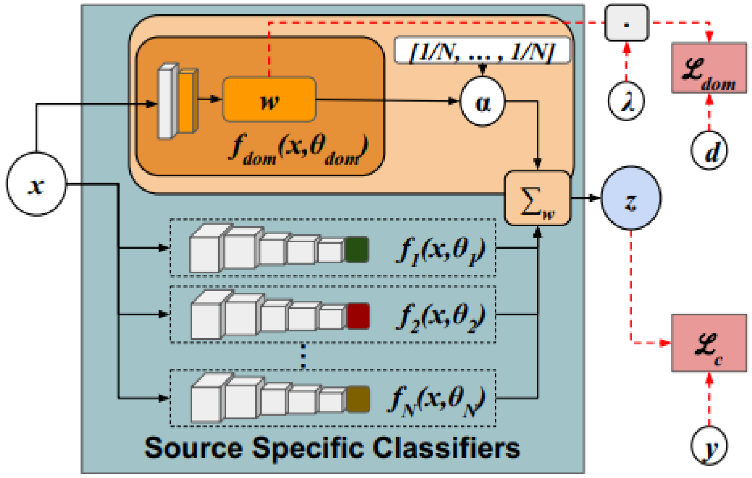
- Ensemble learning allows for more diversities in feature and classifier learning
- *The power from the crowd*

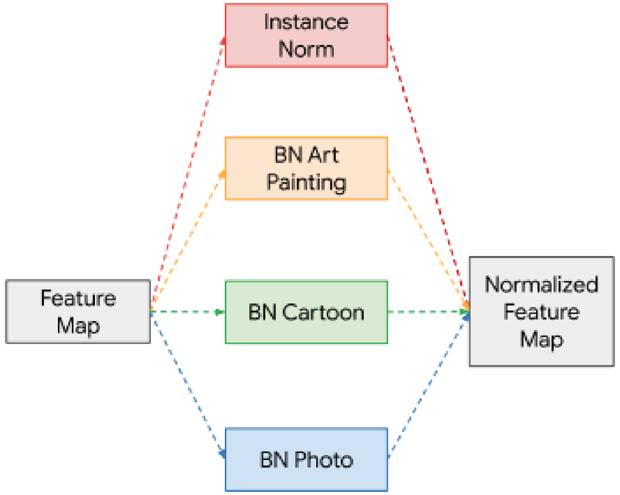
Ensemble learning for DG

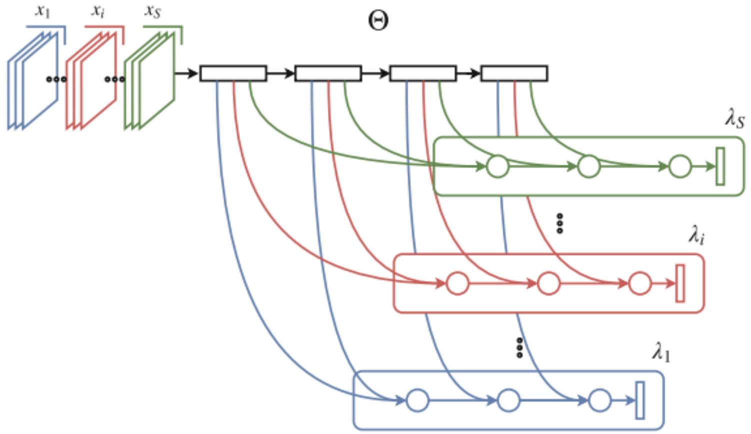
- Ensemble-learned DG representations
  - Feature weighting

Feature combination

Feature attention





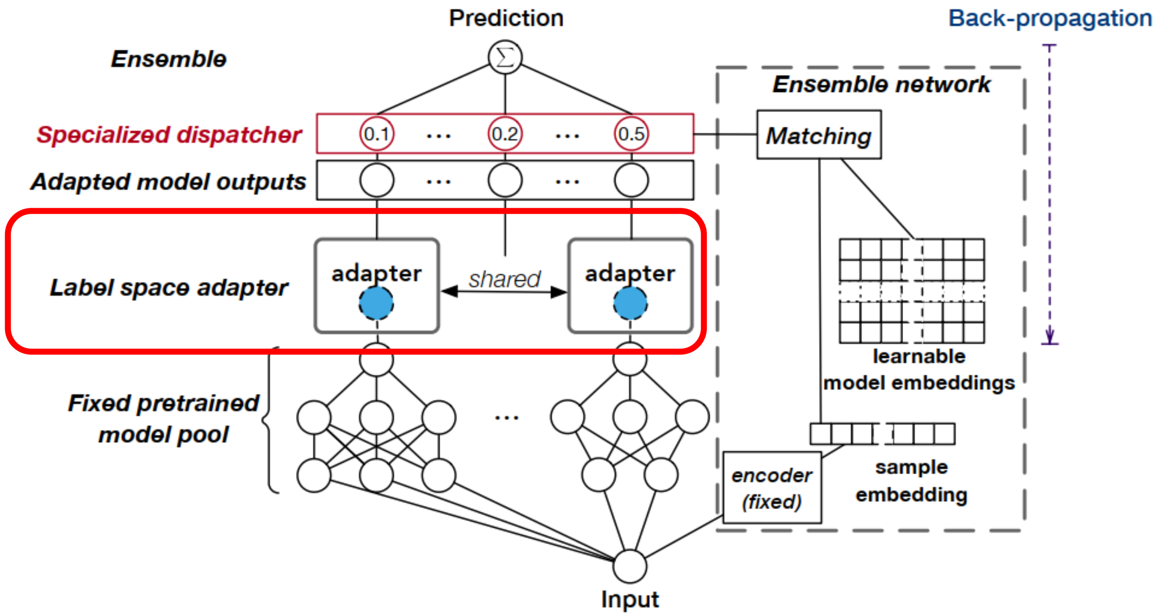


$$z_i = f(x_i, \Theta) = \sum_{j=1}^{N} w_{i,j} f_j(x_i, \theta_j)$$

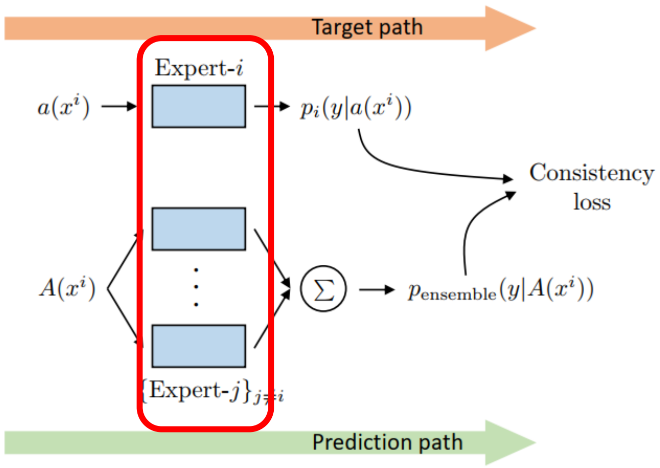
- Mancini M, Bulo S R, Caputo B, et al. Best sources forward: domain generalization through source-specific nets. ICIP 2018.
- Segu M, Tonioni A, Tombari F. Batch normalization embeddings for deep domain generalization[J]. arXiv preprint arXiv:2011.12672, 2020.
- D'Innocente A, Caputo B. Domain generalization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

Ensemble learning for DG

- Ensemble learning for classifier learning
  - SEDGE: ensemble of pre-trained models for classifier learning



DAEL: domain adaptive ensemble learning

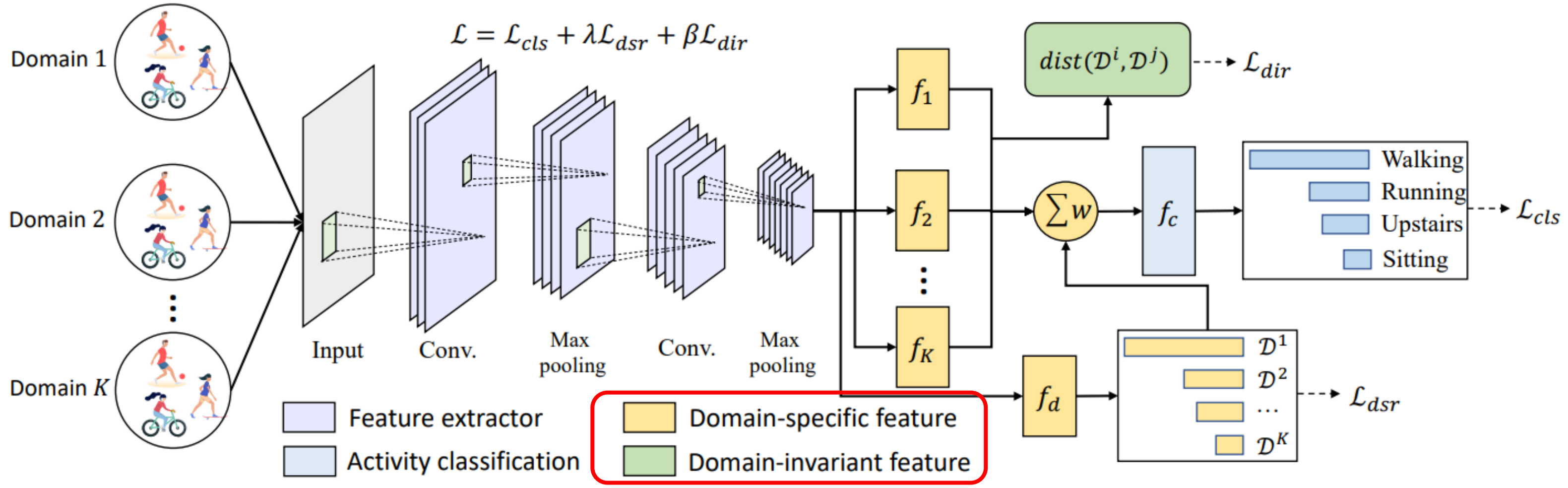


$$w_k = \frac{e^{(\zeta(\mathbf{W}(\mathbf{s})))_k}}{\sum_{j=1}^{K} e^{(\zeta(\mathbf{W}(\mathbf{s})))_j}}$$

Li Z, Ren K, Jiang X, et al. Domain Generalization using Pretrained Models without Fine-tuning[J]. arXiv preprint arXiv:2203.04600, 2022.
Zhou K, Yang Y, Qiao Y, et al. Domain adaptive ensemble learning[J]. IEEE TIP 2021.

Ensemble learning for DG

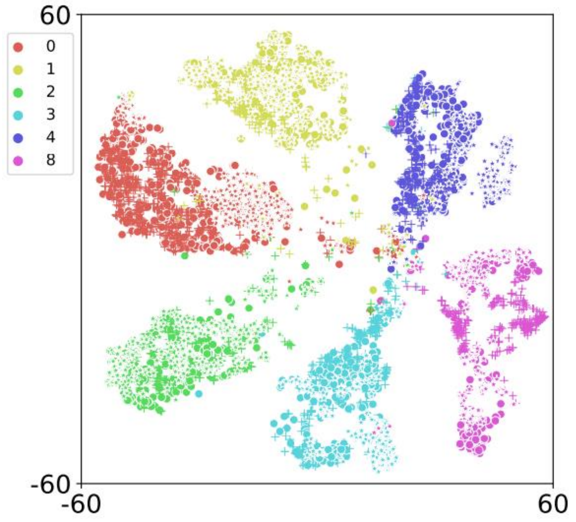
- Is ensemble learning enough for DG?
  - No. Ensemble → domain-specific knowledge
  - We also need a balance with domain-invariant knowledge
  - AFFAR: Adaptive Feature Fusion



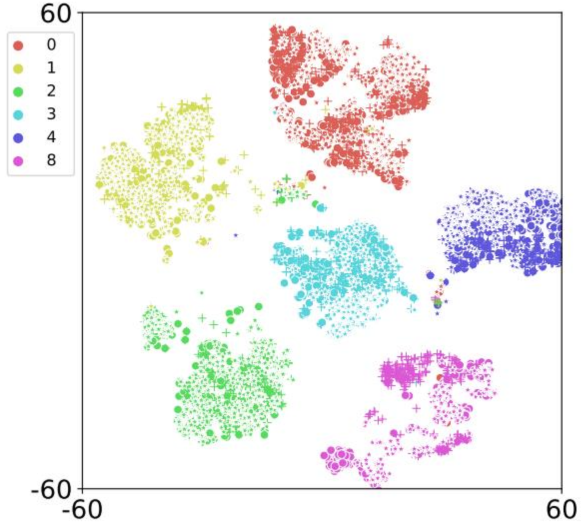
$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{dsr} + \beta\mathcal{L}_{dir}$$

Qin et al. Domain generalization for activity recognition via adaptive feature fusion. ACM TIST 2022.

Domain-specific features



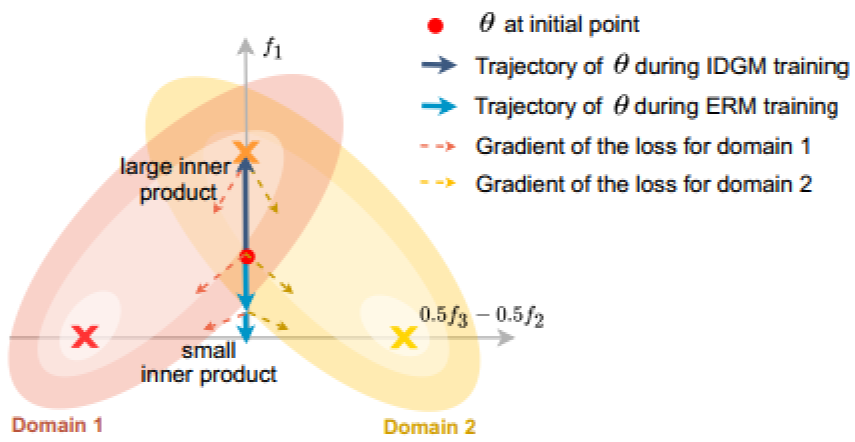
+Domain-invariant features



Gradient operation for DG

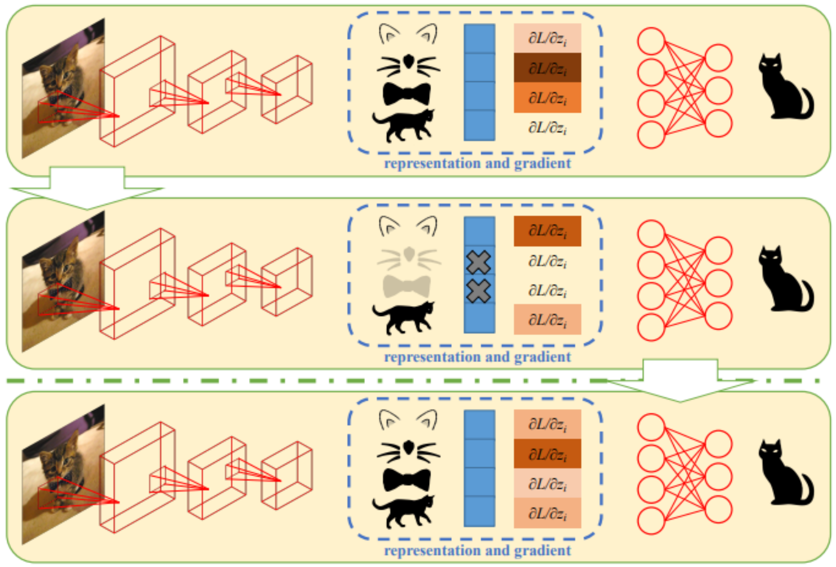
- Model the interactions between cross-domain gradients

Fish: gradient inner product



$$\mathcal{L}_{\text{idgm}} = \mathcal{L}_{\text{erm}}(\mathcal{D}_{tr};\theta) - \gamma \underbrace{\frac{2}{S(S-1)} \sum_{i,j \in S}^{i \neq j} G_i \cdot G_j}_{\text{GIP, denote as } \widehat{G}}$$

RSC: self-challenging for gradient



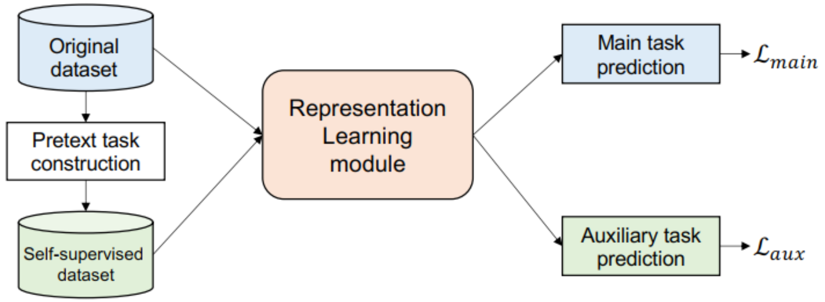
$$\mathbf{m}(i) = \begin{cases} 0, & \text{if} \quad \mathbf{g_z}(i) \geq q_p \\ 1, & \text{otherwise} \end{cases}$$

- Shi Y, Seely J, Torr P H S, et al. Gradient matching for domain generalization. ICLR 2022.
- Huang Z, Wang H, Xing E P, et al. Self-challenging improves cross-domain generalization. ECCV 2020.

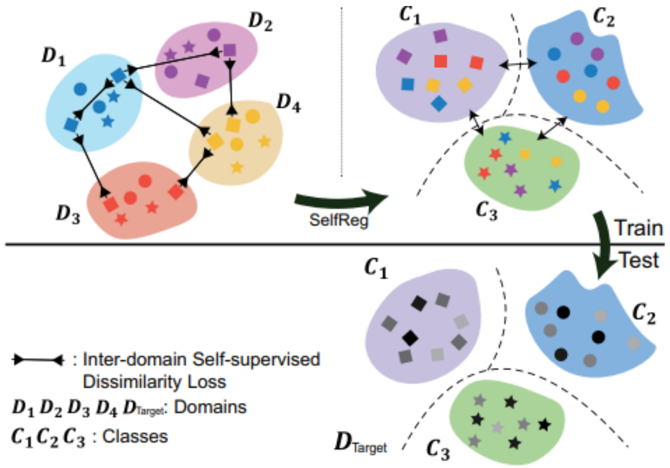
Self-supervised learning for DG

- Construct pretext tasks for general representation learning

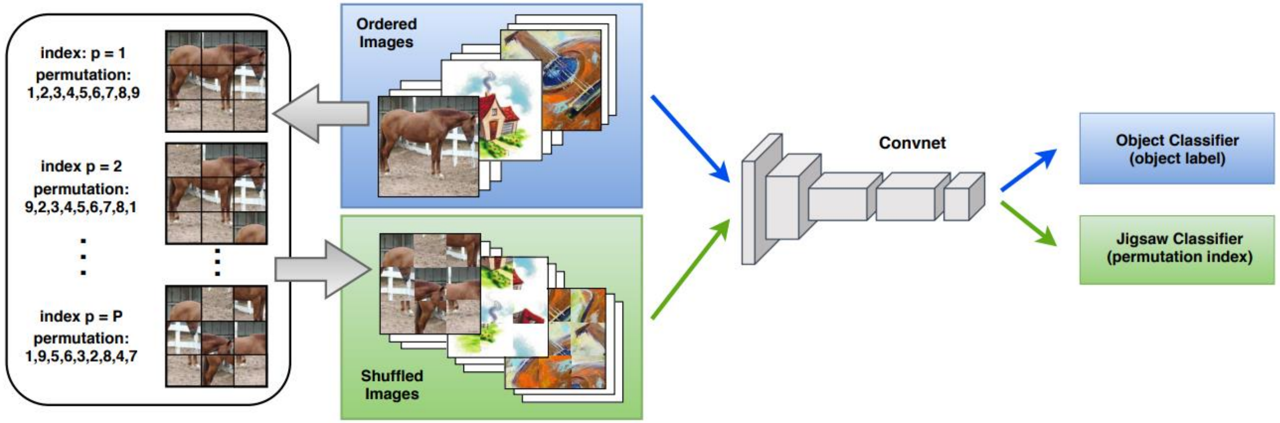
Self-supervised learning



Selfreg: self-supervised contrastive loss



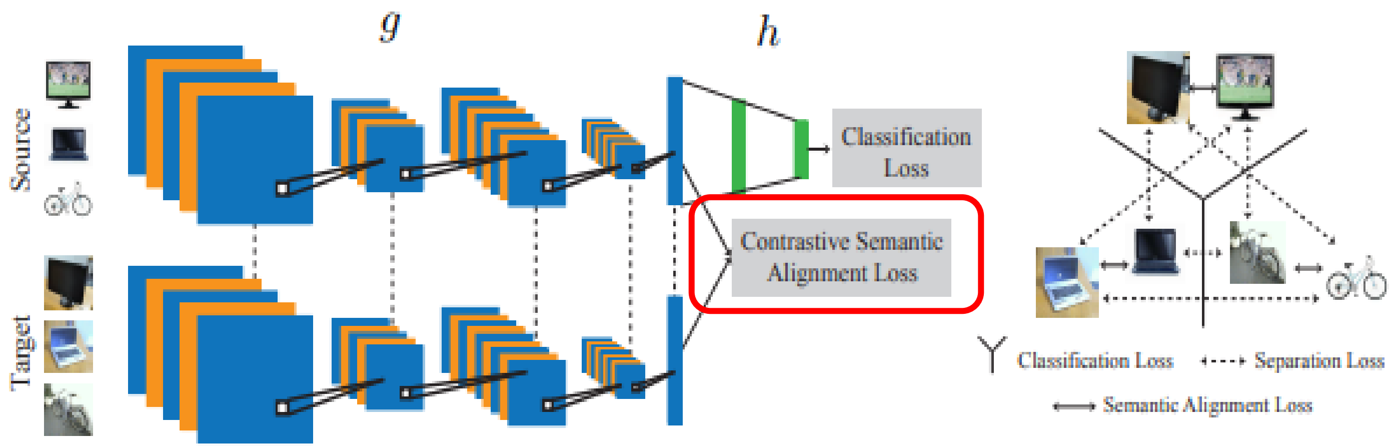
JiGen: Jigsaw puzzle + DG

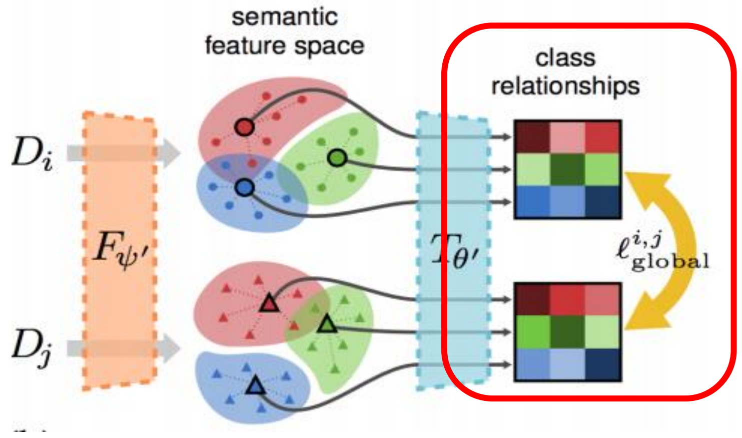


- Carlucci F M, D'Innocente A, Bucci S, et al. Domain generalization by solving jigsaw puzzles. CVPR 2019.
- Kim D, Yoo Y, Park S, et al. Selfreg: Self-supervised contrastive regularization for domain generalization. ICCV 2021.

Contrastive Learning

Minimizing/Maximizing feature distance among samples from with same/different category information from different domains



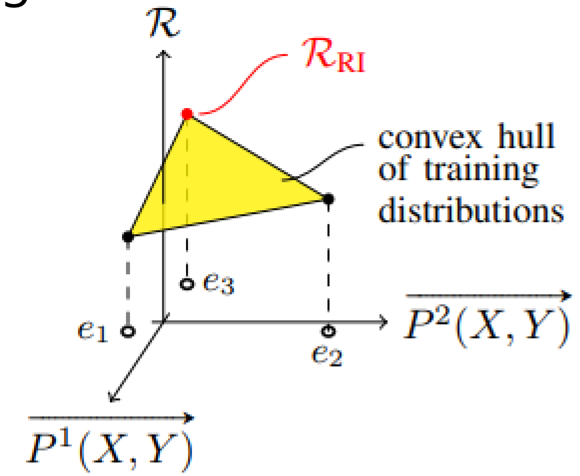


- Motiian, et al., Unified Deep Supervised Domain Adaptation and Generalization, ICCV'17
- Dou, et al., Domain Generalization via Model-Agnostic Learning of Semantic Features, NeurIPS'19

Distributionally robust optimization for DG

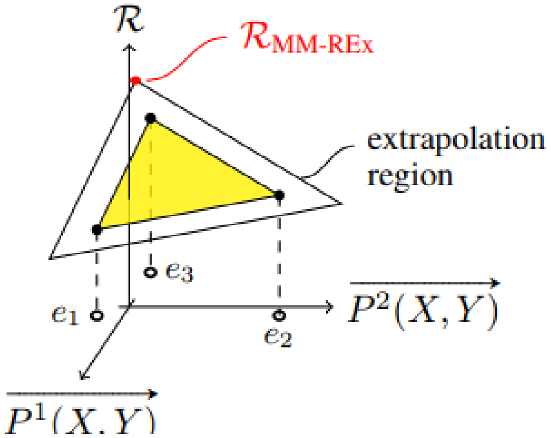
- Learn a model at worst-case distribution scenario

GroupDRO: Convex hull + Regularization



$$\min_{\theta\in\Theta} \sup_{q\in\Delta_m} \sum_{g=1}^{m} q_g \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta;(x,y))]$$

VRex: Convex hull + Perturbation + Risk variance



$$\mathcal{R}_{\text{MM-REx}}(\theta) \doteq \max_{\substack{\Sigma_e\lambda_e=1\\ \lambda_e\geq\lambda_{\min}}} \sum_{e=1}^{m} \lambda_e \mathcal{R}_e(\theta) = (1-m\lambda_{\min}) \max_e \mathcal{R}_e(\theta) + \lambda_{\min}\sum_{e=1}^{m}\mathcal{R}_e(\theta)$$

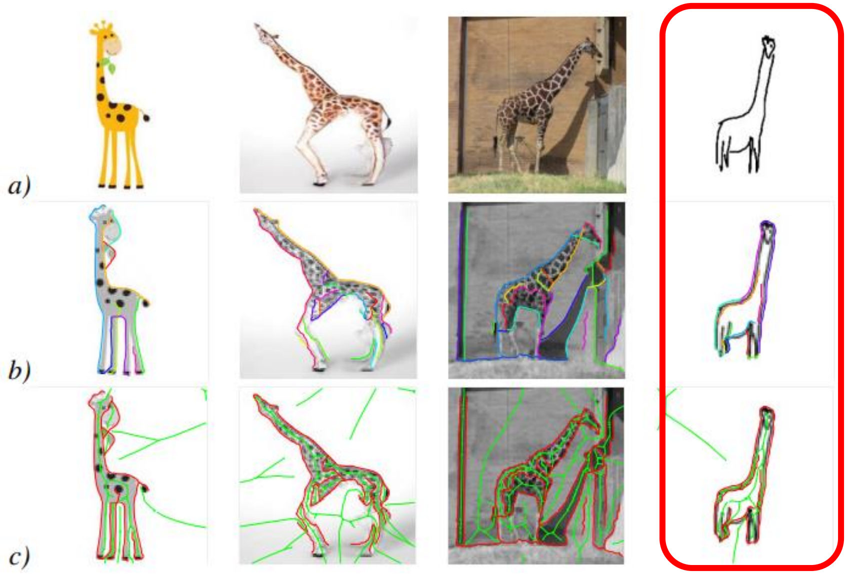
$$\mathcal{R}_{\text{V-REx}}(\theta) \doteq \beta\ \mathrm{Var}(\{\mathcal{R}_1(\theta),...,\mathcal{R}_m(\theta)\}) + \sum_{e=1}^{m}\mathcal{R}_e(\theta)$$

- S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in ICLR, 2020.
- D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in ICML, 2021, pp. 5815–5826.

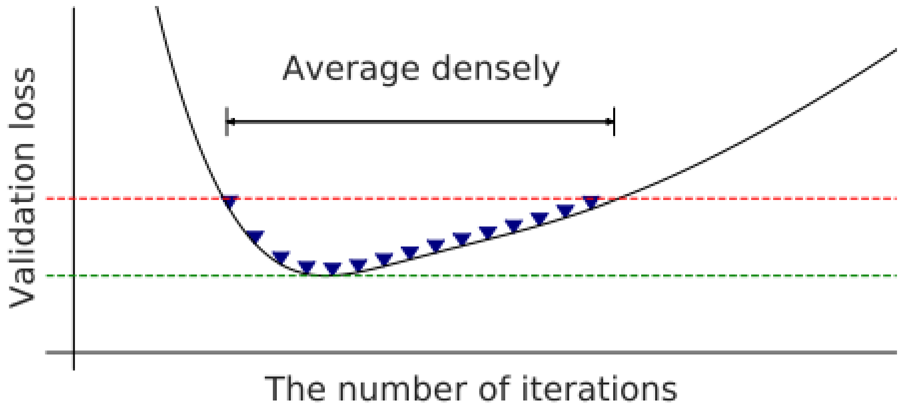
Other learning strategy

- Other interesting learning strategy for DG

Shapelet feature: invariant across domains



SWAD: Smooth training loss



- Narayanan M, Rajendran V, Kimia B. Shape-biased domain generalization via shock graph embeddings. ICCV 2021.
- Cha J, Chun S, Lee K, et al. Swad: Domain generalization by seeking flat minima. NeurIPS 2021.

Introduction of DG & related areas

Methodology

Applications

Datasets, benchmark & evaluation

Theory & challenges

Applications for DG

- Wide applications across CV, NLP, RL, and others

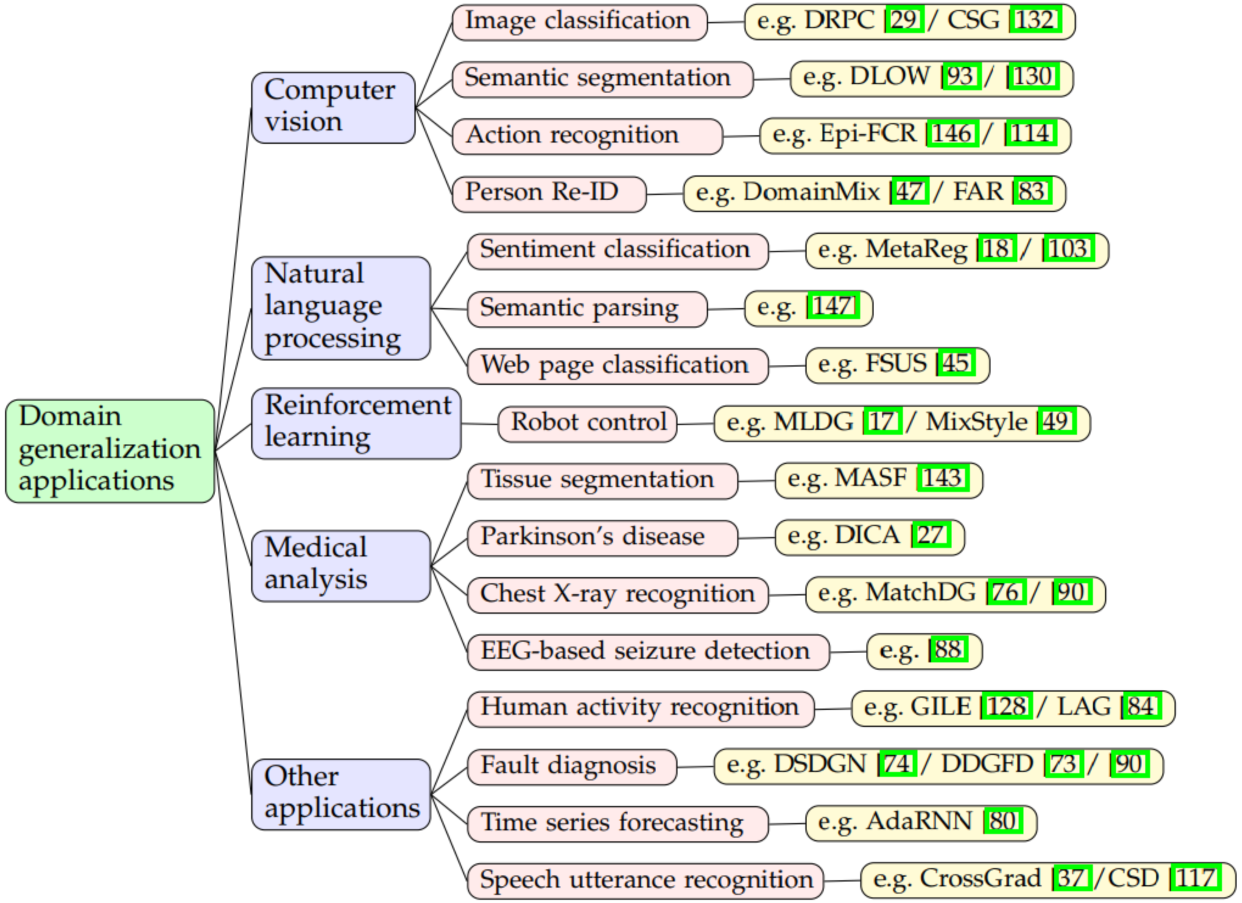
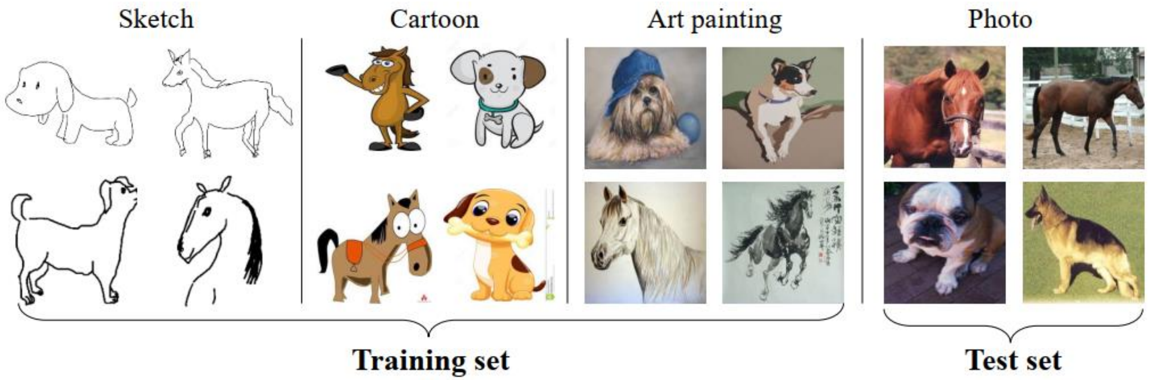


Figure credit: DG survey by Wang et al. (TKDE'22)

Wide applications of DG

- Computer vision

Image classification



Action recognition

Style transfer

Semantic segmentation

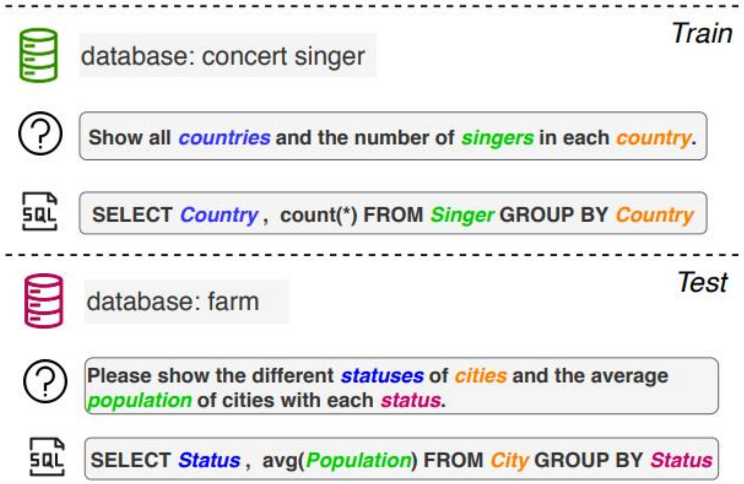
Person ReID

Wide applications of DG

- Natural language processing

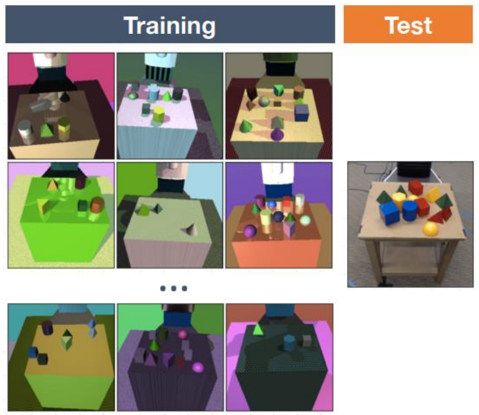
Sentiment classification

Semantic parsing



- Reinforcement learning

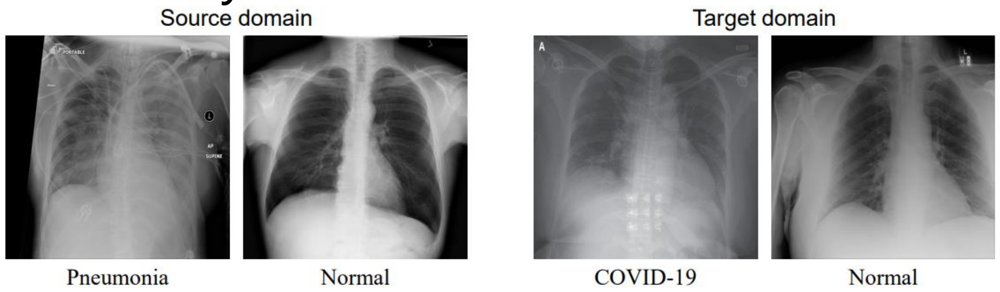
Sim-to-real
Robot
control



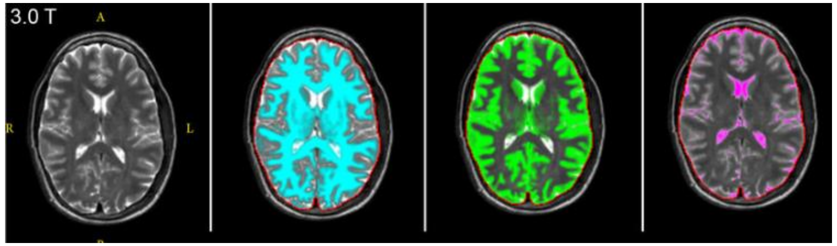
Wide applications of DG

- Medical applications

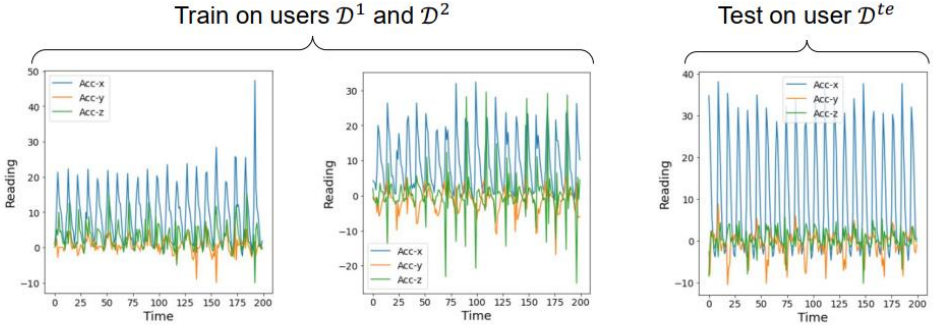
COVID X-ray classification



Tissue segmentation

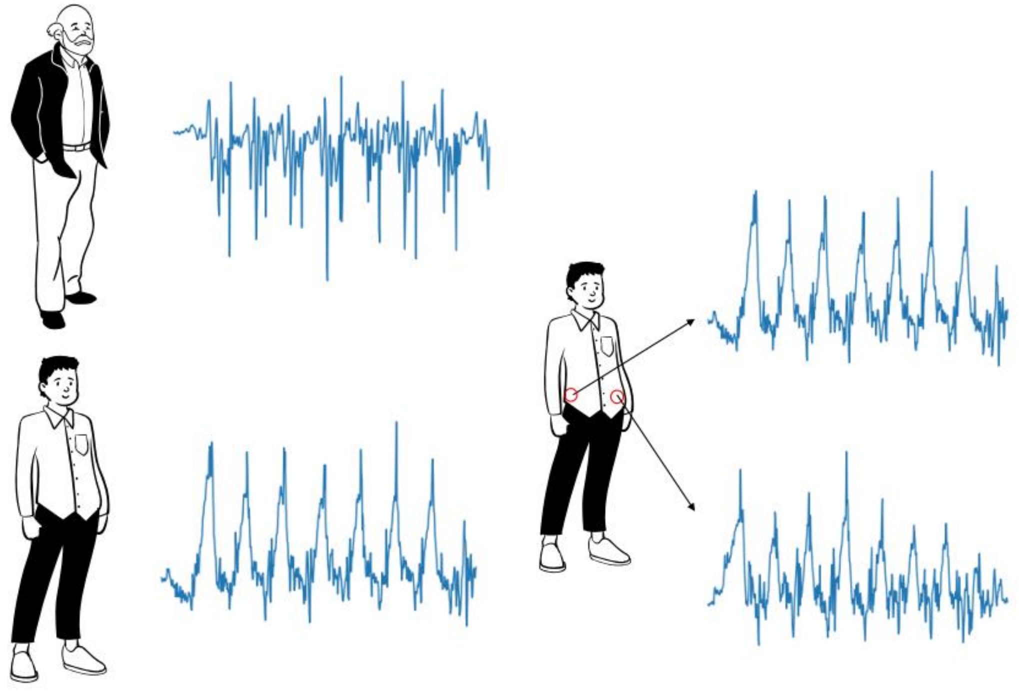


Parkinson’s disease diagnosis

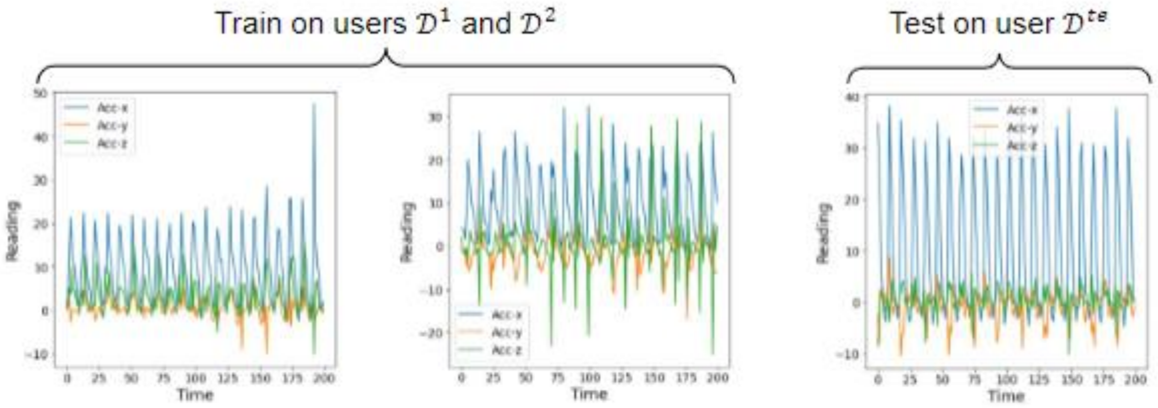


Wide applications of DG

- Sensor-based human activity recognition
  - Create a model that learns generalizable representations for different age groups
  - Different people/device locations/activity patterns generate different sensor readings



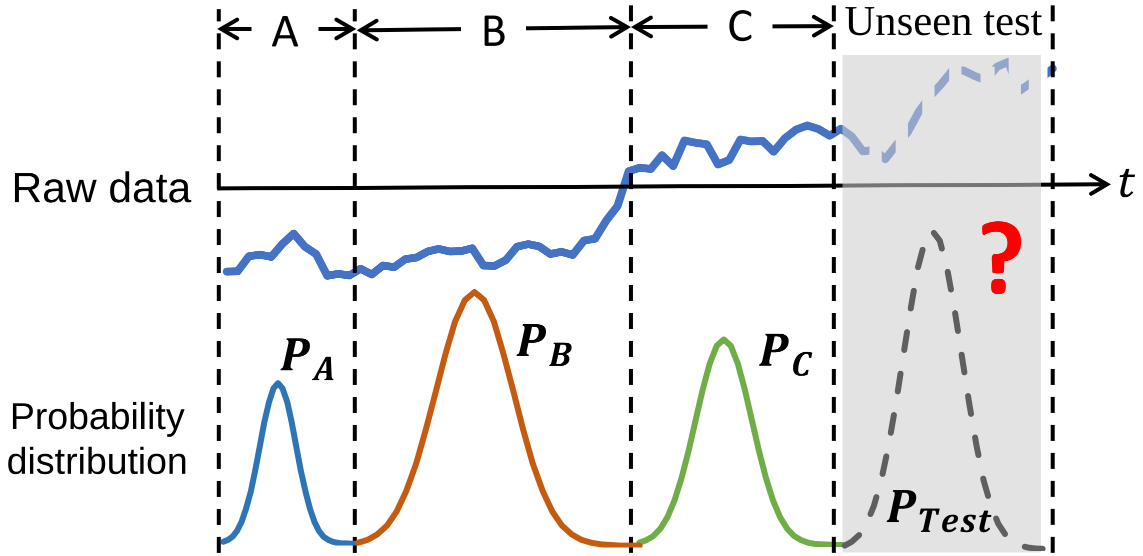
(a) Different sensor readings on different subjects. (b) Different sensor readings on different positions.



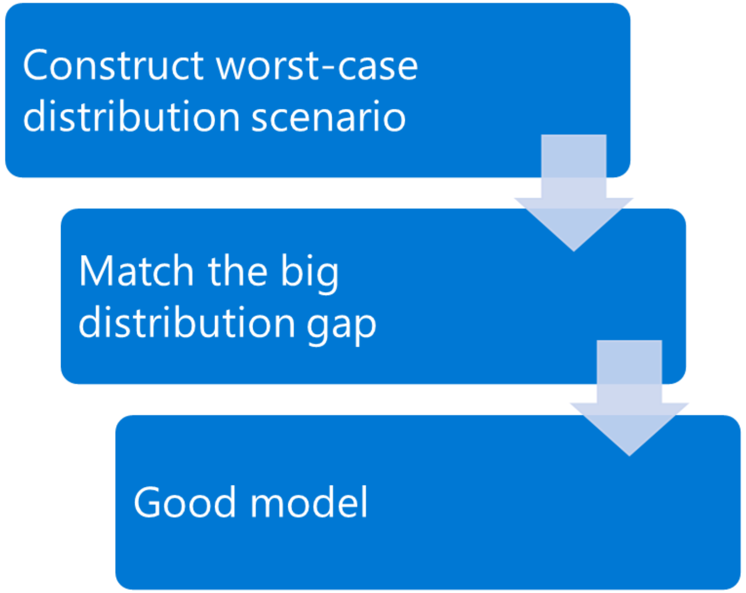
- Lu et al. Local and global alignments for generalizable sensor-based human activity recognition. ICASSP 2022.
- Lu et al. Semantic-discriminative mixup for Generalizable Sensor-based Cross-domain Activity Recognition. ACM IMWUT 2022.

Wide applications of DG

- Time series forecasting
  - AdaRNN: adaptive forecasting of time series using DG



Temporal Covariate Shift: $P_A \neq P_B \neq P_C \neq P_{Test}$

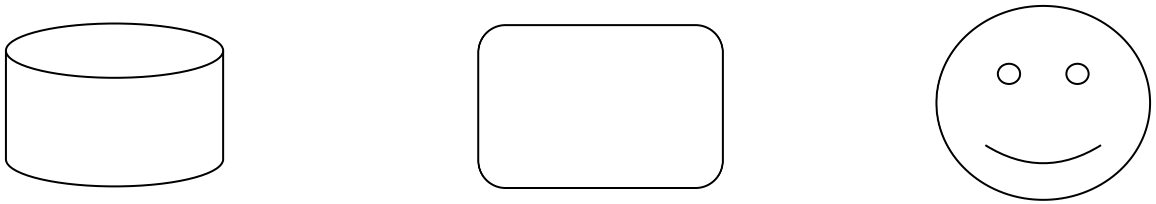


Du et al. AdaRNN: adaptive learning and forecasting of time series. CIKM 2021.

- Introduction of DG & related areas
- Methodology
- Applications
- **Datasets, benchmark & evaluation**
- Theory & challenges

Benchmarks for DG

- Important consideration for DG benchmarks:



Which dataset? Which codebase? Which metric?

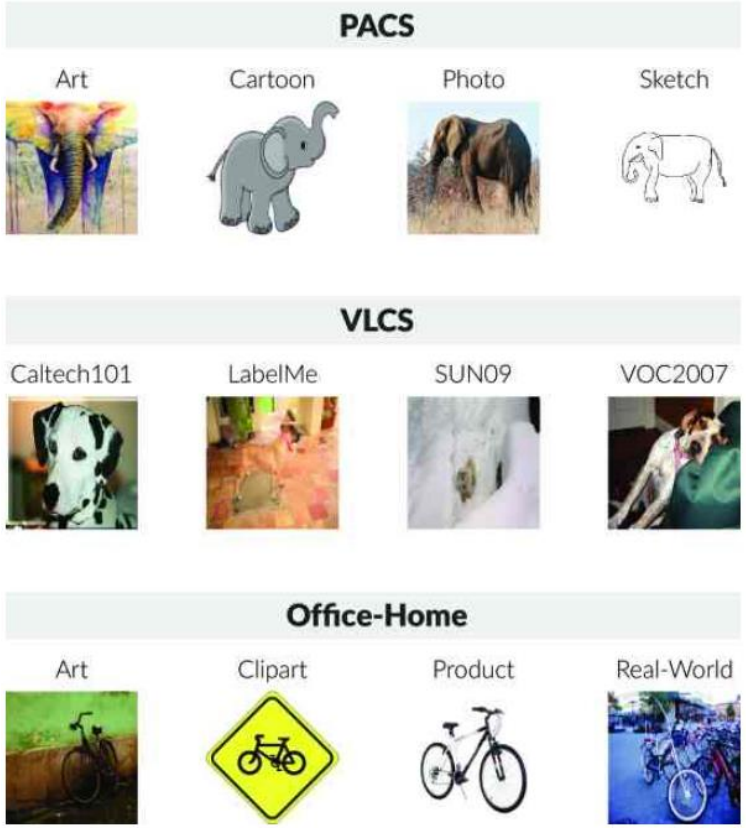
  - Popular datasets
  - Common benchmarks and codebases
  - Evaluation strategy: model selection

Note:

- Technically, ***any*** application settings that fits in DG scenario can be considered as a good test bed.
- There exists ***no*** “golden-standard” for benchmarking and evaluation.

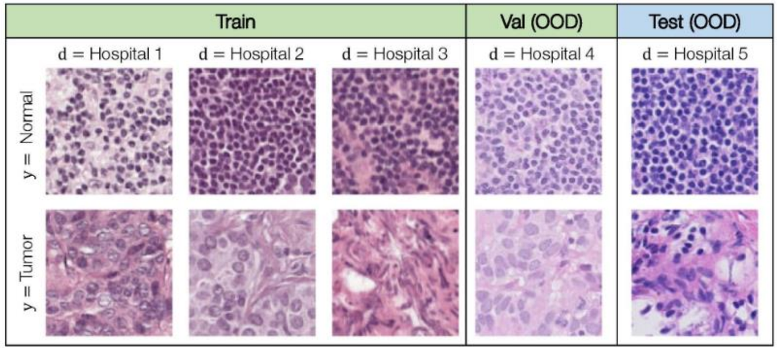
Datasets for DG

- Common benchmarks

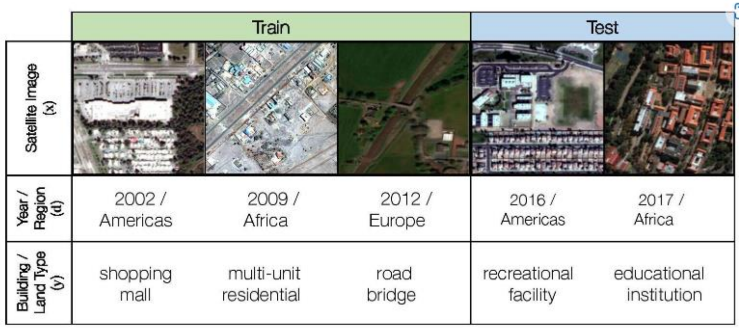


Camelyon17

WILDS



FMoW



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
iWildCam2020-wilds	323	182	203,029	Species classification across different camera traps
Camelyon17-wilds	5	2	45,000	Tumor identification across five different hospitals
RxRx1-wilds	51	1,139	84,898	Genetic perturbation classification across experimental batches
OGB-MolPCBA	120,084	128	400,000	Molecular property prediction across different scaffolds
GlobalWheat-wilds	47	bounding boxes	6,515	Wheat head detection across regions of the world
CivilComments-wilds	-	2	450,000	Toxicity classification across demographic identities
FMoW-wilds	80	62	118,886	Land use classification across different regions and years
PovertyMap-wilds	46	real value	19,669	Poverty mapping across different countries
Amazon-wilds	3920	5	539,502	Sentiment classification across different users
Py150-wilds	8,421	next token	150,000	Code completion across different codebases

Benchmark and codebase

- DomainBed
  - A unified benchmark for domain generalization

Available datasets

The currently available datasets are:

- RotatedMNIST (Ghifary et al., 2015)
- ColoredMNIST (Arjovsky et al., 2019)
- VLCS (Fang et al., 2013)
- PACS (Li et al., 2017)
- Office-Home (Venkateswara et al., 2017)
- A TerraIncognita (Beery et al., 2018) subset
- DomainNet (Peng et al., 2019)
- A SVIRO (Dias Da Cruz et al., 2020) subset
- WILDS (Koh et al., 2020) FMoW (Christie et al., 2018) about satellite images
- WILDS (Koh et al., 2020) Camelyon17 (Bandi et al., 2019) about tumor detection in tissues

Algorithm	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Average
ERM	51.5 ± 0.1	98.0 ± 0.0	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	66.6
IRM	52.0 ± 0.1	97.7 ± 0.1	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	65.4
GroupDRO	52.1 ± 0.0	98.0 ± 0.0	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	64.8
Mixup	52.1 ± 0.2	98.0 ± 0.1	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	39.2 ± 0.1	66.7
MLDG	51.5 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.1	66.7
CORAL	51.5 ± 0.1	98.0 ± 0.1	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.1	67.5
MMD	51.5 ± 0.2	97.9 ± 0.0	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	63.3
DANN	51.5 ± 0.3	97.8 ± 0.1	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	38.3 ± 0.1	66.1
CDANN	51.7 ± 0.1	97.9 ± 0.1	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	38.3 ± 0.3	65.6
MTL	51.4 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	40.6 ± 0.1	66.2
SagNet	51.7 ± 0.0	98.0 ± 0.0	77.8 ± 0.5	86.3 ± 0.2	68.1 ± 0.1	48.6 ± 1.0	40.3 ± 0.1	67.2
ARM	56.2 ± 0.2	98.2 ± 0.1	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	35.5 ± 0.2	66.1
VREx	51.8 ± 0.1	97.9 ± 0.1	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 2.9	65.6
RSC	51.7 ± 0.2	97.6 ± 0.1	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	38.9 ± 0.5	66.1

Model selection: training-domain validation set

Interesting results: DomainBed found that there ***are not*** significant improvements for recent DG algorithms. *Is it the case?*

https://github.com/facebookresearch/DomainBed

Benchmark and codebase

- DeepDG
  - Built by borrowing the knowledge from DomainBed, but *faster*, and *easier* to use

Implemented Algorithms

We currently support the following algoirthms. We are working on r

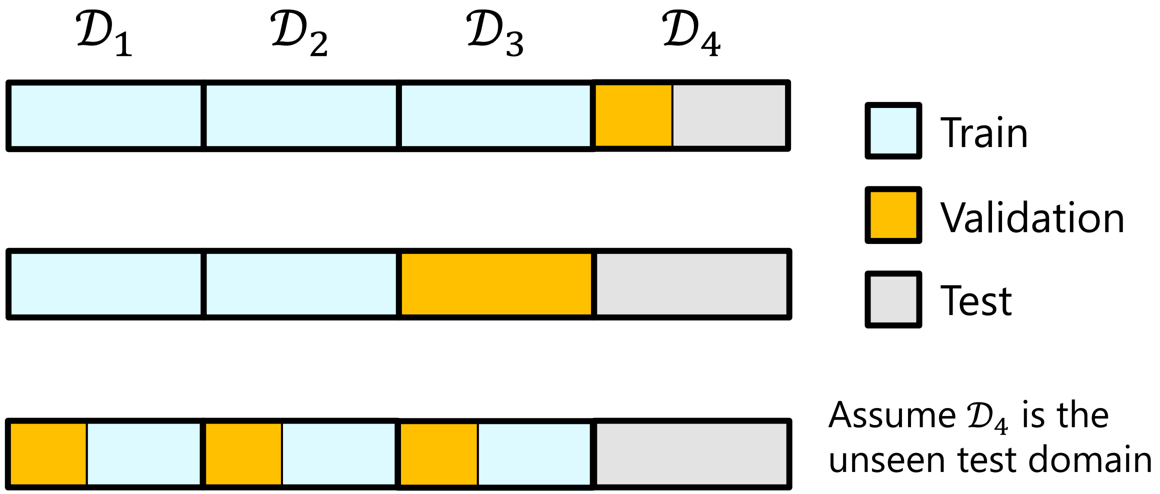
1. ERM
2. DDC (Deep Domain Confusion, arXiv 2014) [1]
3. CORAL (COrrelation Alignment, ECCV-16) [2]
4. DANN (Domain-adversarial Neural Network, JMLR-16) [3]
5. MLDG (Meta-learning Domain Generalization, AAAI-18) [4]
6. Mixup (ICLR-18) [5]
7. RSC (Representation Self-Challenging, ECCV-20) [6]
8. GroupDRO (ICLR-20) [7]
9. ANDMask (ICLR-21) [8]
10. VREx (ICML-21) [9]

- Avoids huge hyperparameter tuning
- More friendly interface
- Better customization

https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG

Model selection

- Model selection in DomainBed
  - Test-domain validation set (oracle)
    - Use part of test domain as the validation
  - Leave-one-domain-out cross-validation
    - One domain as testing domain for validation
  - Training-domain validation set ***(popular)***
    - Leave some part of the training data as the validation set



- Q: is it reasonable to use training-domain validation for model selection?
- A: ***no***. Since the validation distribution cannot represent the test distribution.

Discussion about the performance of DG

- Performance should be restricted to certain applications
  - Cross-dataset human activity recognition[1]

Source	Target	DeepALL	DANN	CORAL	ANDMask	GroupDRO	RSC	Mixup	SDMix
1,2,3,4	0	41.52	45.45	33.22	47.51	27.12	46.56	**48.77**	47.50
0,2,3,4	1	26.73	25.36	25.18	31.06	26.66	27.37	34.19	**36.10**
0,1,3,4	2	35.81	38.06	25.81	39.17	24.34	35.93	37.49	**42.53**
0,1,2,4	3	21.45	28.89	22.32	30.22	18.39	27.04	29.50	**34.52**
0,1,2,3	4	27.28	25.05	20.64	29.90	24.82	29.82	29.95	**30.93**
AVG	-	30.56	32.56	25.43	35.57	24.27	33.34	35.98	**38.32**

  - Cross-dataset object detection[2]

Setting	Method	Cityscapes→Foggy Cityscapes								
		person	rider	car	truck	bus	train	mcycle	bicycle	mAP
DG	Faster R-CNN [52]	17.8	23.6	27.1	11.9	23.8	9.1	14.4	22.8	18.8
	SNR-Faster R-CNN	**20.3**	**24.6**	**33.6**	**15.9**	**26.3**	**14.4**	**16.8**	**26.8**	**22.3**
UDA	DA Faster R-CNN [72]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
	SNR-DA Faster R-CNN	**27.3**	**34.6**	**44.6**	**23.9**	**38.1**	**25.4**	**21.3**	**29.7**	**30.6**

Hint: maybe we should develop application-oriented evaluation benchmarks?

[1] Lu et al. Semantic-discriminative mixup for generalizable cross-domain sensor-based human activity recognition. ACM IMWUT 2022.
[2] Jin X, Lan C, Zeng W, et al. Style normalization and restitution for domain generalization and adaptation. IEEE TMM 2021.

- Introduction of DG & related areas
- Methodology
- Applications
- Datasets, benchmark & evaluation
- **Theory & challenges**

Theory

- Domain adaptation error bound
  - The error on target domain is bounded by:

$$\epsilon^t(h) \leq \epsilon^s(h) + d_{\mathcal{H}\Delta\mathcal{H}}\left(P_X^s, P_X^t\right) + \lambda_{\mathcal{H}}$$

Target risk

Source risk

Complexity of $\mathcal{H}$

Source-target distribution divergence

$\mathcal{H}$-divergence: $d_{\mathcal{H}}(P,Q) = 2\sup_{h\in\mathcal{H}} |Pr_P[\mathbb{I}(h)] - Pr_Q[\mathbb{I}(h)]|$

$\mathcal{H}\Delta\mathcal{H}$-distance: $d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) = 2\sup_{h,h'\in\mathcal{H}} |Pr_{x\sim P}[h(x) \neq h'(x)] - Pr_{x\sim Q}[h(x) \neq h'(x)]|$

Discrepancy distance: $disc_L(P,Q) = \max_{h,h'\in\mathcal{H}} |L_P(h,h') - L_Q(h,h')|$

- Ben-David S, Blitzer J, Crammer K, et al. Analysis of representations for domain adaptation. NIPS 2016.
- Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains[J]. Machine learning, 2010, 79(1): 151-175.
- Mansour Y, Mohri M, Rostamizadeh A. Domain adaptation with multiple sources. NIPS 2009.

Theory for DG

- Assumption 1: convex hull
  - Key: approximate target domain using the convex hull of source distributions

$\Lambda := \{\sum_{i=1}^{M} \pi_i P_X^i \mid \pi \in \Delta_M\}$

Ideal joint risk (best source vs. target)

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

Target risk

Weighted source risk

$\gamma := \min_{\pi \in \Delta_M} d_{\mathcal{H}}(P_X^t, \sum_{i=1}^{M} \pi_i P_X^i)$

Distance between target and convex hull

$\rho := \sup_{P_X', P_X'' \in \Lambda} d_{\mathcal{H}}(P_X', P_X'')$

Diameter of Λ

Albuquerque I, Monteiro J, Darvishi M, et al. Adversarial target-invariant representation learning for domain generalization[J]. 2020.

Theory for DG

- Assumption 2: classifier variation
  - Key: the gap between available environments and all invariants

$$\mathrm{err}(f) = \mathcal{L}(\mathcal{E}_{all}, f) - \mathcal{L}(\mathcal{E}_{avail}, f)$$

Theorem 4.1 (Main Theorem). *Suppose we have learned a classifier $f(x) = g(h(x))$ such that $\forall e \in \mathcal{E}_{all}$ and $\forall y \in \mathcal{Y}$, $p_{h^e|Y^e}(h|y) \in L^2(\mathbb{R}^d)$. Denote the characteristic function of random variable $h^e|Y^e$ as $\hat{p}_{h^e|Y^e}(t|y) = \mathbb{E}[\exp\{i\langle t, h^e\rangle\}|Y^e = y]$. Assume the hypothetical space $\mathcal{F}$ satisfies the following regularity conditions that $\exists \alpha, M_1, M_2 > 0, \forall f \in \mathcal{F}, \forall e \in \mathcal{E}_{all}, y \in \mathcal{Y}$,*

$$\int_{h \in \mathbb{R}^d} p_{h^e|Y^e}(h|y)|h|^\alpha \mathrm{d}h \le M_1 \quad \text{and} \quad \int_{t \in \mathbb{R}^d} |\hat{p}_{h^e|Y^e}(t|y)||t|^\alpha \mathrm{d}t \le M_2. \tag{4}$$

If $(\mathcal{E}_{avail}, \mathcal{E}_{all})$ is $\left(s(\cdot), \mathcal{I}^{inf}(h, \mathcal{E}_{avail})\right)$-learnable under Φ with Total Variation ρ[3] *then we have*

$$\mathrm{err}(f) \le O\left(\boxed{s\left(\mathcal{V}_\rho^{sup}(h, \mathcal{E}_{avail})\right)^{\frac{\alpha^2}{(\alpha+d)^2}}}\right). \tag{5}$$

Variation

Here ρ is total variation distance, and $O(\cdot)$ depends on d, C, α, M_1, M_2.

Ye H, Xie C, Cai T, et al. Towards a theoretical framework of out-of-distribution generalization. NeurIPS 2021.

Theory of DG

- Assumption 3: subpopulation shift
  - Key: Gaussian mixture model to contain all sub-distributions

Theorem 1 (Error comparison with subpopulation shifts) *Consider n independent samples generated from model (4), $\pi^{(R)} = \pi^{(1)} = 1/2$, $\pi^{(0,R)} = \pi^{(1,G)} = \alpha < 1/4$, $\max_{y,d} \|\mu^{(y,d)}\|_2 \leq C$, and Σ is positive definite. Suppose (ξ, α) satisfies that $\xi < \min\{\frac{\|\tilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, 1\} - C\alpha$ for some large enough constant C and $\|\tilde{\Delta}\|_\Sigma \leq \sqrt{\frac{2\mathbb{E}[\lambda_i^2]}{\max\{3var(\lambda_i), 1/4\}}}$. Then for any $p_{sel} \in [0,1]$,*

$$\widehat{E}^{(wst)}_{\text{LISA}} < \boxed{\min\{\widehat{E}^{(wst)}_{\text{ERM}}, \widehat{E}^{(wst)}_{\text{mix}}\}} + O_P\left(\frac{p \log n}{n} + \frac{p}{\alpha n}\right).$$

Gaussian mixture distribution

Yao H, Wang Y, Li S, et al. Improving Out-of-Distribution Robustness via Selective Augmentation. ICML 2022.

Theory of DG

- Other theory
  - Adversarial training and pretrained model is good for DG[1]
  - DG can be bounded under kernel learning conditions[2]
- Current progress
  - The research on DG theory is still on the go

- [1] Yi M, Hou L, Sun J, et al. Improved OOD Generalization via Adversarial Training and Pretraing. ICML 2021.
- [2] Deshmukh A A, Lei Y, Sharma S, et al. A generalization error bound for multi-class domain generalization[J]. arXiv preprint arXiv:1905.10392, 2019.

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

Conclusion

General ML $\xrightarrow{\text{Non-IID}}$ Domain adaptation $\xrightarrow{\text{Unseen target}}$ Domain generalization

DG Roadmap

- Introduction and background
- Relation with existing area: transfer learning, domain adaptation, multi-task learning…
- Algorithm
  - Data manipulation: augmentation, or generation
  - Representation learning: domain-invariant learning, disentanglement
  - Learning strategy: meta-learning, ensemble learning, gradient, DRO, SSL…
- Applications: CV, NLP, RL, medical…
- Datasets, benchmark, evaluation
- Theory and future challenges

Microsoft

Thanks

Contact: jindong.wang@microsoft.com, haoliang.li@cityu.edu.hk

Tutorial website: https://dgresearch.github.io/

DG survey paper: https://arxiv.org/abs/2103.03097

Codebase: https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG