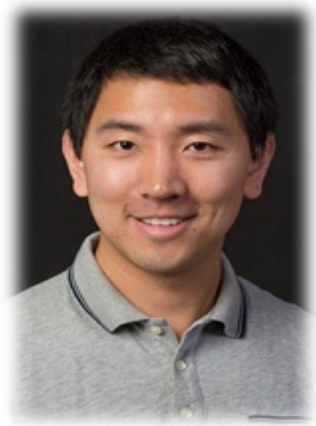
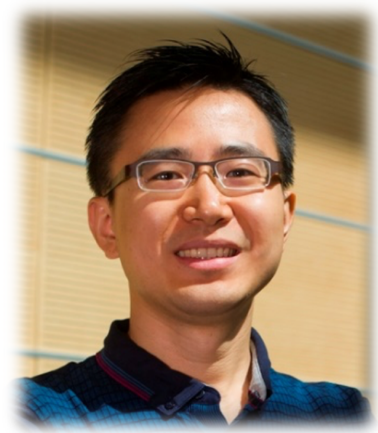
Northeastern University

Multi-view Visual Data Analytics

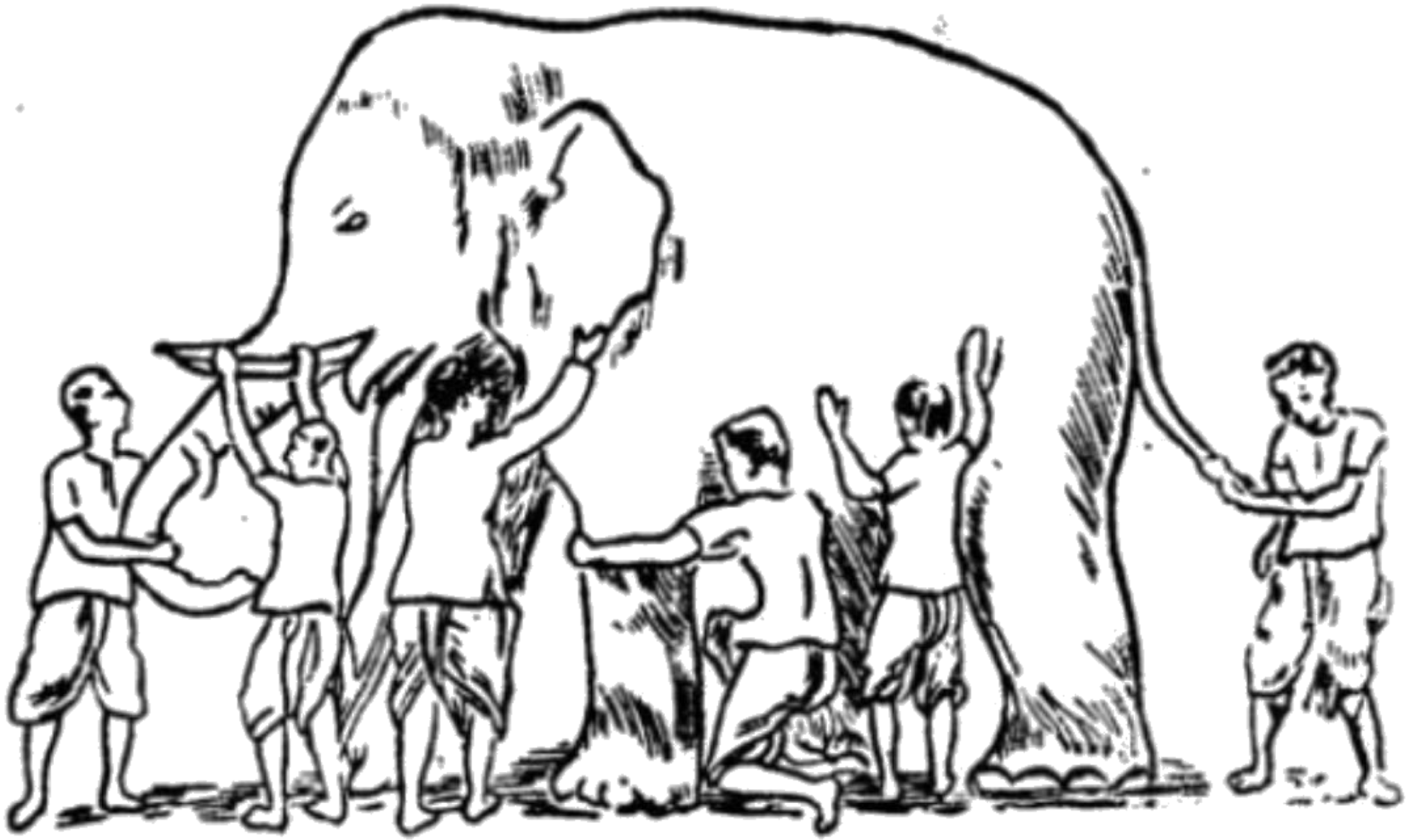
---CVPR-2018 Tutorial



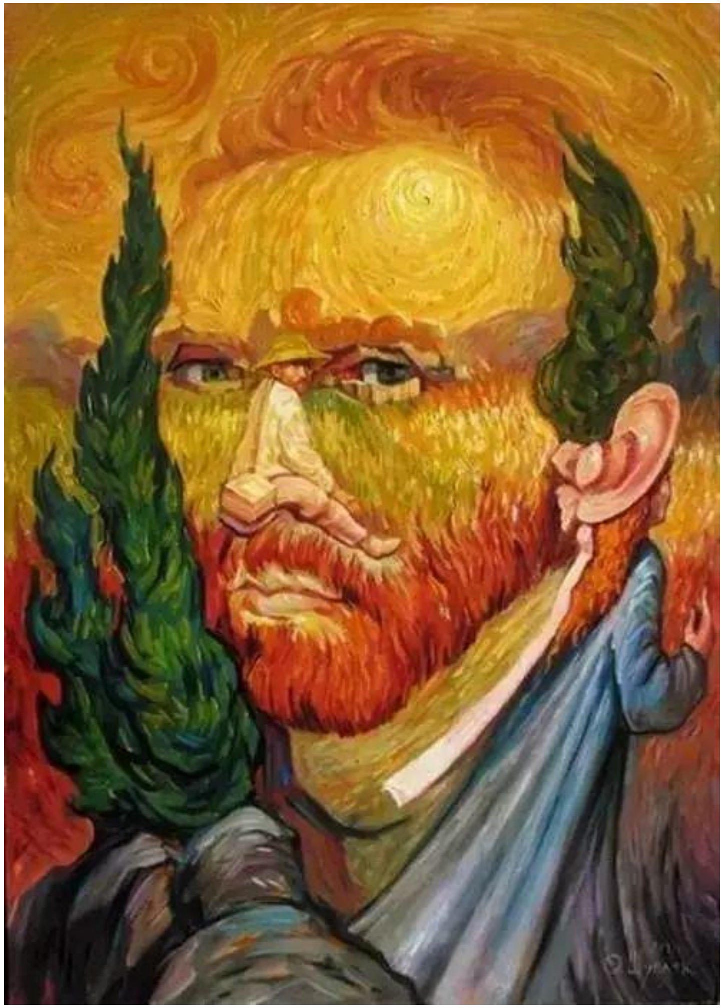


Zhengming Ding **Ming Shao** **Yun Fu**

An Example from Life



An Example from Art



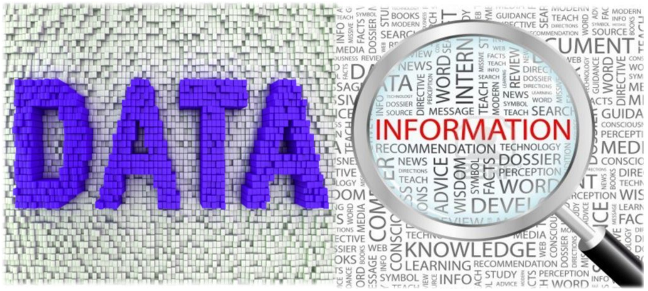
Outline

Northeastern University

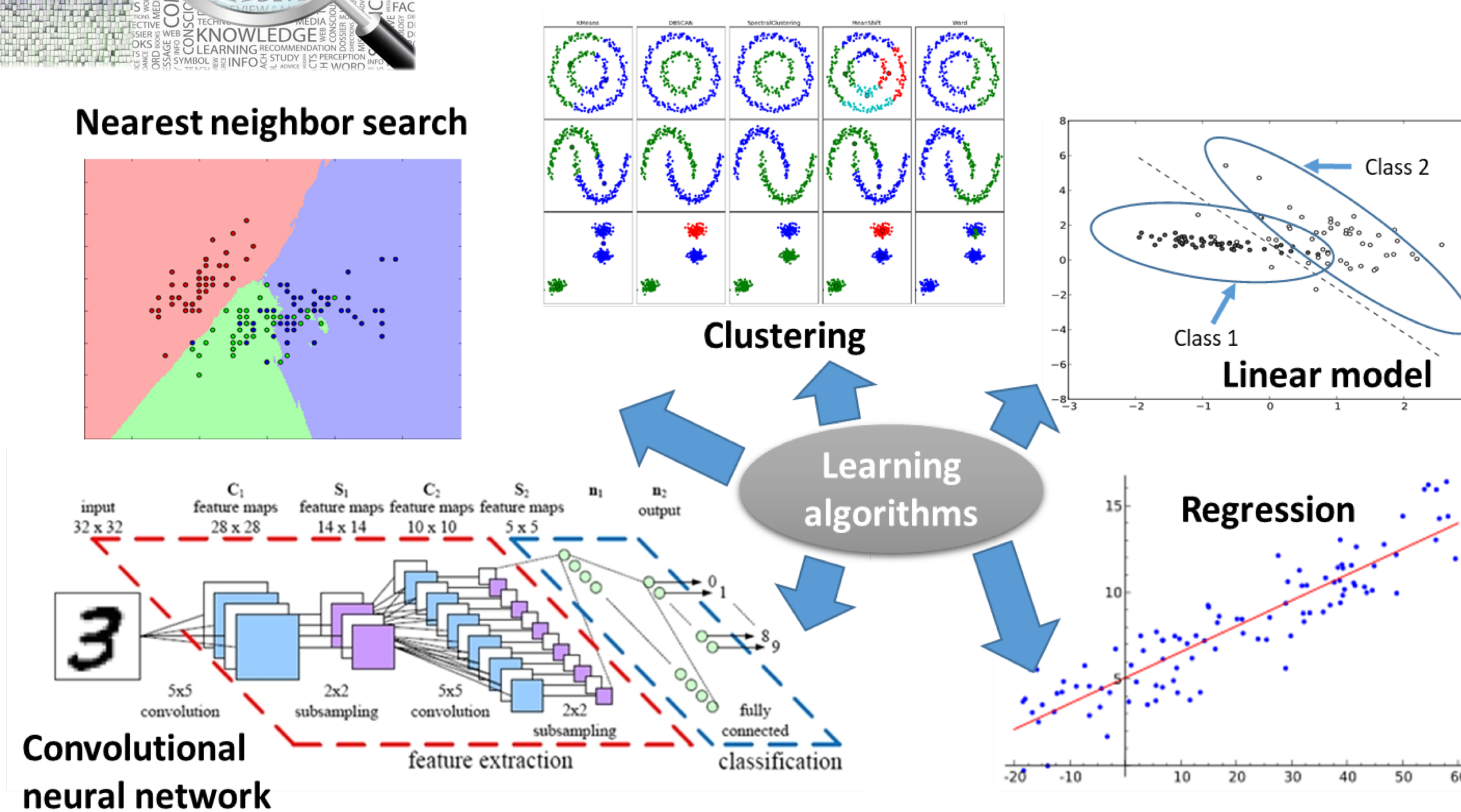
Synergetic Media Learning Lab

- **Introduction & Background**
  - Multi-view Visual Data
  - Multi-view Learning Problems
  - Multi-view Learning Taxonomy
- **Multi-view Learning**
  - Projection and Embedding
  - Knowledge Fusion
  - Multi-view Clustering
  - Supervised Multi-view Learning → Zero-shot Learning
- **Domain Adaptation**
  - Transfer Learning → Domain Adaptation
  - Multi-Source Domain Adaptation & Domain Generalization
- **Conclusion**

What is Multi-view Learning in General?

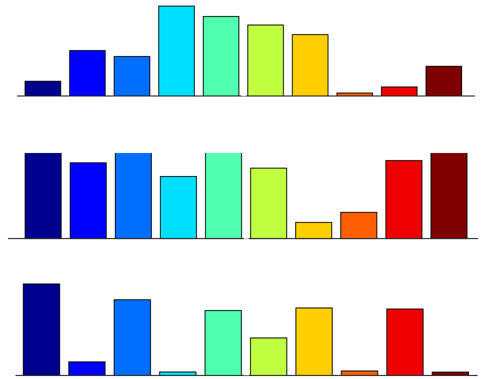


- What **Data** we are interested in??
- What **Learning Problems** we will formulate??



Multi-view Visual Data

LBP
SIFT
HOG



Multiple features

NBC NEWS

CNN

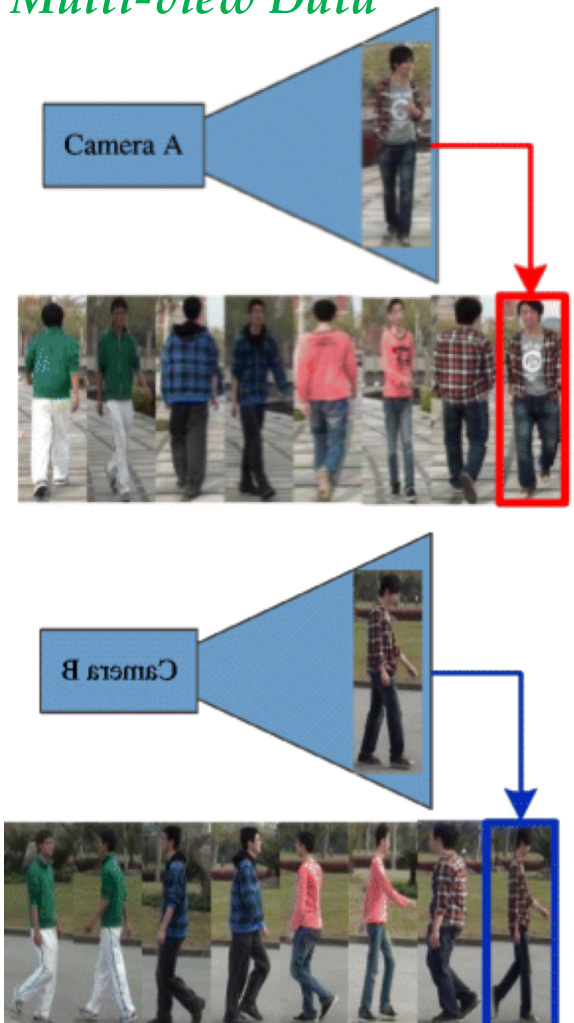
BBC

REUTERS

YAHOO!
NEWS

Multi-view Visual Data

Multi-view Data

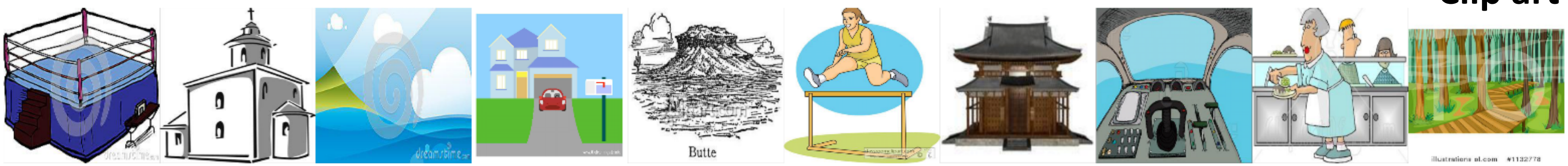


Multi-view Visual Data

Northeastern University
Smile lab
Synergetic Media Learning Lab

Multi-modal Data

Clip art



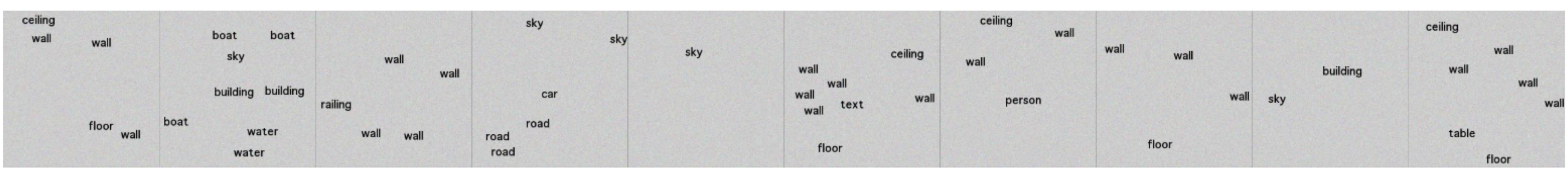
Sketches

Descriptions

There is a bed with a striped bedspread. Beside this is a nightstand with a drawer. There is also a tall dresser and a chair with a blue cushion. On the dresser is a jewelry box and a clock.

I am inside a room surrounded by my favorite things. This room is filled with pillows and a comfortable bed. There are stuffed animals everywhere. I have posters on the walls. My jewelry box is on the dresser.

There are brightly colored wooden tables with little chairs. There is a rug in one corner with ABC blocks on it. There is a bookcase with picture books, a larger teacher's desk and a chalkboard.

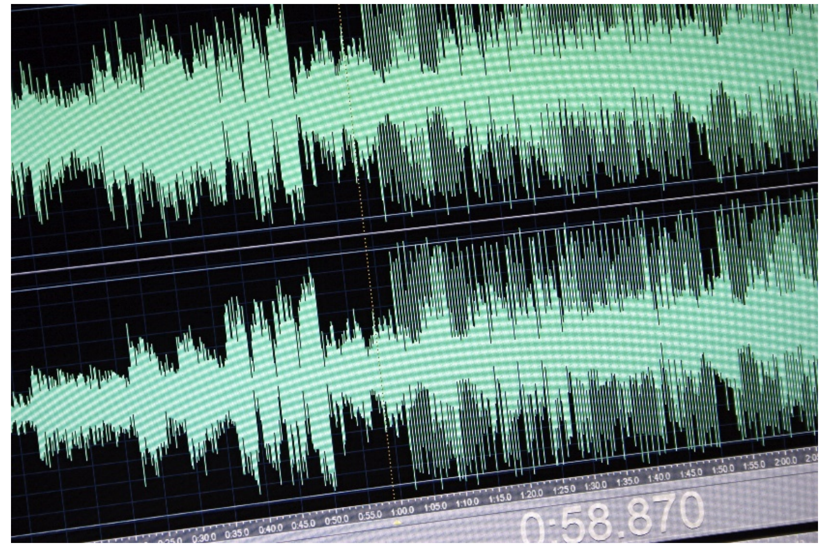


Spatial text

http://projects.csail.mit.edu/cmplaces/

Multi-view Visual Data

Multi-modal Data

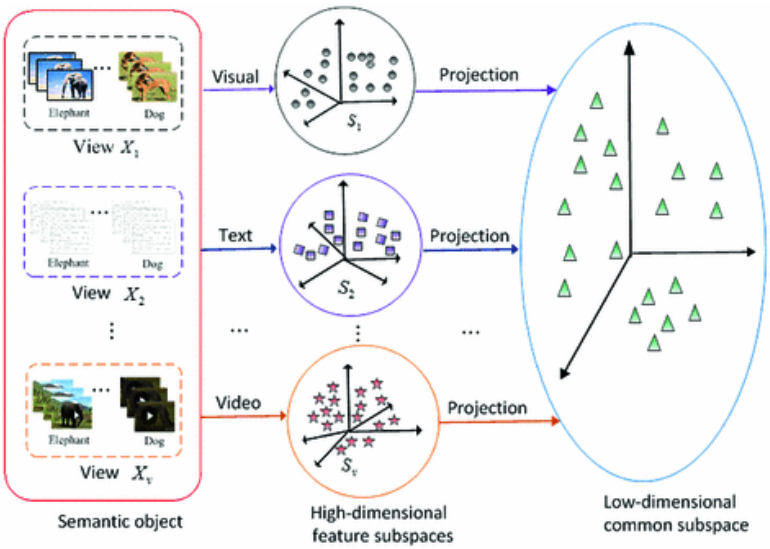


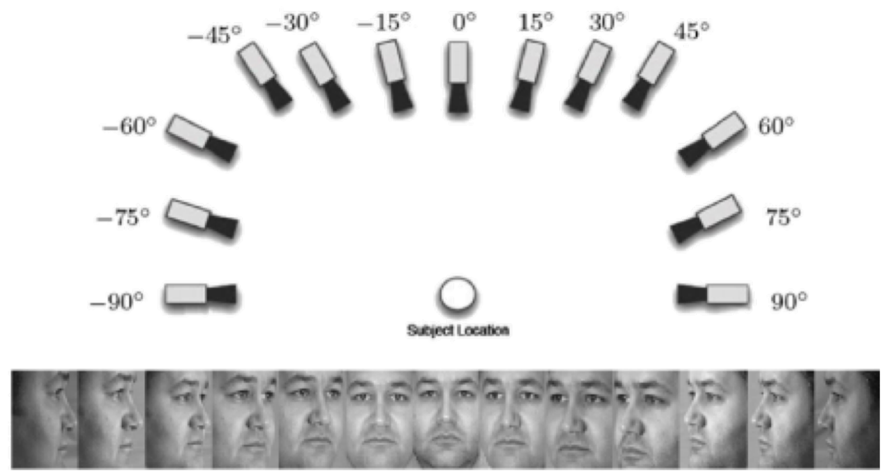
Multi-view Problems

Northeastern University

Smile lab — Synergetic Media Learning Lab

- ❑ Multi-view Projection/Embedding





- ❑ Multi-view Clustering
- ❑ Supervised Multi-view Learning
  - Multi-view Classification
  - Zero-Shot Learning

This is a cat. Its name is Sam. It is grey. It is fluffy.

It has got a little head and two big yellow eyes. It has got two little ears, a brown nose and a pink mouth. It has got four long legs. It has got a long fluffy tail.

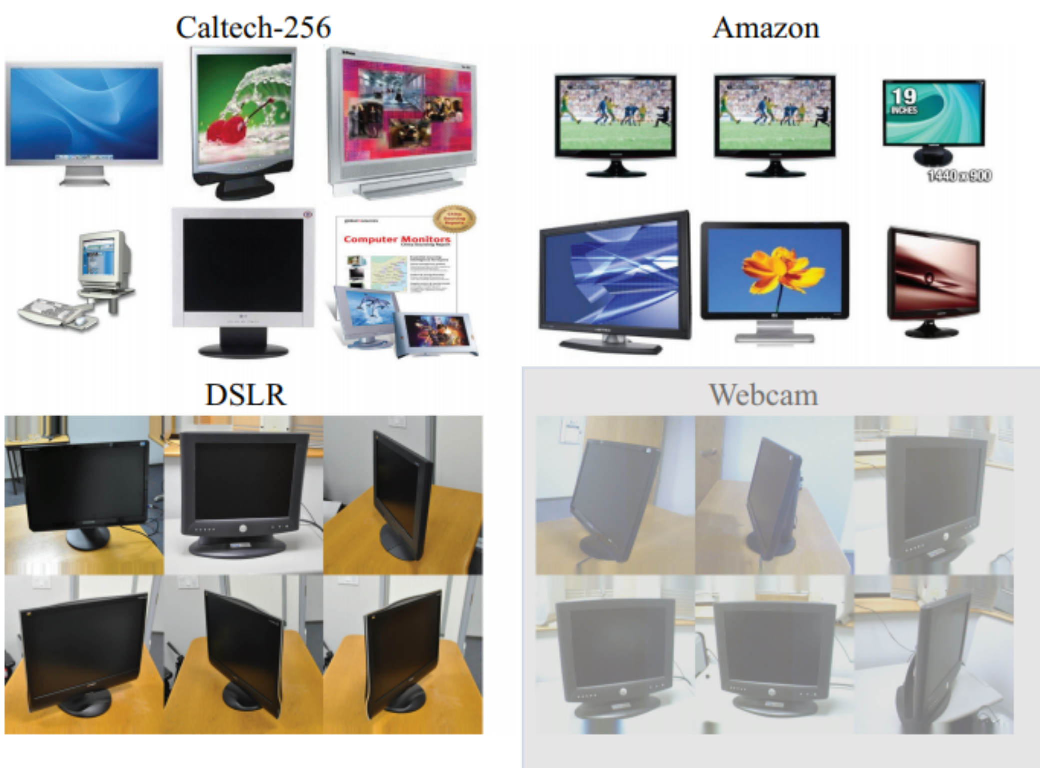
Sam is very funny. It can run, jump and climb. It can't swim and fly.

It likes fish and milk.

fluffy - пухнастий, tail - хвостик

Multi-view Problems

- ❑ Transfer Learning

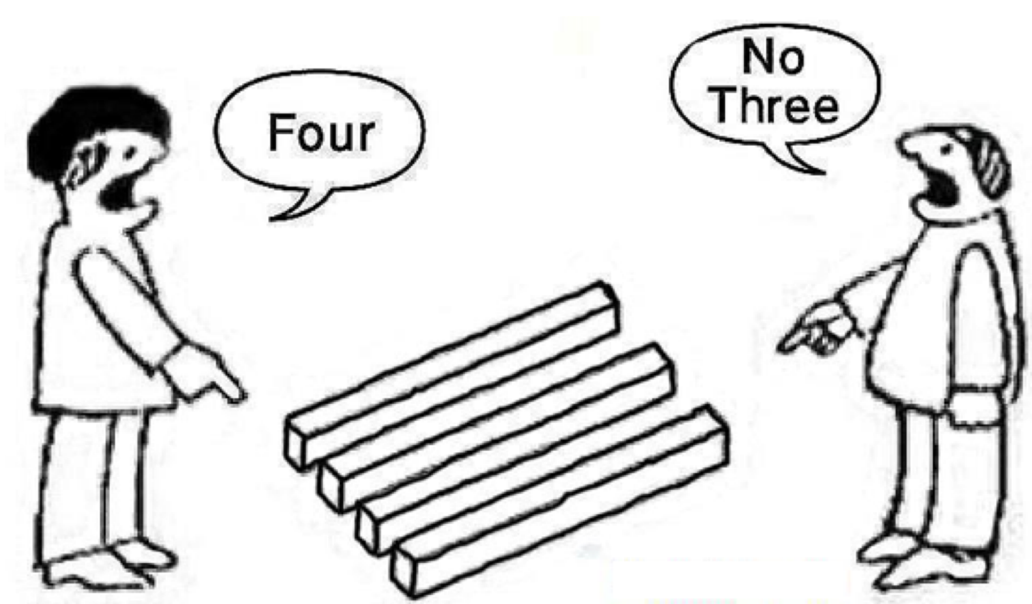


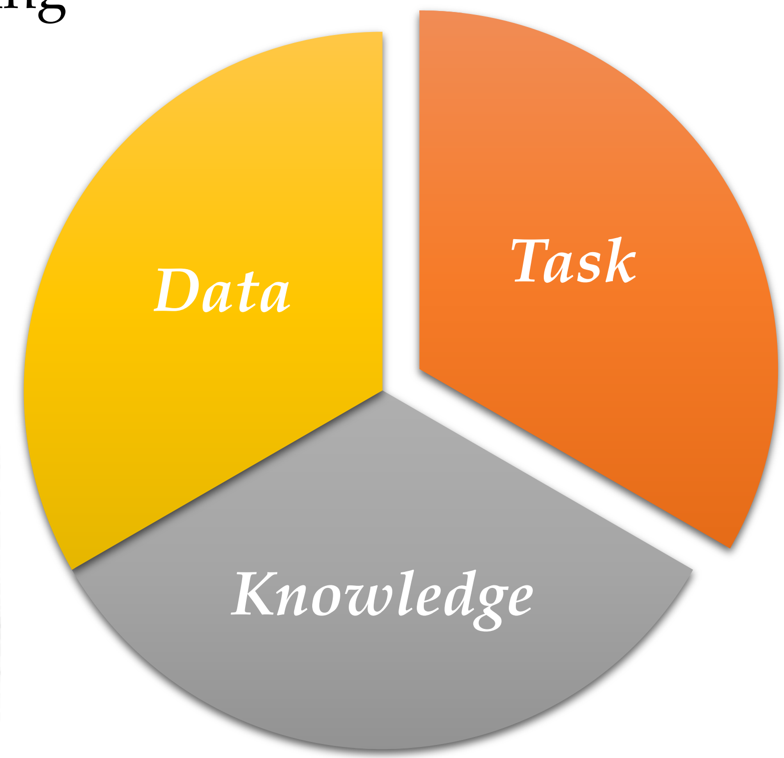
- ❑ Domain Adaptation
  - → Multi-source Domain Adaptation
  - → Domain Generalization

Multi-view Learning: How to Categorize??

Northeastern University

- Different Data and Different Learning Problems, and possibly more…
- We will offer a better Taxonomy:





Taxonomy
[*Data View*]

Northeastern University
Synergetic Media Learning Lab

- **Category 1 [Sample-Wise Correspondence] (Multi-view Learning)**
  - **Multiple Features,** e.g., LBP, SIFT, HOG…

    Goal: fuse various knowledge from multiple features to boost the final tasks

  - **Multi-Modal Visual Data**

    Goal: seek a view-invariant space to mitigate the view divergence to facilitate the final task

    *(**adapt** knowledge across different views)*

- **Category 2 [Class-wise Correspondence] (Transfer learning, Projection/Embedding)**
  - **Multi-Feature/Multi-Pose/Multi-Modal Visual data**

    Goal: transfer knowledge from well-labeled source views to unlabeled target views

Taxonomy
[*Task View*]

Northeastern University

Synergetic Media Learning Lab

- **Unsupervised Learning [Clustering, Projection/Embedding]**

  Goal: fuse various knowledge from multiple features

  Data: multiple features [***unlabeled, sample-wise correspondence***]

- **Supervised Learning [Projection/Embedding, Classification]**

  Goal: seek a view-invariant space to mitigate the view divergence to facilitate the final task *(**adapt** knowledge across different views)*

  - **Sub-Category 1 (Multi-view Learning) [sample-wise correspondence]**

    Training Stage: multiple labeled view data

    Test Stage: some labeled views, some unlabeled views

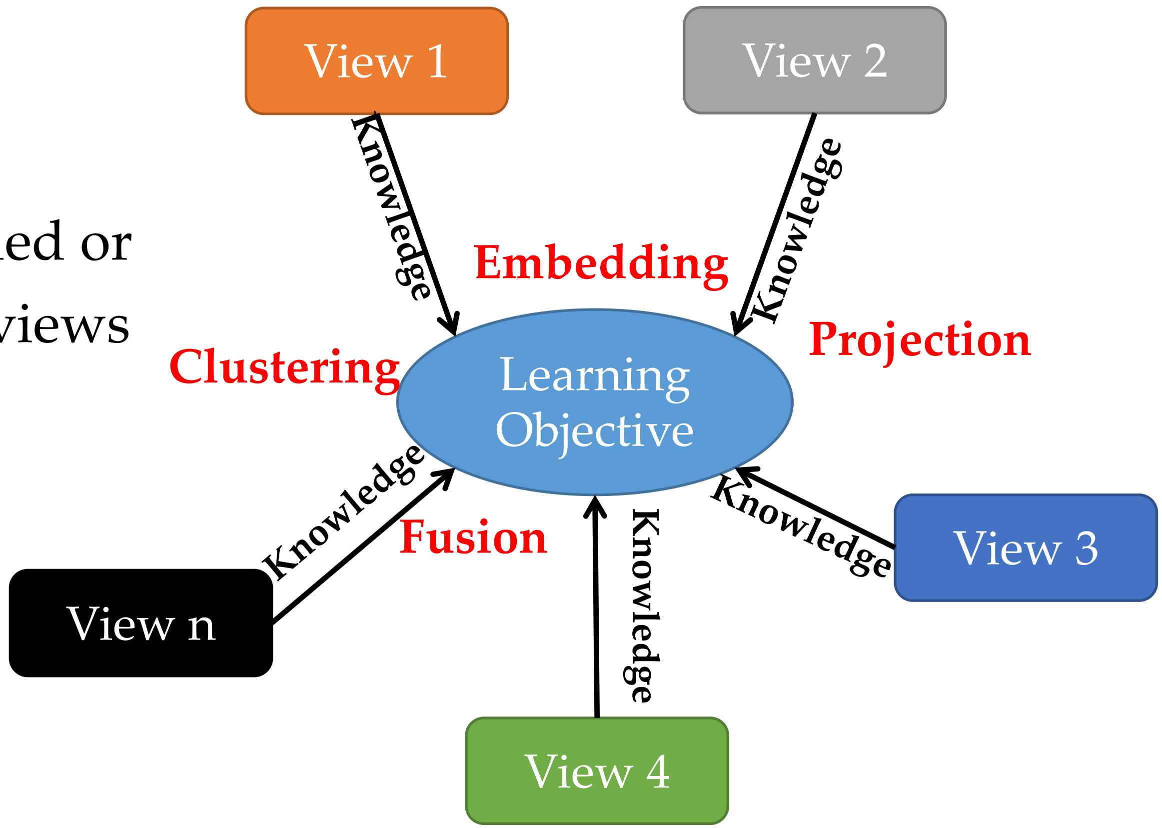
  - **Sub-Category 2 (Transfer learning) [class-wise correspondence]**

    Training Stage: some source labeled views & some target unlabeled views

Taxonomy [Knowledge View]

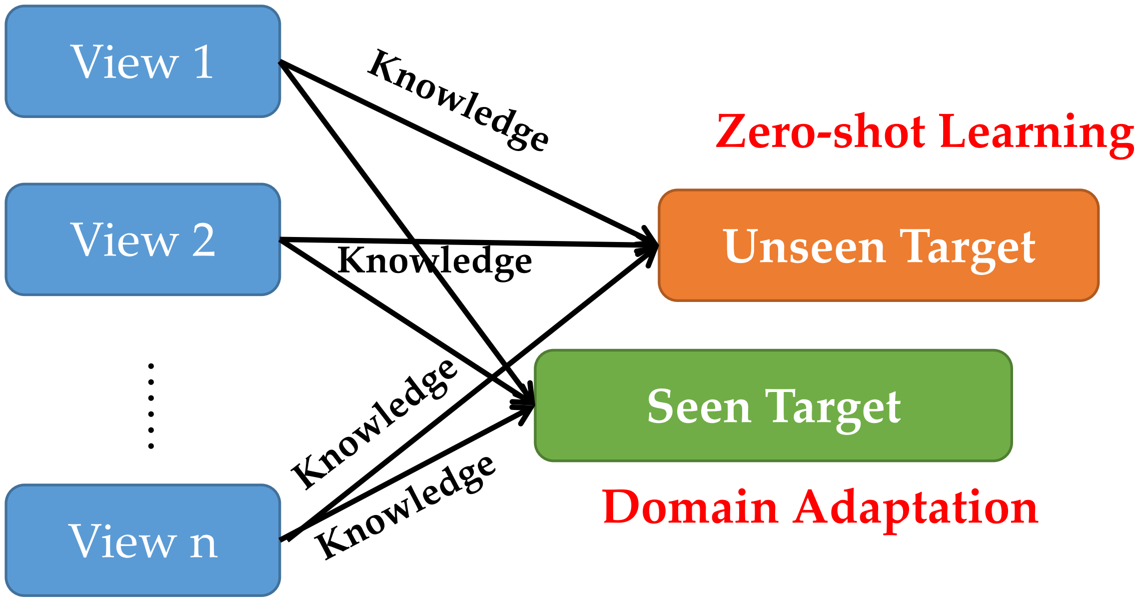
Northeastern University

- Knowledge integration from different views
  - **Goal:** Find a way to better integrate knowledge (labeled or unlabeled) from different views for a common aim:
    - Multi-view Clustering
    - Multi-view Projection and Embedding
    - Knowledge Fusion



Taxonomy [Knowledge View]

- ❑ Knowledge transfer from one view(s) to another view(s)
- **Goal:** Reuse knowledge from well-established data in a new problem (seen or unseen)
  - Transfer Learning
  - Domain Adaptation
  - Zero-shot Learning



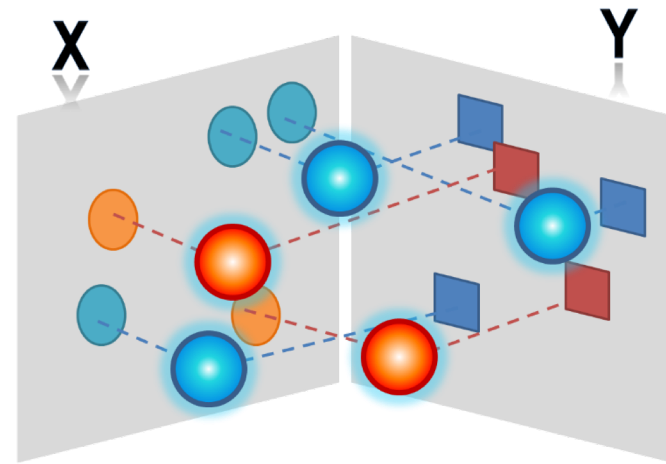
Multi-view Algorithms

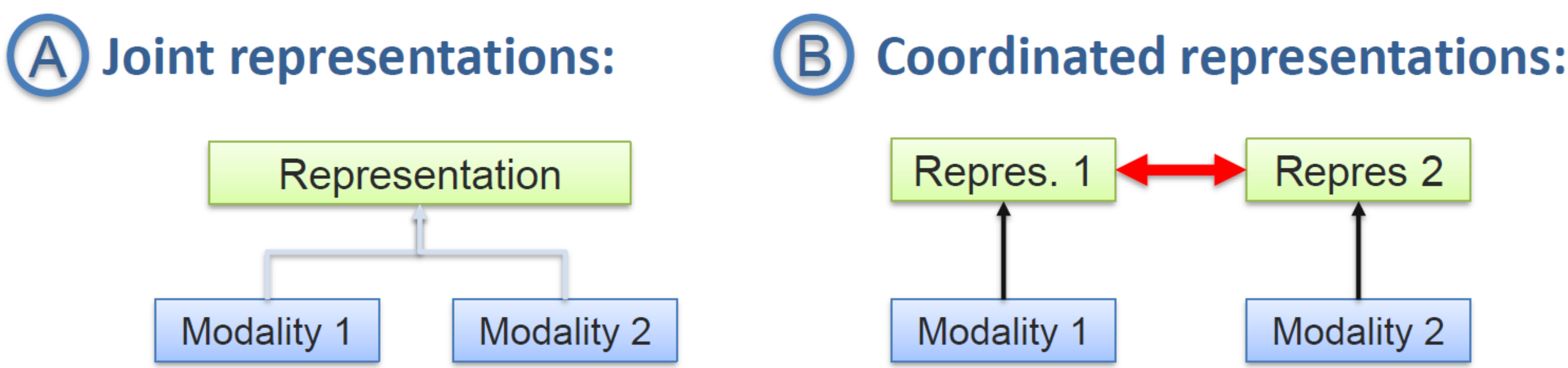
Northeastern University

Smile lab — Synergetic Media Learning Lab

❖ Representation learning

- Joint representations
- Coordinated representations
- Projection and embedding
- Similarity metrics, dictionary learning
- Multi-view AutoEncoder, CNN, GAN



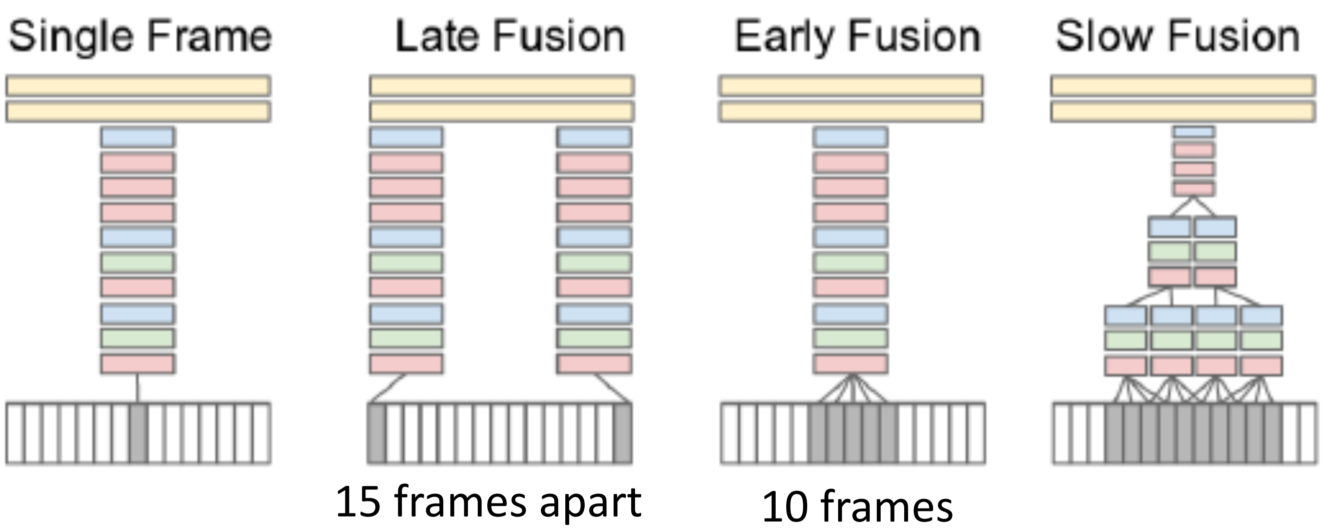


Tadas Baltrusaitis, Chaitanya Ahuja, Louis-Philippe Morency:
Multimodal Machine Learning: A Survey and Taxonomy. CoRR abs/1705.09406 (2017)

Multi-view Algorithms

❖ Fusion

To join information from two or more views to perform a prediction task



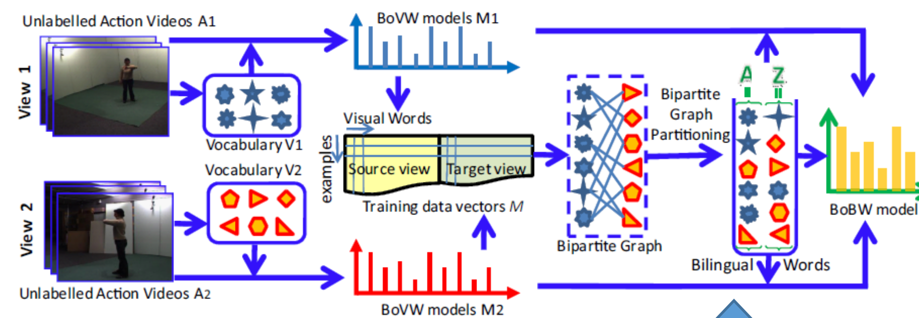
[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Fei-Fei Li: Large-Scale Video Classification with Convolutional Neural Networks. CVPR 2014: 1725-1732

[2] Vrigkas, Michalis, Christophoros Nikou, and Ioannis A. Kakadiaris. "A review of human activity recognition methods." *Frontiers in Robotics and AI* 2 (2015): 28.

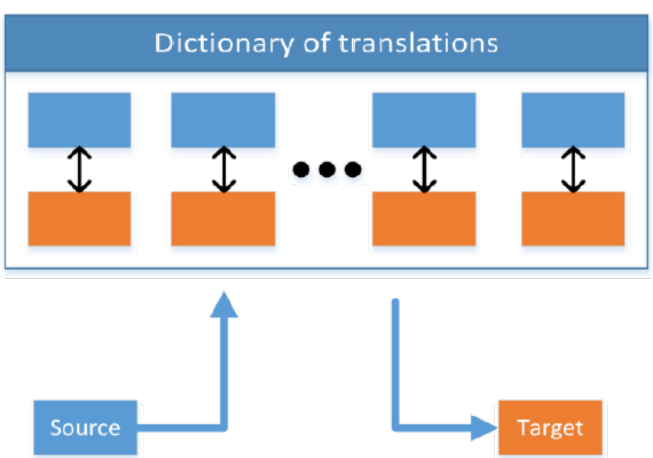
Multi-view Algorithms

❖ Translation

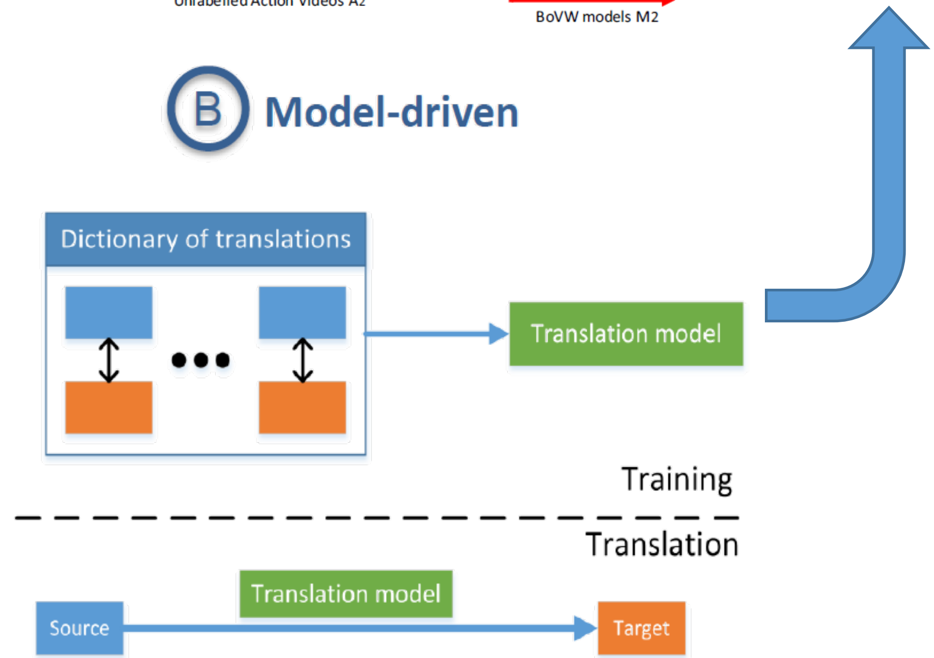
Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.



A Example-based



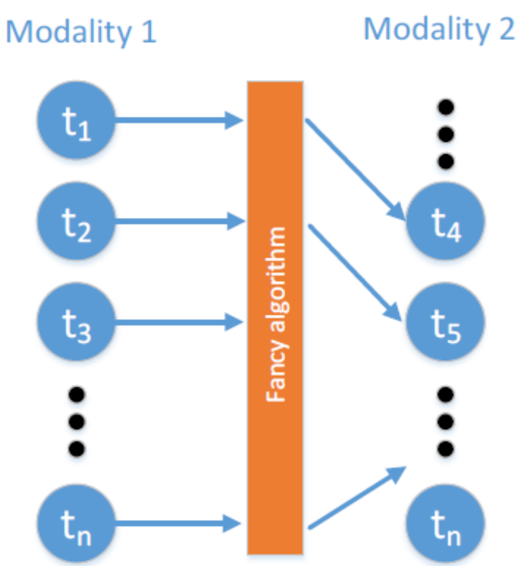
B Model-driven

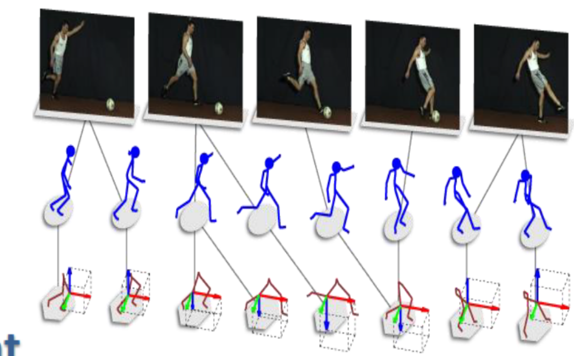


Multi-view Algorithms

❖ Alignment

Identify the direct relations between (sub)elements from two or more different modalities.



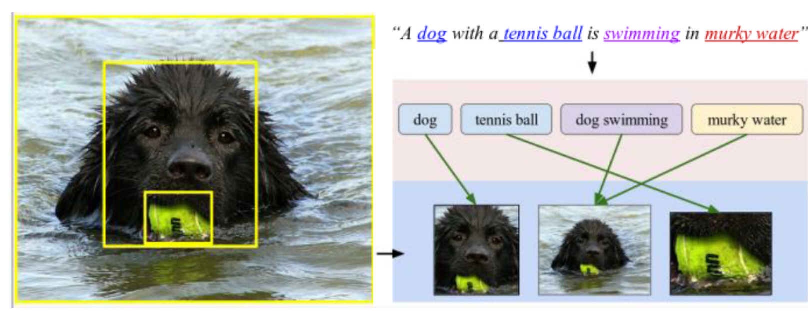


A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

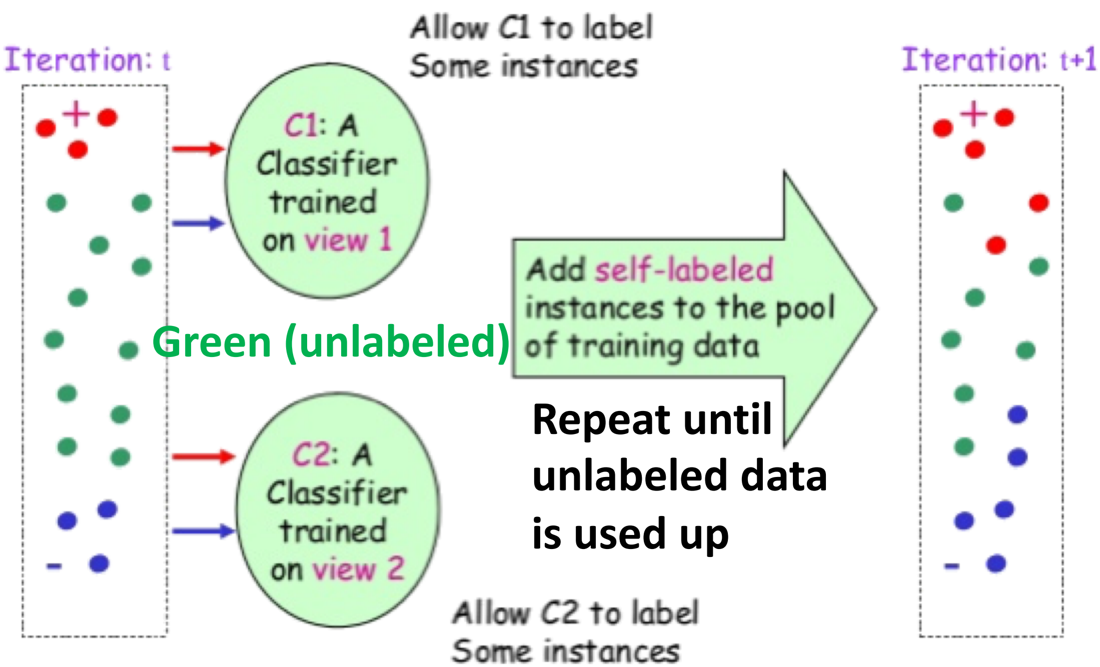
Uses internally latent alignment of modalities in order to better solve a different problem



Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Multi-view Algorithms

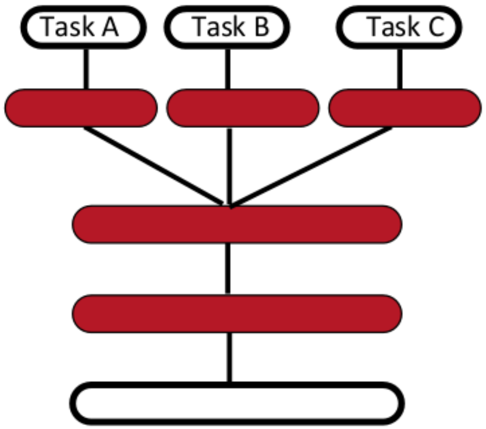
Co-Learning & Multi-task Learning



Green (unlabeled)

Repeat until unlabeled data is used up

Two views are conditionally independent



Learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better

Unified Model

- The Tutorial will focus on the

  Representation Learning and Knowledge Fusion/Alignment

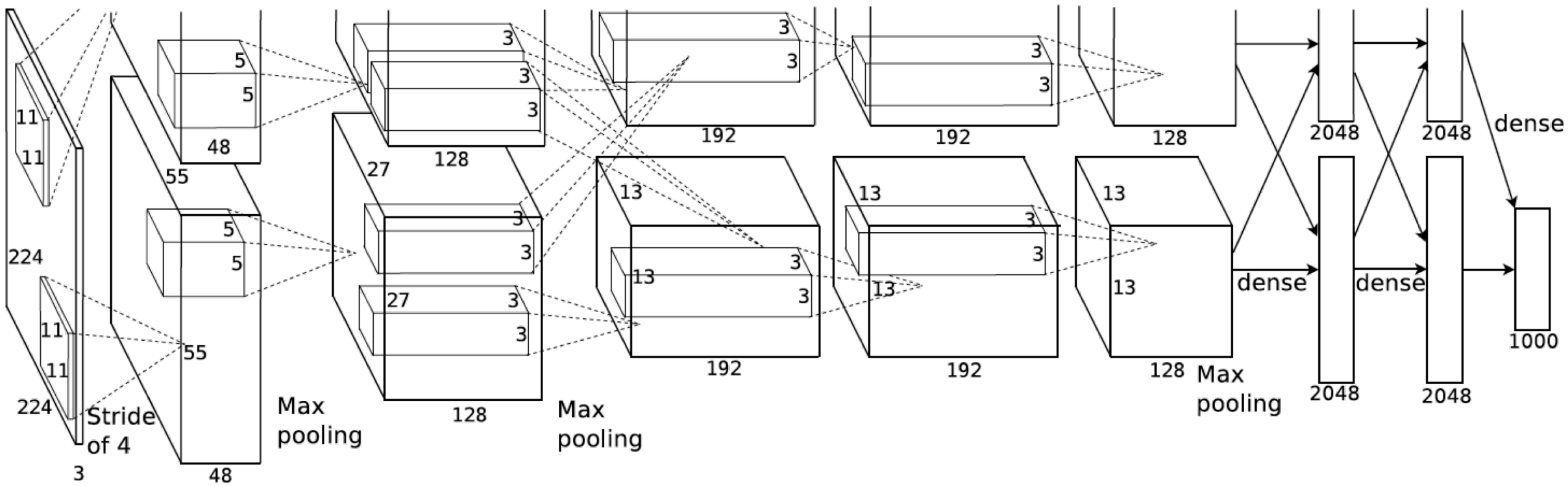
- Assume we have more than one data domains: $X_1, X_2, \ldots X_i$. In general, the discussed methods are organized in three lines:

  - Modeling Features Representation for each source: $f_i(X_i)$
  - Modeling Coherence/Alignment between different sources: $A(X_i, X_j)$
  - Regularization term regarding the label information and data underlying distribution, and semantics: $R(X_i)$

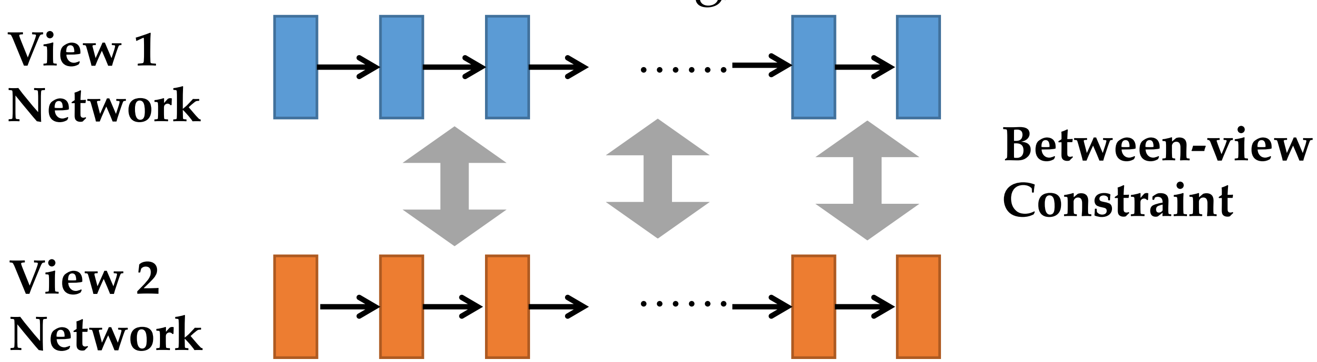
$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \mathcal{A}(f_i(X_i), f_j(X_j)) + \lambda \sum_{k=1}^{v} \mathcal{R}(f_k(X_k))$$

Unified Model

- We barely consider the deep features in the learning process…



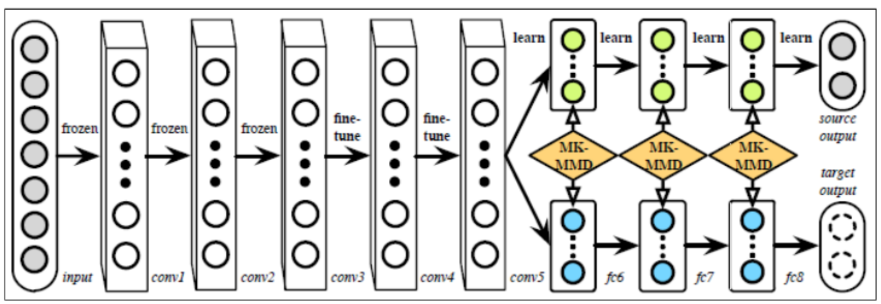
- When consider multi-view learning, what that would be??



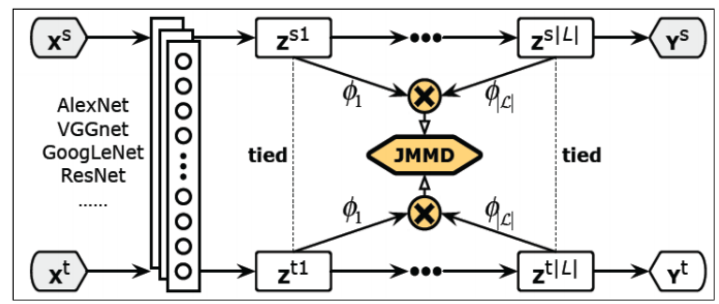
Unified Model

❑ From Shallow to Deep Learning

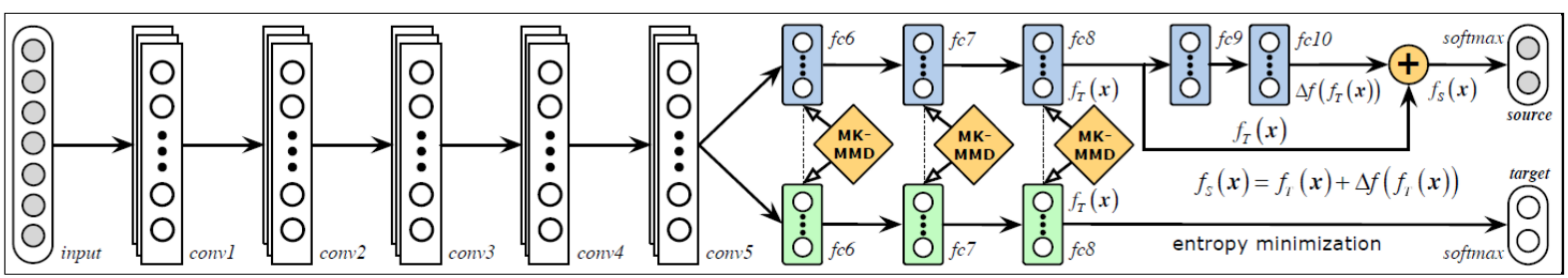
- Redistribute the constraints above into the deep structure, then which layer??
- Usually modeled on top layers, but may be from the first layer
- Multiple networks to model different aspects, e.g., one for different domains, the other for different classes



(a) The Deep Adaptation Network (DAN) architecture



(b) The Joint Adaptation Network (JAN) architecture



(c) The Residual Transfer Network (RTN) architecture

Unified Model

Northeastern University
Smile lab
Synergetic Media Learning Lab

Representation Learning + Multi-view Alignment (Fusion)

$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \boxed{\mathcal{A}(\boxed{f_i(X_i)}, f_j(X_j))} + \lambda \sum_{k=1}^{v} \mathcal{R}(f_k(X_k))$$

Linear Mapping → Kernel → Tensor

Dictionary Learning (Sparse/Low-Rank Coding)

Auto-Encoder & Neural Networks

Convolutional Neural Networks

- **Category 1 [Sample-Wise Correspondence] (Multi-view Learning)**
- **Category 2 [Class-wise Correspondence] (Transfer learning)**
  - Maximum Mean Discrepancy
  - Reconstruction-based Alignment
  - Adversarial loss [0/1]

Unified Model

Northeastern University

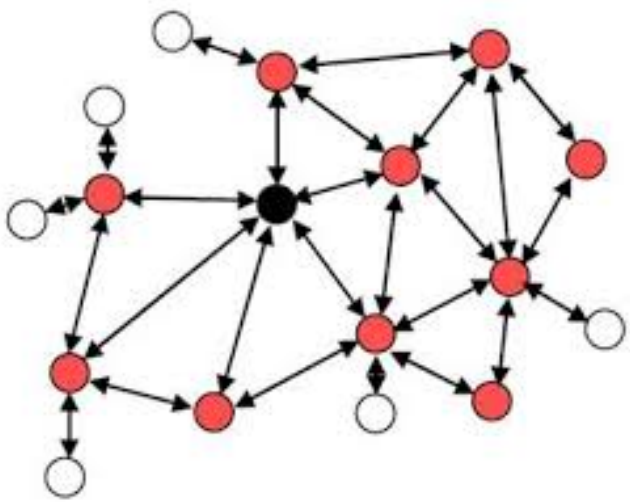
Smile lab

Synergetic Media Learning Lab

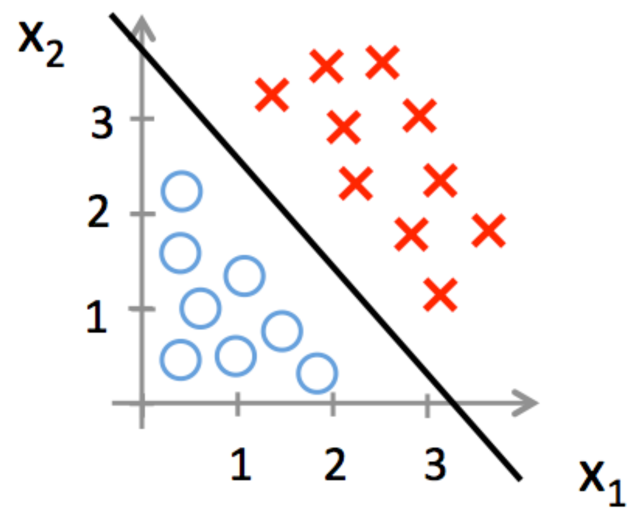
Representation Learning + Multi-view Alignment (Fusion)

$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \mathcal{A}(f_i(X_i), f_j(X_j)) + \lambda \sum_{k=1}^{v} \boxed{\mathcal{R}(f_k(X_k))}$$

Graph Regularizer



Regression model



Outline

Northeastern University

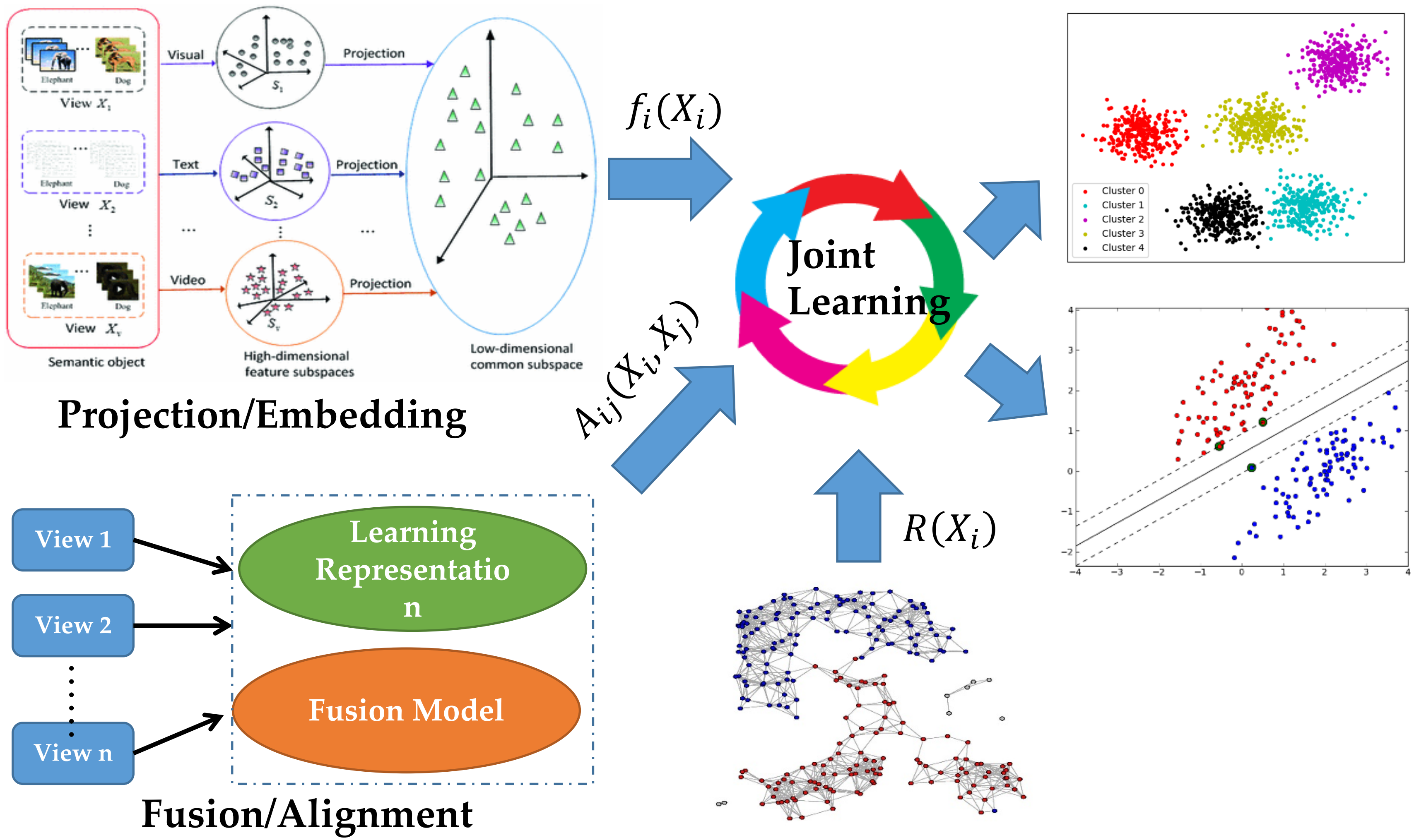
- **Introduction & Background**
  - Multi-view Visual Data
  - Multi-view Learning Problems
  - Multi-view Learning Taxonomy
- **Multi-view Learning**
  - Projection and Embedding
  - Knowledge Fusion
  - Multi-view Clustering
  - Supervised Multi-view Learning → Zero-shot Learning
- **Domain Adaptation**
  - Transfer Learning → Domain Adaptation
  - Domain Generalization → Zero-shot Learning
- **Conclusion**

Multi-view Learning RoadMap

Northeastern University

Smile lab

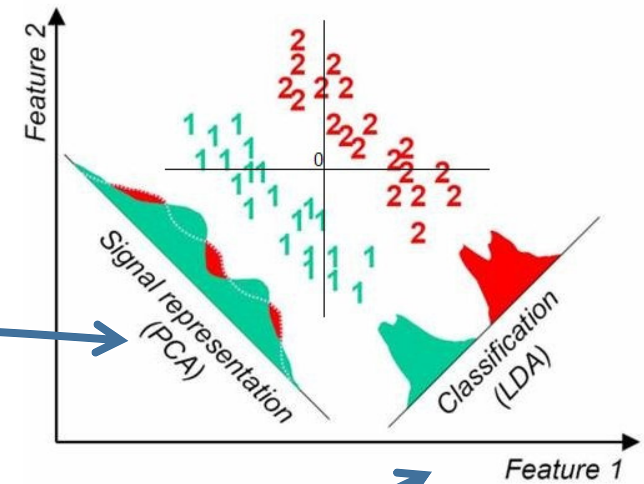
Synergetic Media Learning Lab

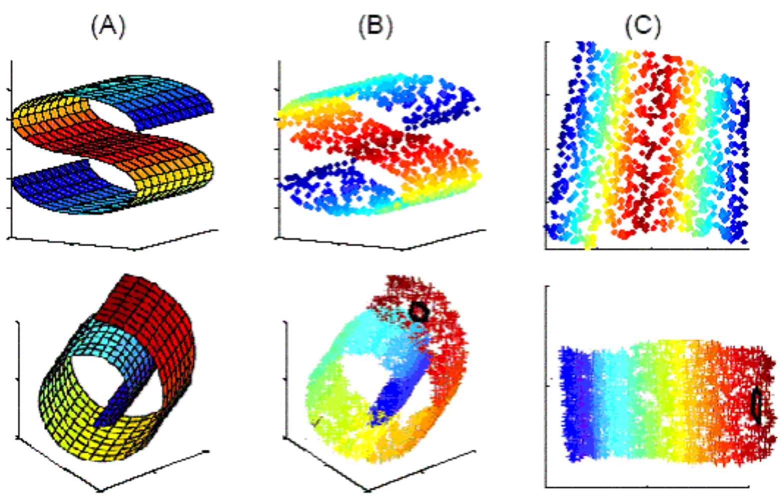


Background (Dimensionality Reduction)

Northeastern University

- **Projection**:
  - Given the input data $X \in \mathrm{R}^{n \times D}$, find a Linear Mapping: $f(X) = Y$ that will map X into a lower dimensional space $Y \in \mathrm{R}^{n \times d}$, and f is usually a projection matrix.
  - Typical methods: PCA, LDA, LPP





- **Nonlinear Embedding:**
  - Given the input data $X \in \mathrm{R}^{n \times D}$, find a function: $f(X) = Y$ that will map X into a lower dimensional space $Y \in \mathrm{R}^{n \times d}$, and f is usually a mapping with implicit formulation.
  - Typical methods: LLE, ISOMAP

Multi-view Projection: Canonical Correlation Analysis (CCA)

Northeastern University

Smile lab

Synergetic Media Learning Lab

Canonical Correlation Analysis (CCA)

Formally, for two views $X \in R^{d\times n}$ and $Y \in R^{k\times n}$, CCA computes two projection vectors, $w_x \in R^d$ and $w_y \in R^k$, such that the following correlation coefficient is maximized:

$$\rho = \frac{w_x^T XY^T w_y}{\sqrt{(w_x^T XX^T w_x)(w_y^T YY^Y w_y)}}$$

Since ρ is invariant to the scaling of w_x and w_y, CCA can be formulated equivalently as

$$\max_{w_x, w_y} \quad w_x^T XY^T w_y$$

$$\texttt{s.t.} \quad w_x^T XX^T w_x = 1, \quad w_y^T YY^T w_y = 1$$

✓ It has a sense of consensus.

Relations between two sets of variates. Biometrika, 28(3/4):321–377, 1936

H. Hotelling.

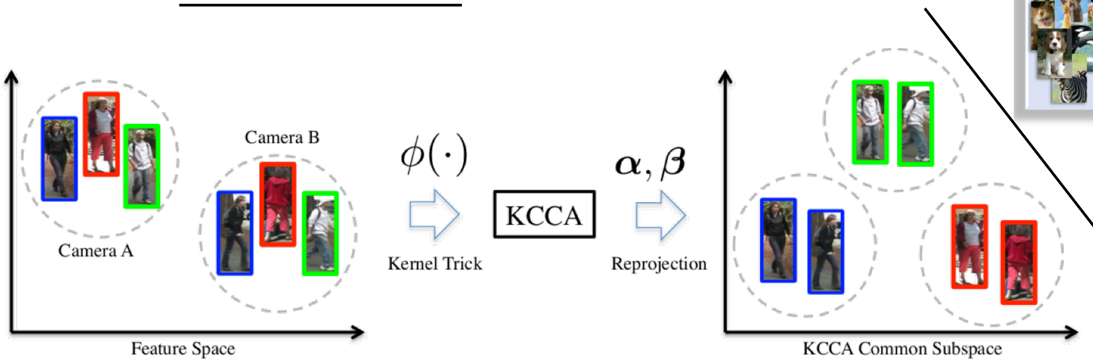
Multi-view Projection: Canonical Correlation Analysis (CCA)

Northeastern University

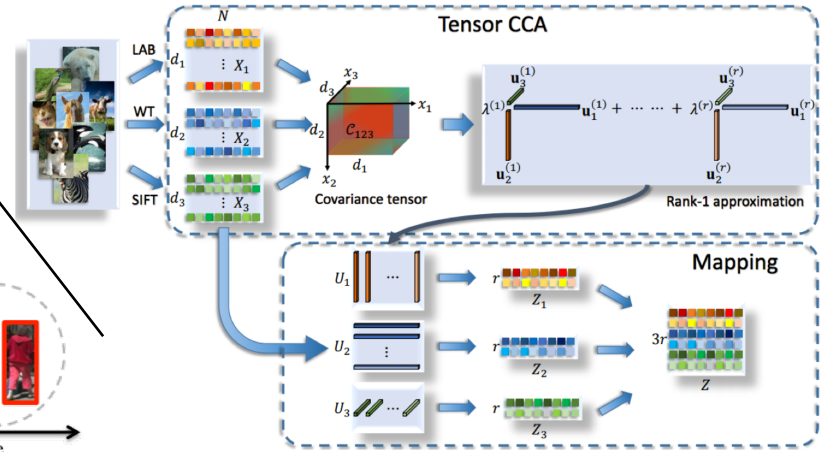
Smile lab
Synergetic Media Learning Lab

Extensions of CCA

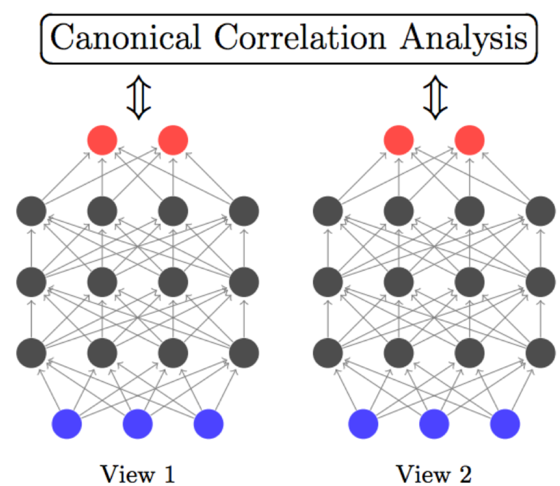
Kernel-based



Tensor-based



DNN-based



[**Kernel-based**] David R. Hardoon, Sándor Szedmák, John Shawe-Taylor: Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation 16(12): 2639-2664 (2004)

[**Tensor-based**] Tensor Canonical Correlation Analysis for Multi-view Dimension Reduction, Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen, IEEE Transactions on Knowledge and Data Engineering (T-KDE), vol. 27, no. 11, pp. 3111-3124, 2015.

[**DeepNN-based**] Galen Andrew, Raman Arora, Jeff A. Bilmes, Karen Livescu: Deep Canonical Correlation Analysis. ICML (3) 2013: 1247-1255

Multi-view Projection: Canonical Correlation Analysis (CCA)

Northeastern University
Smile lab
Synergetic Media Learning Lab

Extensions of CCA

Even more...

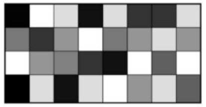
Multiple Sets

Assume we will apply CCA on multiple views (more than two)

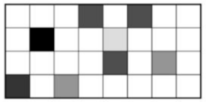
We will conduct this in a pairwise fashion: X and Y, Y and Z, Z and Y

$$\rho = \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T C_{xx} w_x)(w_y^T C_{yy} w_y)}} + \frac{w_y^T C_{yz} w_z}{\sqrt{(w_y^T C_{yy} w_z)(w_z^T C_{zz} w_z)}} + \frac{w_x^T C_{xz} w_z}{\sqrt{(w_x^T C_{xx} w_x)(w_z^T C_{zz} w_z)}}$$

Sparsity



Dense Projection



Sparse Projection

Motivated by good feature selection and stability of the features, we pursue sparsity in CCA:

$$\rho = \max_{w_x, w_y} \frac{w_x^\top C_{xy} w_y}{\sqrt{w_x C_{xx} w_x w_y C_y w_y}}$$

$$\text{s.t.} \quad ||w_x||_0 \leq s_x, \quad ||w_y||_0 \leq s_y.$$

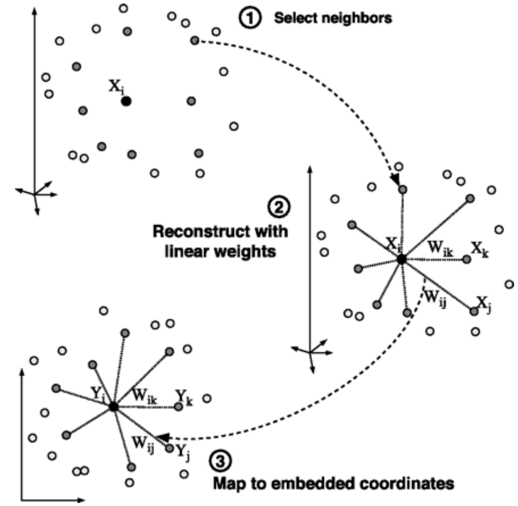
[**Multiple Sets**] Kettenring, J.R. Canonical analysis of several sets of variables. Biometrika (1971)

[**Sparsity**] A. Wiesel, M. Kliger, and A. O. Hero, III, "A greedy approach to sparse canonical correlation analysis," ArXiv e-prints, 2008.

Multi-view Embedding

- Embedding offers a way to find the low-dimensional representation through an implicit mapping
- For multi-view data, how to do embedding in a joint manner?

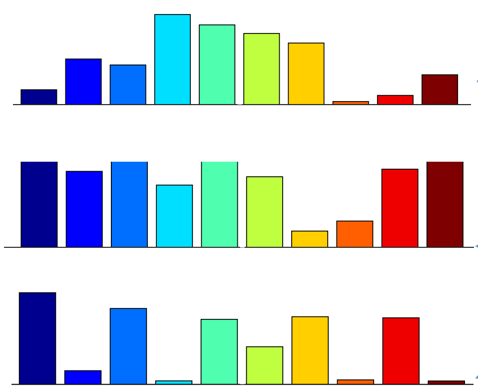
From nD to 2D



LBP

SIFT

HOG



$$\arg\min_{Y,\vec{c}} \sum_{v=1}^{V} \boxed{c_v^r} tr(\boxed{Y} L^v Y^T)$$

$$s.t.\ YY^T = I; \sum_{v=1}^{V} c_v = 1, c_v \geq 0.$$

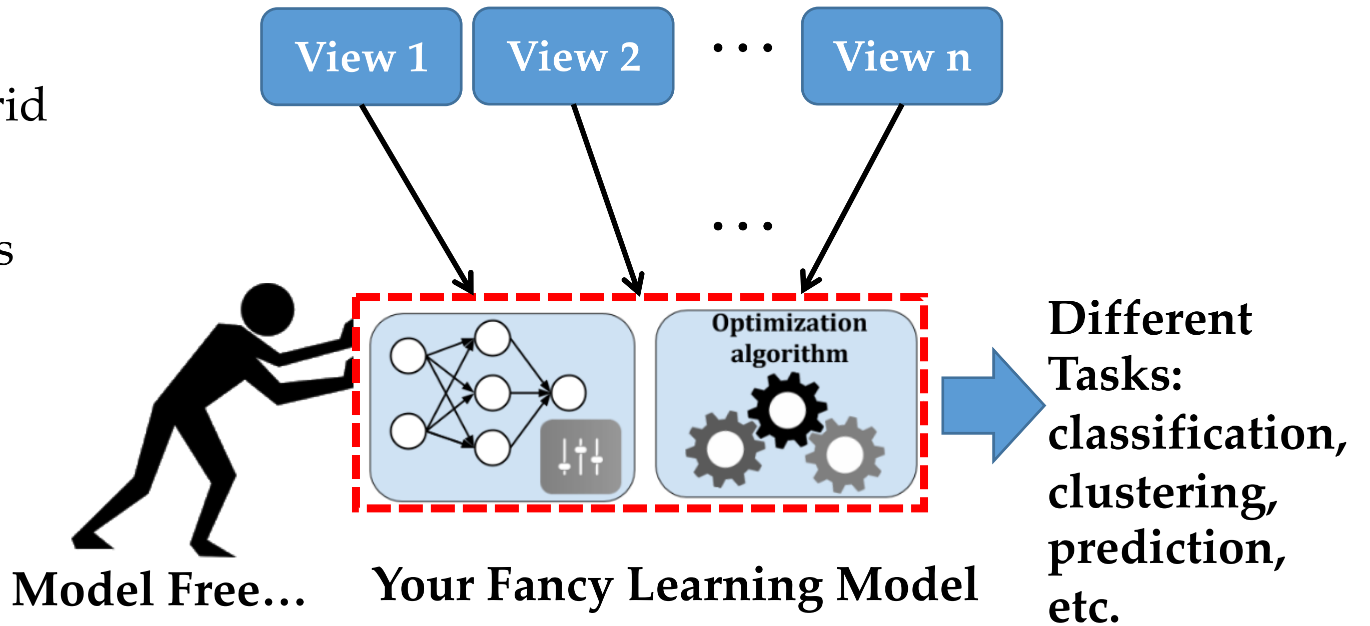
Solved in an alternative fashion

[1] Tian Xia, Dacheng Tao, Tao Mei, Yongdong Zhang: Multiview Spectral Embedding. IEEE Trans. Systems, Man, and Cybernetics, Part B 40(6): 1438-1446 (2010)

[2] Shen H, Tao D, Ma D (2013) Multiview Locally Linear Embedding for Effective Medical Image Retrieval. PLoS ONE 8(12): e82409.

Knowledge Fusion

- Process of joining information from two or more modalities to perform a prediction
- One of the earlier and more established problems, e.g. audio-visual speech recognition, multimedia event detection, multimodal emotion recognition
- Model Free
  - Early, late, hybrid
- Model Based
  - Kernel Methods
  - Graph



[1] Tadas Baltrušaitis, Chaitanya Ahuja, Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy, https://arxiv.org/abs/1705.09406

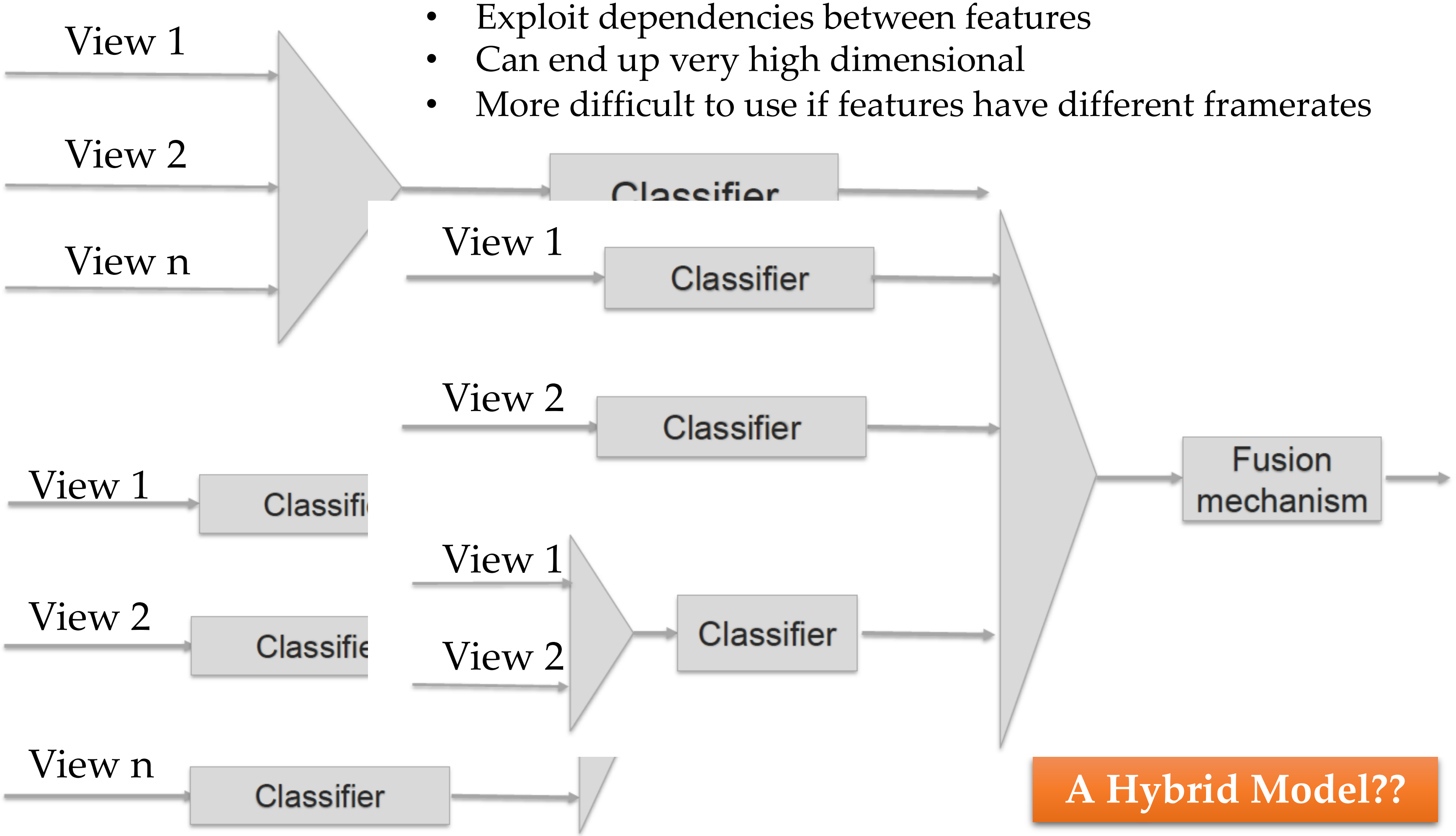
Model Free Approaches – Early vs. Late Fusion

Northeastern University

Smile lab

Synergetic Media Learning Lab

- Easy to implement – just concatenate the features
- Exploit dependencies between features
- Can end up very high dimensional
- More difficult to use if features have different framerates



A Hybrid Model??

Northeastern University

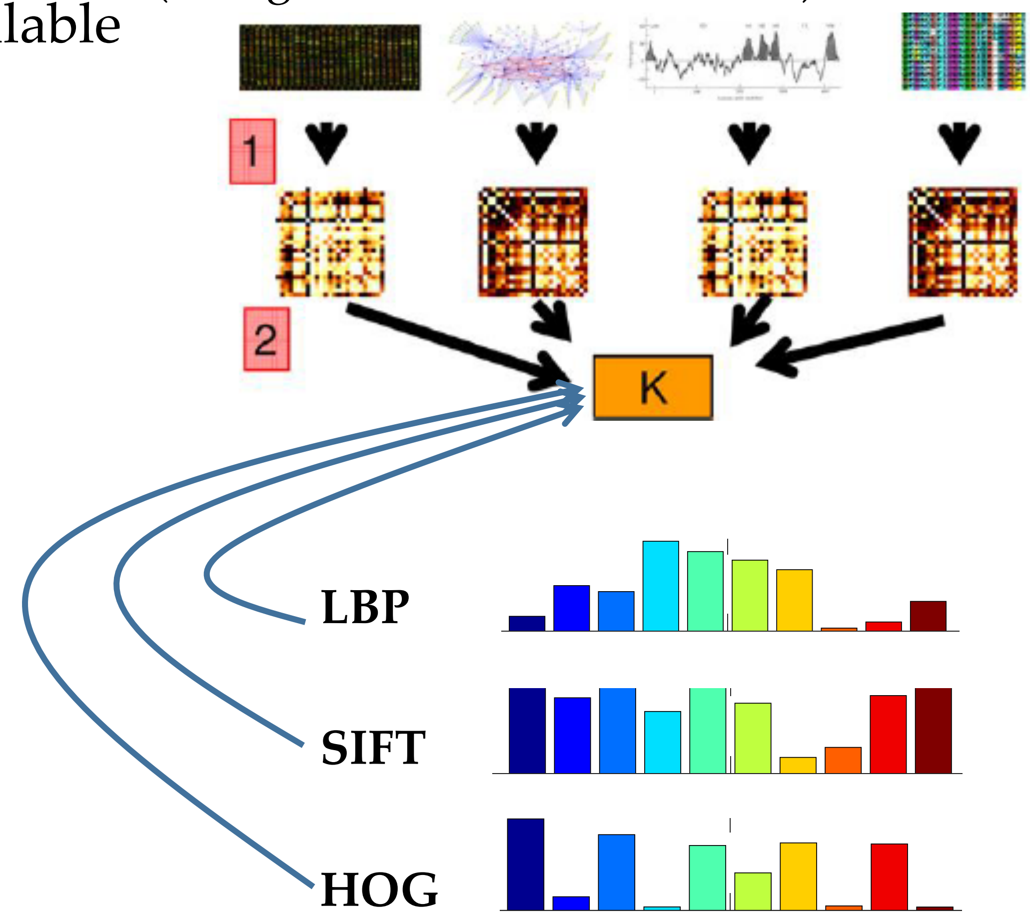
Multi-Kernel Learning

- The overall MKL framework:
- Extract features from all available sources
- Construct kernel matrices
  - Different features
  - Different kernel types
  - Different kernel parameters
- Find the optimal kernel combination and the kernel classifier, e.g., SVM

$$\mathbf{K}(\boldsymbol{\beta}) = \sum_{j=1}^{s} \beta_j \mathbf{K}_j$$

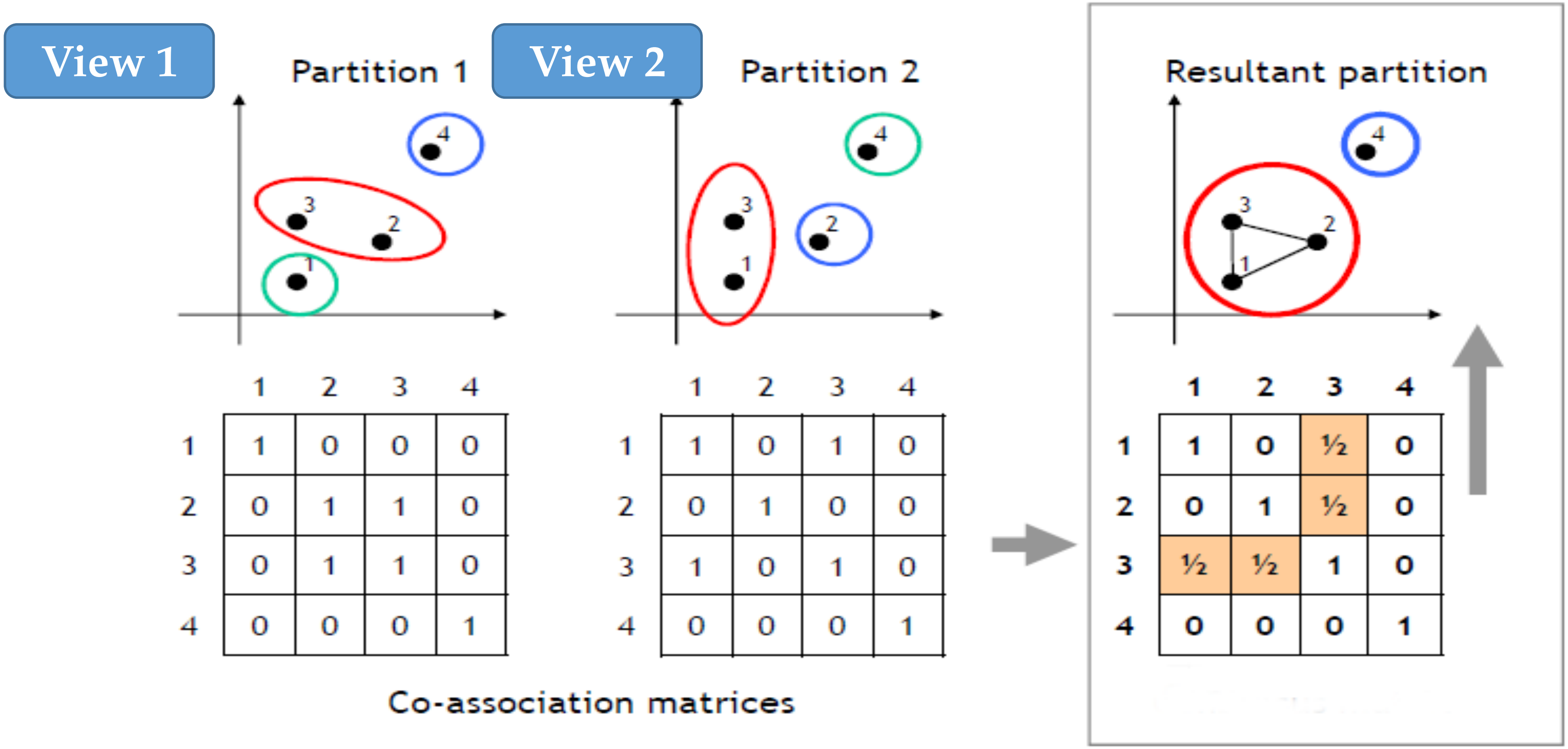
The new kernel is a linear combination of different kernels

Create individual kernels for each source (string kernel, diffusion kernel)



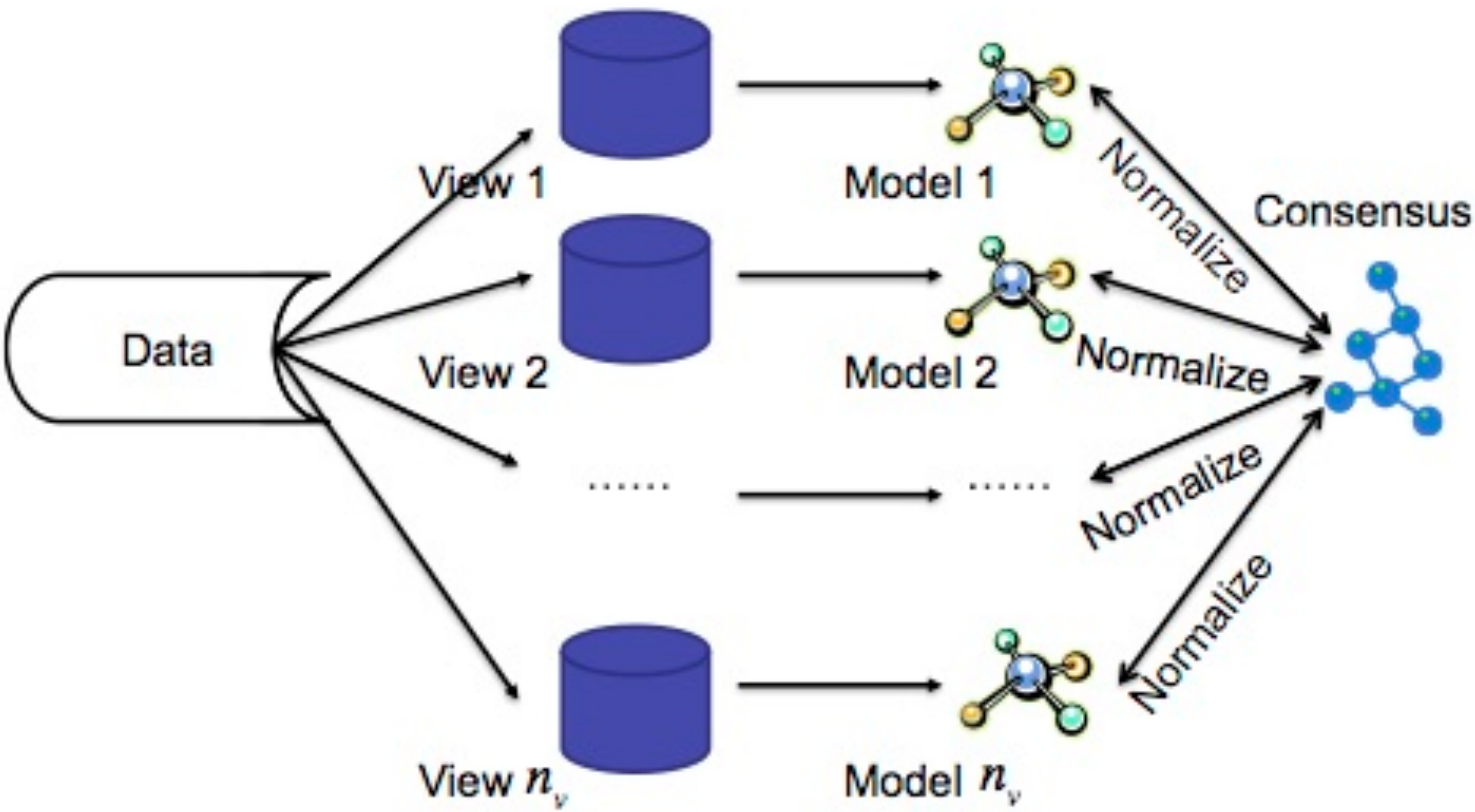
Graph based Fusion

- We may have more than one partitions from different views or features, and fusing them using graph (co-association matrix) is straightforward



Fred & Jain, 2005; Iam-On et. al, 2008; Vega-Pons & Ruiz-Shulcloper, 2009; Wang et. al, 2009; Li et. al, 2007; etc

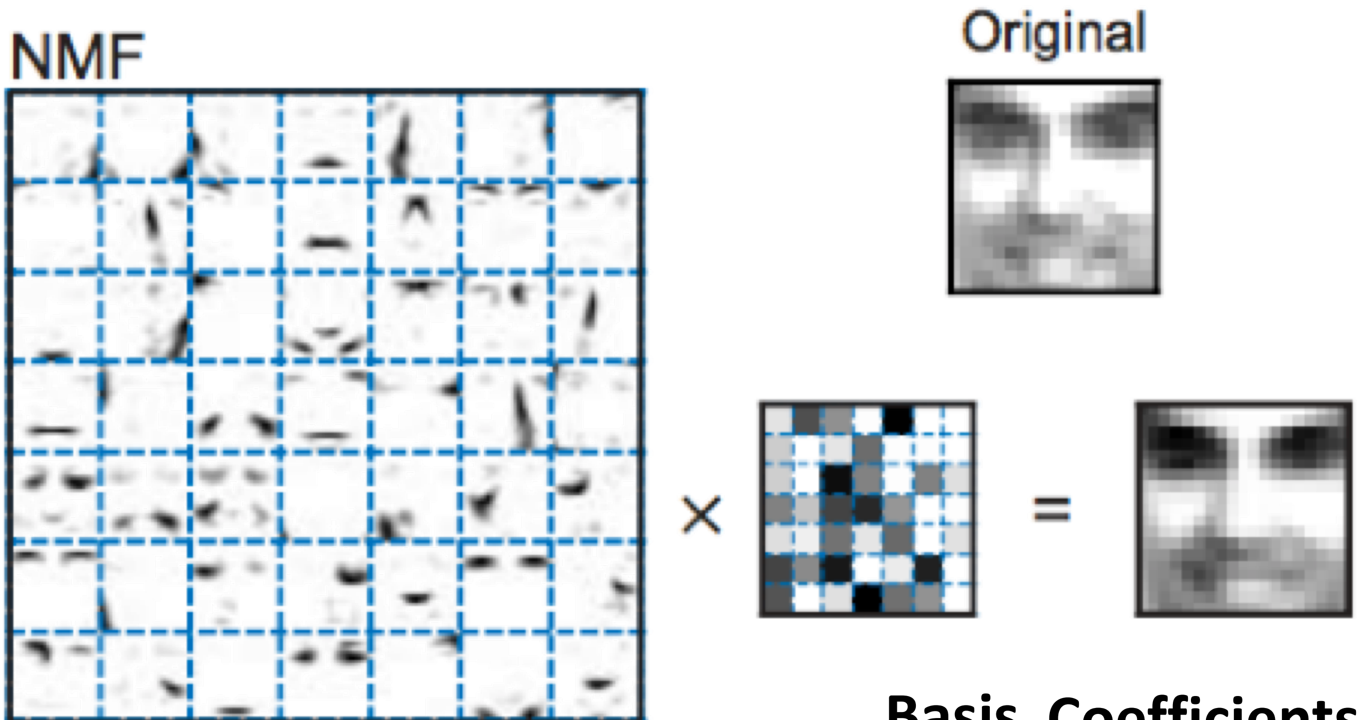
Multi-view Clustering



Methodology

- Multi-view Clustering

Multi-view Clustering via Joint **Nonnegative Matrix Factorization**



Basis **Coefficients**

Objective function: $\min_{U,V} \|\underset{M\times N}{\boxed{X}} - UV^T\|_F^2, \; s.t. \; \underset{M\times K}{\boxed{U}} \geq 0, \underset{N\times K}{\boxed{V}} \geq 0$

Learning the parts of objects by non-negative matrix factorization – Nature'99

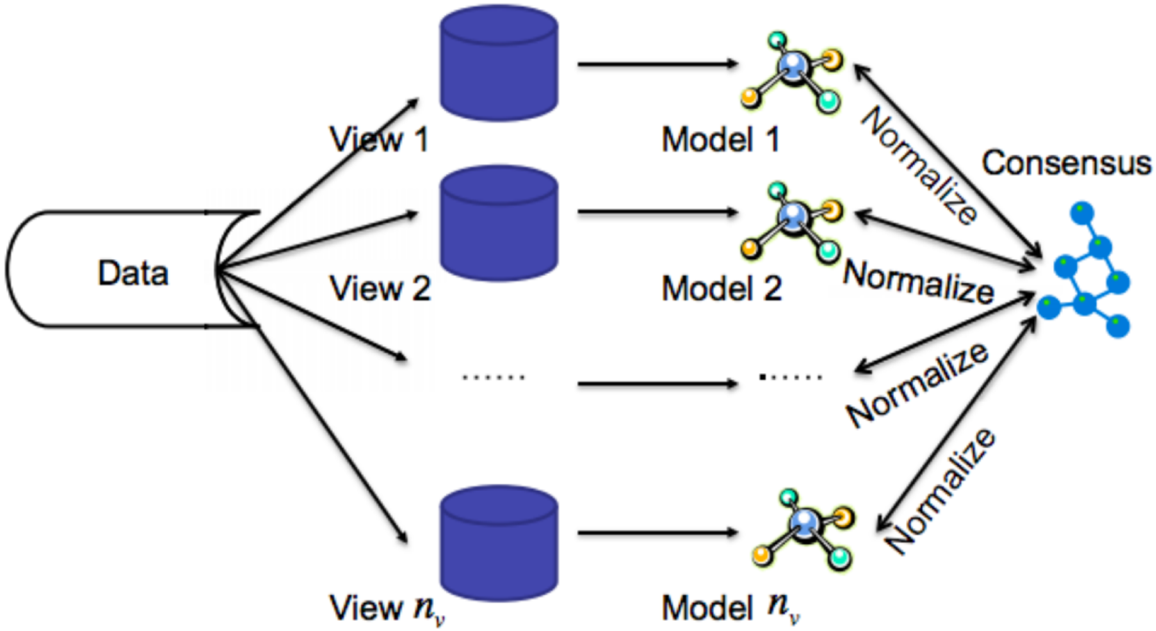
Daniel D Lee, H Sebastian Seung

Methodology

- Multi-view Clustering

Northeastern University

Multi-view Clustering via Joint Nonnegative Matrix Factorization



NMF has a good interpretability, and it is reported to achieve competitive performance compared with most of the state-of-the-art unsupervised algorithms.

The latent representations $V^{(v)}$ in different views are forced to be close to the consensus one V^*.

Objective function:

$$\sum_{v=1}^{n_v} \boxed{\|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2} + \boxed{\sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2}$$

$$s.t. \quad \forall 1 \le k \le K, \|U^{(v)}_{\cdot,k}\|_1 = 1, \boxed{U^{(v)}, V^{(v)}, V^* \ge 0}$$

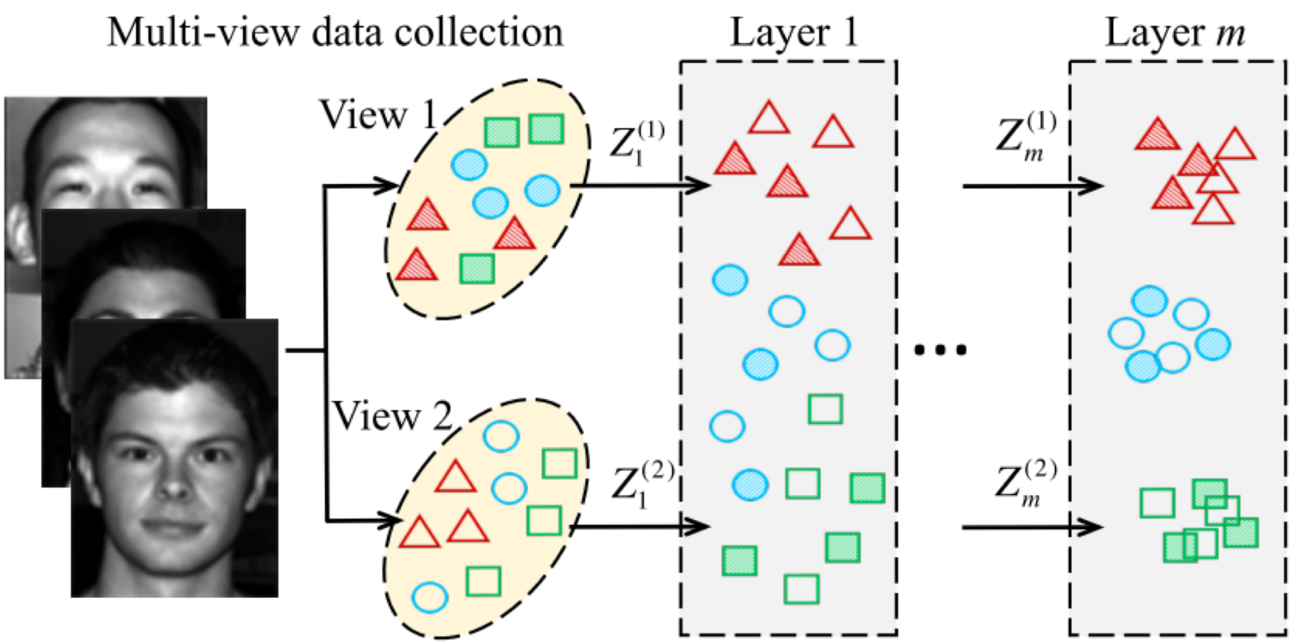
Multi-View Clustering via Joint NMF – SDM’13

Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han

Methodology
- Multi-view Clustering

Northeastern University

Multi-View Clustering via Deep Matrix Factorization



Motivation:

To learn the hierarchical semantics of multi-view data in a layer-wise fashion, semi-nonnegative matrix factorization is adopted.

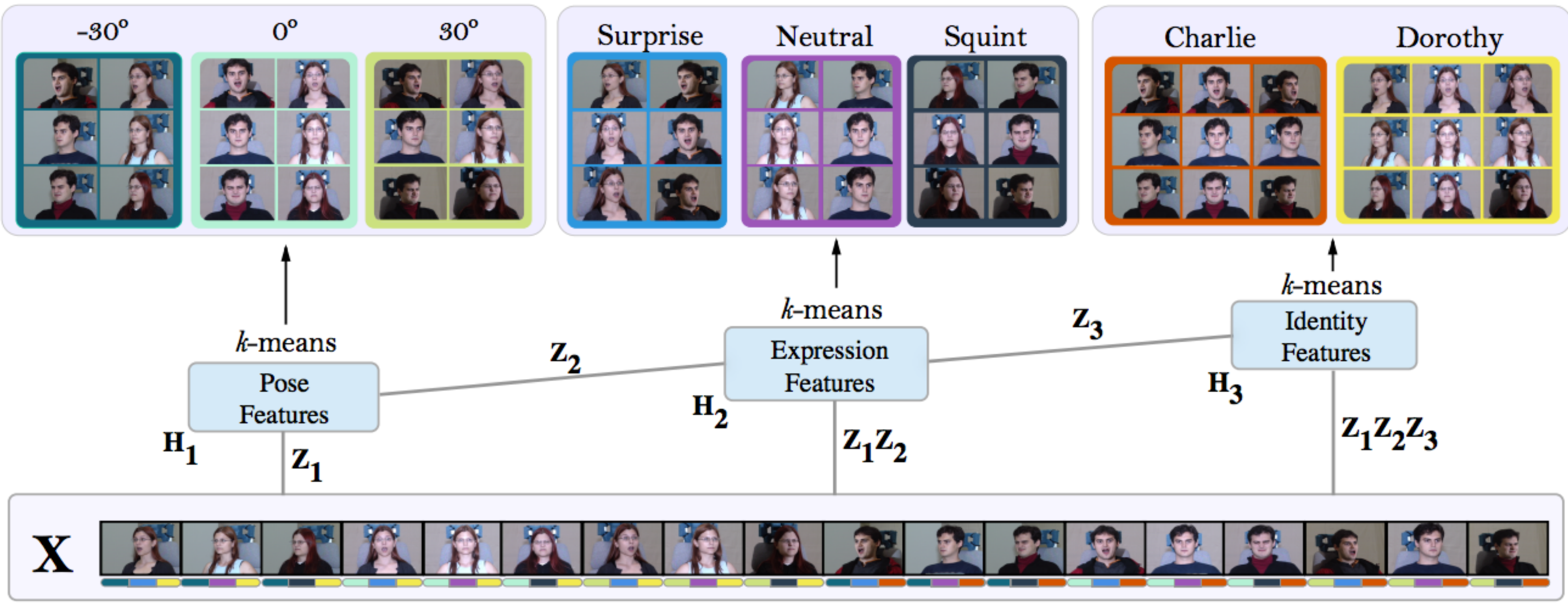
Methodology

- Multi-view Clustering

Northeastern University

Smile lab
Synergetic Media Learning Lab

Single-View Clustering via Deep Semi-NMF



Layer-wise formulation:

$$X^{\pm} \approx Z_1^{\pm} H_1^{+}$$
$$X^{\pm} \approx Z_1^{\pm} Z_2^{\pm} H_2^{+} \quad \Rightarrow \quad X^{\pm} \approx Z_1^{\pm} Z_2^{\pm} \cdots Z_m^{\pm} H_m^{+}$$
$$X^{\pm} \approx Z_1^{\pm} Z_2^{\pm} Z_3^{\pm} H_3^{+}$$

A Deep Semi-NMF Model for Learning Hidden Representations – ICML'14

George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern W.Schuller

Methodology
- Multi-view Clustering

Northeastern University

Smile lab
Synergetic Media Learning Lab

Multi-View Clustering via Deep Matrix Factorization

Objective function:

$$\min_{\substack{Z_i^{(v)}, H_i^{(v)}, \\ H_m, \alpha^{(v)}}} \sum_{v=1}^{V} (\alpha^{(v)})^\gamma \left(\boxed{\|X^{(v)} - Z_1^{(v)} Z_2^{(v)} \dots Z_m^{(v)} H_m\|_F^2} \right.$$
$$\left. + \boxed{\beta \mathrm{tr}(H_m L^{(v)} H_m^T)} \right)$$

Decomposition on all views, where the representations on the last layer $H_m^{(v)}$ are forced to be same H_m.

$L^{(v)}$ is the graph Laplacian of the graph for view v, where each graph is constructed in k-nearest neighbor fashion.

The hidden representation H are non-negative, with good interpretability.

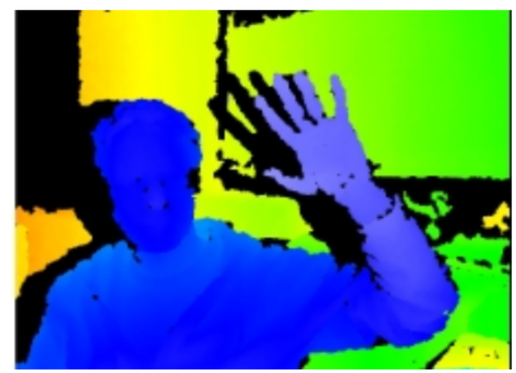
$$\text{s.t.}\ \boxed{H_i^{(v)} \geq 0,\ H_m \geq 0,}\ \sum_{v=1}^{V} \alpha^{(v)} = 1,\ \alpha^{(v)} \geq 0$$

Multi-View Clustering via Deep Matrix Factorization – AAAI'17
Handong Zhao, Zhengming Ding, and Yun Fu

Application
- Multi-view Clustering

Extension: Incomplete Scenario

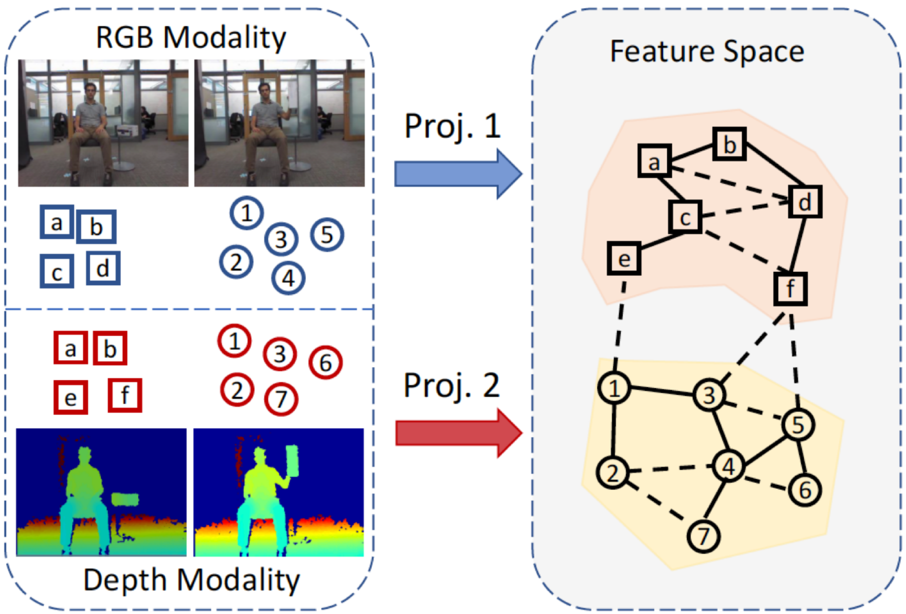
When the data from one modality/more modalities are inaccessible because of sensor failure or other reasons, most traditional MVC methods would inevitably degenerate or even fail.



Application
- Multi-view Clustering

Extension: Incomplete Scenario
Incomplete Multi-Modal Visual Data Grouping

Motivation:



Objective function:

$$\min_{\substack{P_c,\, \hat{P}^{(1)},\, \hat{P}^{(2)} \\ U^{(1)},\, U^{(2)},\, A}} \left\| \begin{bmatrix} X_c^{(1)} \\ \hat{X}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_F^2 + \left\| \begin{bmatrix} X_c^{(2)} \\ \hat{X}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_F^2 + \mathcal{G}(P, A) + \mathcal{R}(U, A).$$

$$\text{s.t. } \forall i \; A_i^T \mathbf{1} = 1, \; A_i \succeq 0.$$

$$\mathcal{G}(P, A) = \beta \operatorname{tr}(P^{\mathrm{T}} L_A P),$$

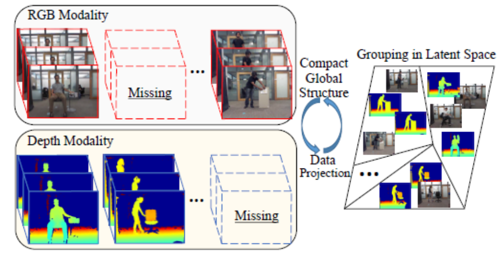
$$\mathcal{R}(U, A) = \lambda(\|U^{(1)}\|_F^2 + \|U^{(2)}\|_F^2) + \gamma \|A\|_F^2$$

Incomplete Multi-Modal Visual Data Grouping – IJCAI'16
Handong Zhao, Hongfu Liu and Yun Fu

Application
- Multi-view Clustering

Northeastern University
Smile lab
Synergetic Media Learning Lab

Extension: Incomplete Scenario
Incomplete Multi-Modal Visual Data Grouping



Objective function:

$$\min_{\substack{P_c, \hat{P}^{(1)}, \hat{P}^{(2)} \\ U^{(1)}, U^{(2)}, A}} \left\| \begin{bmatrix} \boxed{X_c^{(1)}} \\ \hat{X}^{(1)} \end{bmatrix} - \begin{bmatrix} \boxed{P_c} \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_{\mathrm{F}}^2 + \left\| \begin{bmatrix} \boxed{X_c^{(2)}} \\ \hat{X}^{(2)} \end{bmatrix} - \begin{bmatrix} \boxed{P_c} \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_{\mathrm{F}}^2$$

$$+ \mathcal{G}(P, A) + \boxed{\mathcal{R}(U, A)}$$

$$\text{s.t. } \forall i \; A_i^T \mathbf{1} = 1, \; A_i \succeq 0.$$

$$\mathcal{G}(P, A) = \beta \mathrm{tr}(P^{\mathrm{T}} L_A P)$$

$$\boxed{\mathcal{R}(U, A) = \lambda(\|U^{(1)}\|_{\mathrm{F}}^2 + \|U^{(2)}\|_{\mathrm{F}}^2) + \gamma \|A\|_{\mathrm{F}}^2}$$

- Use the shared data $X_c^{(v)}$ in each view to learn the common representation P_c in low-dimensional subspace space.
- Regularizers to prevent trivial solution. Here we choose a simple Frobenius norm. Its alternatives include ℓ_1 and others.

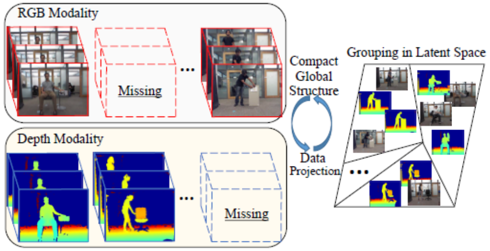
Incomplete Multi-Modal Visual Data Grouping – IJCAI'16
Handong Zhao, Hongfu Liu and Yun Fu

Application
- Multi-view Clustering

Northeastern University

Smile lab Synergetic Media Learning Lab

Extension: Incomplete Scenario
Incomplete Multi-Modal Visual Data Grouping



Objective function:

$$\min_{\substack{P_c, \hat{P}^{(1)}, \hat{P}^{(2)} \\ U^{(1)}, U^{(2)}, A}} \left\| \begin{bmatrix} X_c^{(1)} \\ \hat{X}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_{\mathrm{F}}^2 + \left\| \begin{bmatrix} X_c^{(2)} \\ \hat{X}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_{\mathrm{F}}^2$$

$$+ \boxed{\mathcal{G}(P, A)} + \mathcal{R}(U, A)$$

$$\text{s.t. } \forall i \; A_i^T \mathbf{1} = 1, \; A_i \succeq 0.$$

$$\boxed{\mathcal{G}(P, A) = \beta \mathrm{tr}(P^{\mathrm{T}} L_A P)}$$

$$\mathcal{R}(U, A) = \lambda(\|U^{(1)}\|_{\mathrm{F}}^2 + \|U^{(2)}\|_{\mathrm{F}}^2) + \gamma \|A\|_{\mathrm{F}}^2$$

- $P = [P_c; \hat{P}^{(1)}; \hat{P}^{(2)}]$
- Graph Laplacian term to preserve locality information, where L_A is the Laplacian matrix of similarity matrix A, which is learned on the latent representation P.

Incomplete Multi-Modal Visual Data Grouping – IJCAI'16
Handong Zhao, Hongfu Liu and Yun Fu

Outline

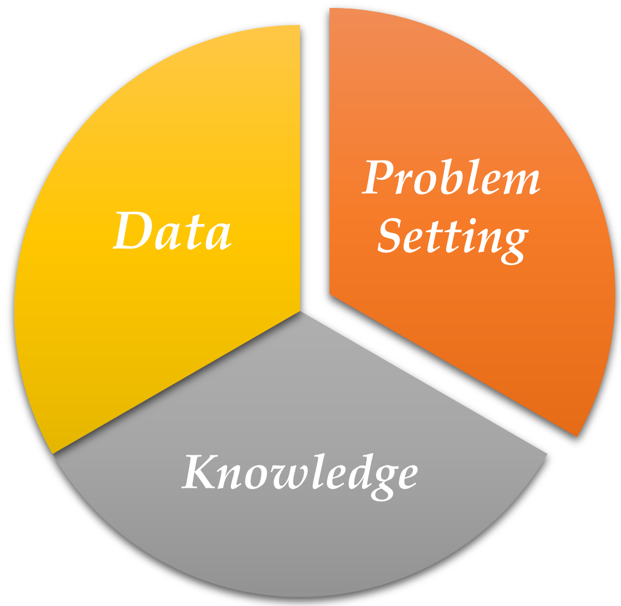
Northeastern University

Synergetic Media Learning Lab

- ❑ **Introduction & Background**
  - Multi-view Visual Data
  - Multi-view Learning Problems
  - Multi-view Learning Taxonomy
- ❑ **Multi-view Learning**
  - Projection and Embedding
  - Knowledge Fusion
  - Multi-view Clustering
  - Supervised Multi-view Learning → Zero-shot Learning
- ❑ **Domain Adaptation**
  - Transfer Learning → Domain Adaptation
  - Multi-Source Domain Adaptation & Domain Generalization
- ❑ **Conclusion**

Supervised Multi-View Learning & Domain Adaptation

- **Supervised Multi-view Learning [sample-wise correspondence]**



Training Stage: multiple labeled view data

Test Stage:

[**Setting 1**]: labeled views → unlabeled views

[**Setting 2**]: multiple unlabeled view data

Goal: **fuse** various knowledge from multiple views

Goal: **adapt** knowledge across different views

- **Domain Adaptation [class-wise correspondence]**

Training Stage: some source labeled domains & some target [un]labeled domains
Test Stage: source to predict target data

Supervised Multi-View Learning

Northeastern University

Smile lab Synergetic Media Learning Lab

Training Stage: $X_s = \{X_s^1, \ldots, X_s^v | y\}$

Test Stage: one view to recognize others $\{y_*\}$ or $\{X_s | y\}$ to recognize test data $\{X_t | y\}$

- Cross-pose Face Recognition
- Multi-modal Recognition

- **Sample-wise Correspondence**
- **Labels Information**

$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \mathcal{A}(f_i(X_i), f_j(X_j)) + \lambda \sum_{k=1}^{v} \boxed{\mathcal{R}(f_k(X_k))}$$

- ❑ Feature Learning
  → Subspace Learning, Deep Learning
- ◆ Alignment & fusion
  → joint & coordinated representation

- *Label information*
  - Supervised Graph
  - Regression loss

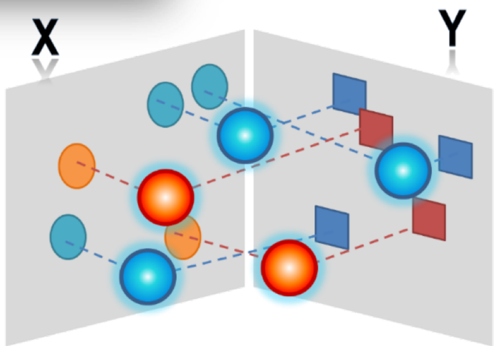
Zhengming Ding, Ming Shao, and Yun Fu. Robust Multi-view Representation: A Unified Perspective from Multi-view Learning to Domain Adaption. **IJCAI**, 2018

CCA → Supervised CCA

CCA

$$\max_{\mathbf{P},\mathbf{Q}} \quad \mathrm{trace}\big(\mathbf{P}^\top \mathbf{X}_T \mathbf{X}_S^\top \mathbf{Q}^\top\big), \quad \Rightarrow \quad ||\mathbf{X}_T^\top \mathbf{P} - \mathbf{X}_S^\top \mathbf{Q}||_F^2$$

$$\text{s.t.} \quad \mathbf{P}^\top \mathbf{X}_T \mathbf{X}_T^\top \mathbf{P} = \mathbf{I}, \quad \mathbf{Q}^\top \mathbf{X}_S \mathbf{X}_S^\top \mathbf{Q} = \mathbf{I}$$



Supervised CCA

$$\min_{\mathbf{w},\mathbf{P},\mathbf{Q}} C \sum_{i=1}^{n_L} \xi_{L,i} + \frac{1}{2}||\mathbf{w}||^2 + \frac{\gamma}{2}||\mathbf{X}_T^\top \mathbf{P} - \mathbf{X}_S^\top \mathbf{Q}||_F^2$$

$$\text{s.t.} \quad \mathbf{P}^\top \mathbf{X}_T \mathbf{X}_T^\top \mathbf{P} = \mathbf{I}, \quad \mathbf{Q}^\top \mathbf{X}_S \mathbf{X}_S^\top \mathbf{Q} = \mathbf{I},$$

pair-wise constraints

cross-entropy loss function

$$\xi_{L,i} = \log\Big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \tilde{\mathbf{x}}_{L,i} - \mathbf{w}_{y_{L,i}}^\top \tilde{\mathbf{x}}_{L,i})\Big)$$

Yan, Yuguang, et al. "Learning Discriminative Correlation Subspace for Heterogeneous Domain Adaptation." IJCAI, 2017.

Multi-view Discriminant Analysis

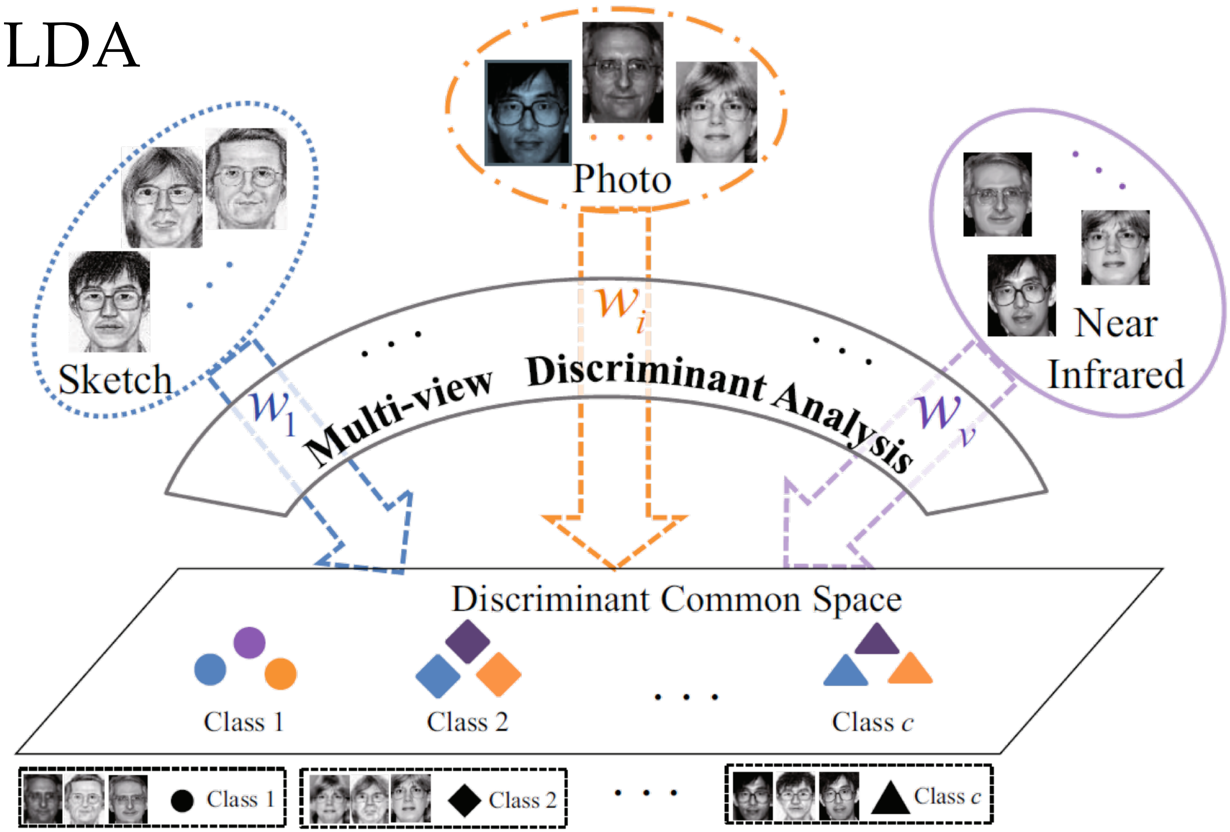
Northeastern University

LDA → Multi-view LDA

multiple projections

+ *Fisher loss*

$$\max_{\mathbf{w}_1,\cdots,\mathbf{w}_v} \frac{Tr\left(\mathbf{S}_B^y\right)}{Tr\left(\mathbf{S}_W^y\right)}$$



S_B^y: between-class scatter matrix S_W^y: within-class scatter matrix

Multi-view Discriminant Analysis – ECCV 2012 & IEEE TPAMI 2016

Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen

Supervised Multi-View Learning

Extension

view-consistent regularizer

$$(\mathbf{w}_1^*, \mathbf{w}_2^*, \ldots, \mathbf{w}_v^*)$$

$$= \arg\max_{\mathbf{w}_1,\ldots,\mathbf{w}_v} \frac{Tr(\mathbf{W}^T\mathbf{D}\mathbf{W})}{Tr(\mathbf{W}^T\mathbf{S}\mathbf{W}) + \lambda \boxed{\sum_{i,j=1}^{v} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2}}$$

$$= \arg \max_{\mathbf{w}_1,\ldots,\mathbf{w}_v} \frac{Tr(\mathbf{W}^T\mathbf{D}\mathbf{W})}{Tr(\mathbf{W}^T(\mathbf{S} + \lambda\mathbf{M})\mathbf{W})}$$

Projection & Data

$$\mathbf{w}_i = \mathbf{X}_i\boldsymbol{\beta}_i$$

$$\mathbf{X}_1 = \mathbf{R}\mathbf{X}_2$$

$$\mathbf{w}_1 = \mathbf{R}\mathbf{w}_2$$

$$\mathbf{X}_1\boldsymbol{\beta}_1 = \mathbf{R}\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{X}_1\boldsymbol{\beta}_2$$

Results on CUFSF and HFB Datasets

		CCA[4]**	CCA[4]+LDA	CDFE[19]	CSR[21]	PLS[6]	U-LDA[31]	GMA[28]	MvDA	MvDA-VC
CUFSF	Photo-Sketch	45.5%	45.0%	45.6%	50.2%	48.6%	46.8%	-	**53.4%**	**56.3%**
	Sketch-Photo	47.5%	50.6%	47.6%	49.0%	51.0%	53.4%	-	**55.5%**	**61.5%**
HFB	NIR-VIS	36.7%	40.0%	40.8%	26.7%	38.3%	39.1%	47.5%	**53.3%**	**59.2%**
	VIS-NIR	30.0%	40.0%	36.7%	32.5%	40.8%	40.0%	45.0%	**50.0%**	**59.2%**

Multi-view Discriminant Analysis. –ECCV 2012 & IEEE TPAMI 2016

Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen

Supervised Multi-View Learning

Deep Multi-View Learning

$$\min_{\mathbf{g}_c, \mathbf{f}_1, \cdots, \mathbf{f}_v} Tr\left(\frac{\mathbf{S}_W^y}{\mathbf{S}_B^y}\right)$$

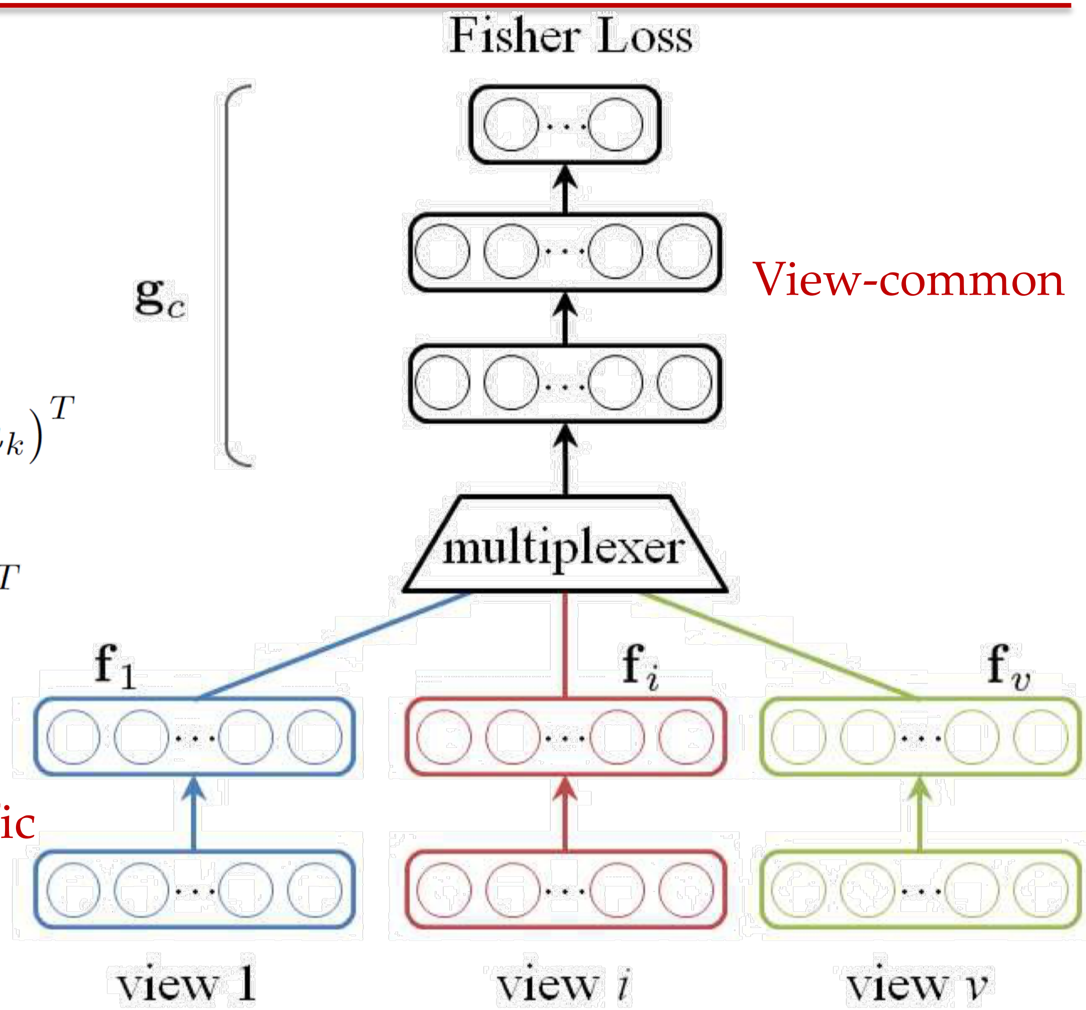
$$\mathbf{S}_W^y = \sum_{k=1}^{c}\sum_{i=1}^{v}\sum_{j=1}^{n_{ki}}\left(\mathbf{y}_{jk}^i - \boldsymbol{\mu}_k\right)\left(\mathbf{y}_{jk}^i - \boldsymbol{\mu}_k\right)^T$$

$$\mathbf{S}_B^y = \sum\nolimits_{k=1}^{c} n_k \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)^T$$

Fisher-like Loss

$$\mathbf{y}_j^i = \mathbf{g}_c\left(\mathbf{f}_i\left(\mathbf{x}_j^i\right)\right)$$

View-specific



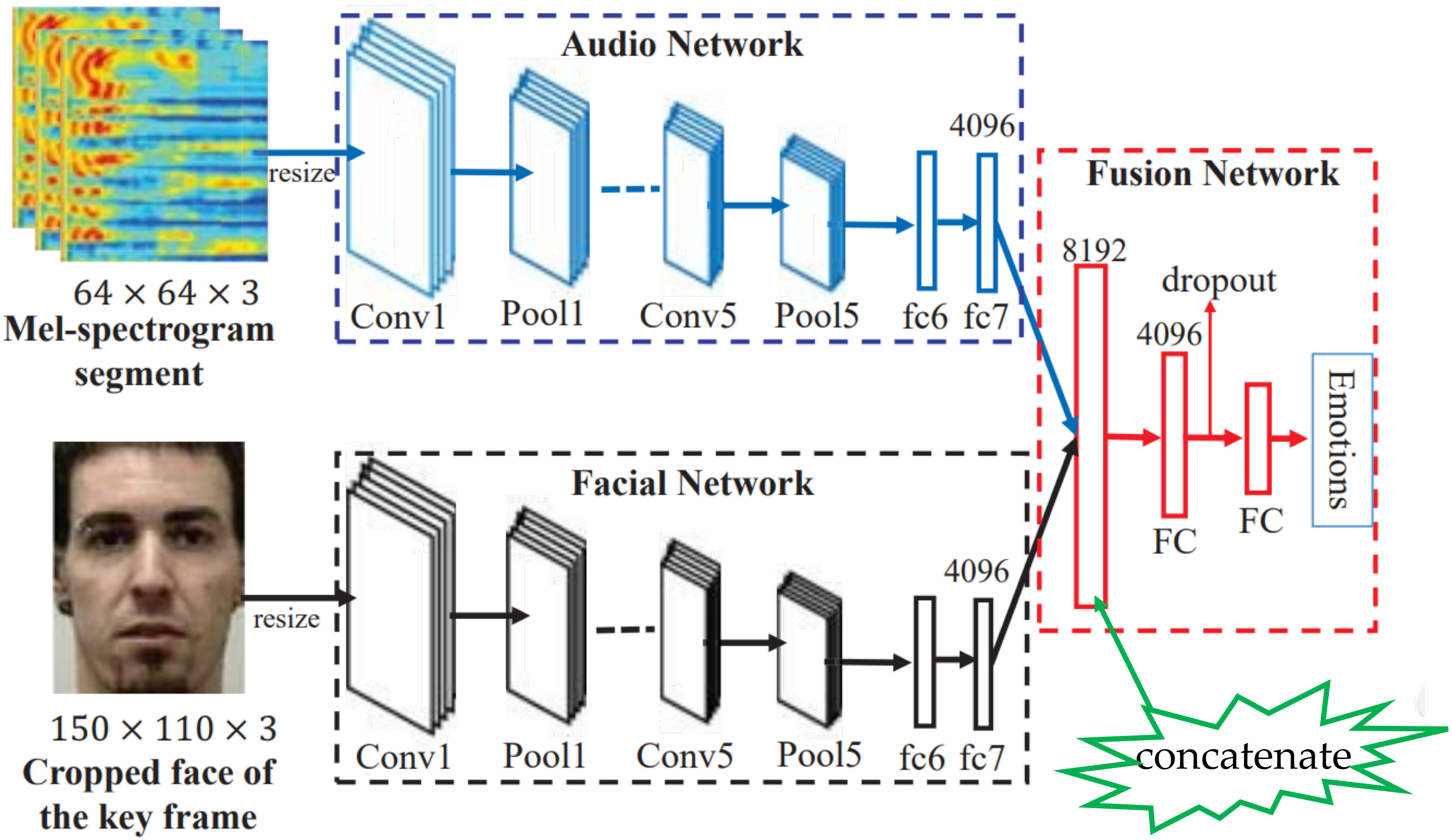
Multi-view Deep Network for Cross-view Classification– CVPR 2016

Meina Kan, Shiguang Shan and Xilin Chen

Supervised Multi-View Learning

Northeastern University

Smile lab Synergetic Media Learning Lab



Zhang, Shiqing, et al. "Multimodal deep convolutional neural network for audio-visual emotion recognition." Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, 2016.

Northeastern University

Special Case: Zero-Shot Learning

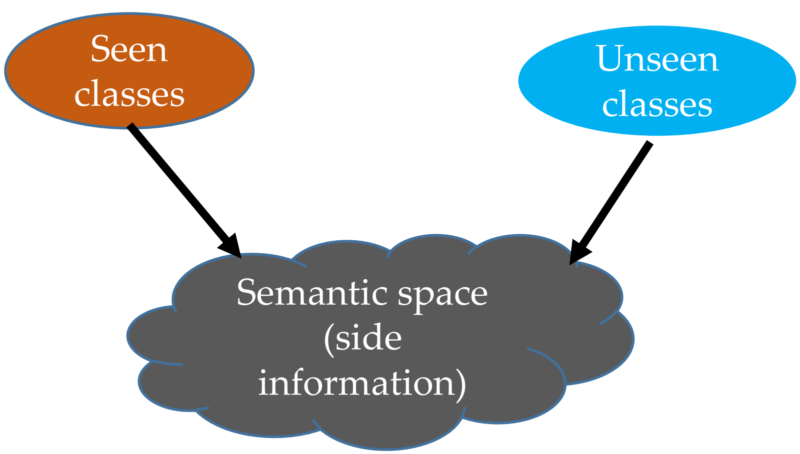
Supervised Multi-View Learning
[**Setting 1**]

Zero-Shot Learning

- **Seen classes:** labeled data available, training classes, 'side' information available
- **Unseen classes:** no training data (zero-shot), test classes

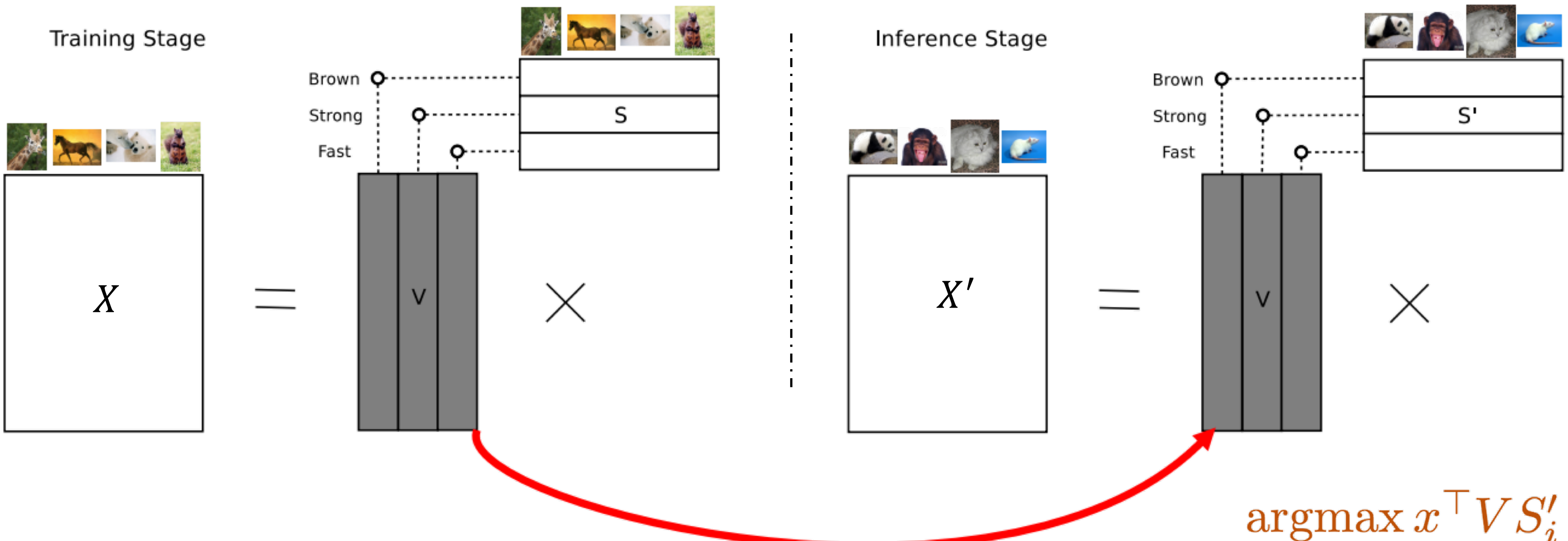
  - Zero-Shot Recognition
  - Zero-shot Annotation
  - Zero-shot Retrieval

find relationship between visual and semantic views



[Reed et.al. CVPR'16, ICML'16, NIPS'16]

Zero-Shot Learning



$$\operatorname*{argmax}_i x^\top V S'_i$$

Sample-wise correlation ship

$Y = \{-1, 1\}$

$$\operatorname*{minimise}_{V \in \mathbb{R}^{d \times a}} L\left(\boxed{X^\top V S}, Y\right) + \Omega\left(V; S, X\right)$$

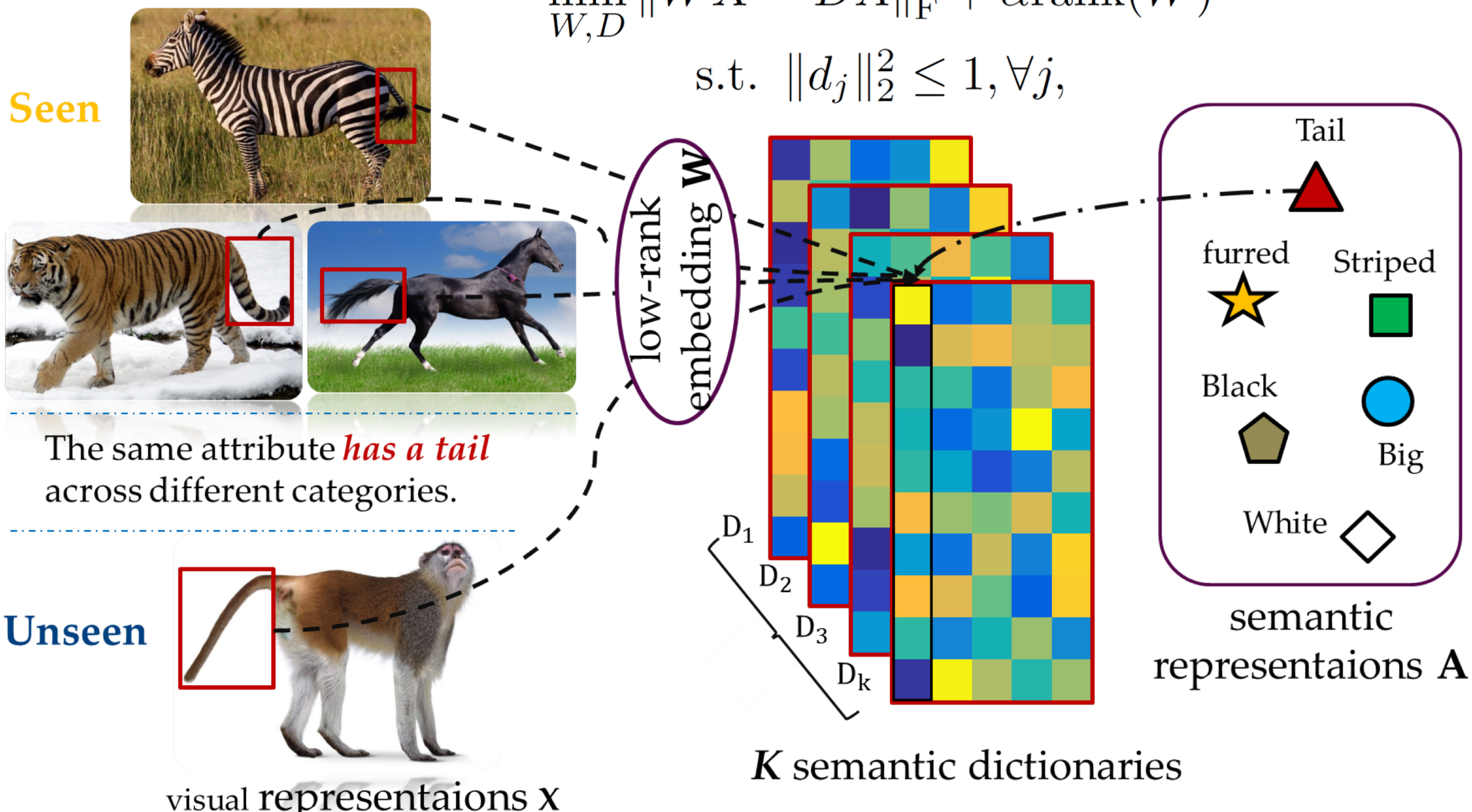
$$\Omega\left(V; S, X\right) = \gamma \left\|V S\right\|^2_{\text{Fro}} + \lambda \left\|X^\top V\right\|^2_{\text{Fro}} + \beta \left\|V\right\|^2_{\text{Fro}}$$

Romera-Paredes, Bernardino, and Philip Torr. "An embarrassingly simple approach to zero-shot learning.“ *International Conference on Machine Learning*. 2015.

Zero-Shot Learning

$$\min_{W,D} \|WX - DA\|_F^2 + \alpha \text{rank}(W)$$

$$\text{s.t. } \|d_j\|_2^2 \le 1, \forall j,$$



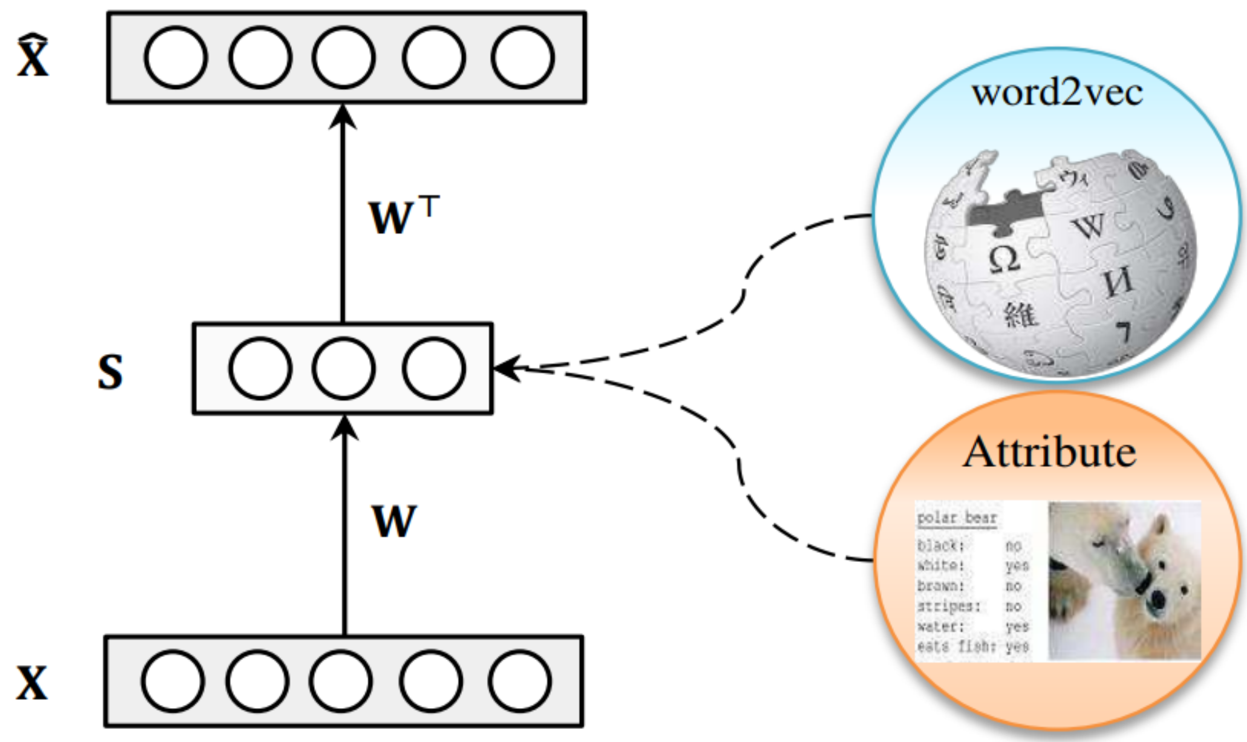
Zhengming Ding, Ming Shao and Yun Fu. Low-Rank Embedded Ensemble Semantic Dictionary for Zero-Shot Learning. **CVPR**, 2017

Zero-Shot Learning

The proposed semantic auto-encoder leverages the semantic side information such as attributes and word vector, while learning an encoder and a decoder

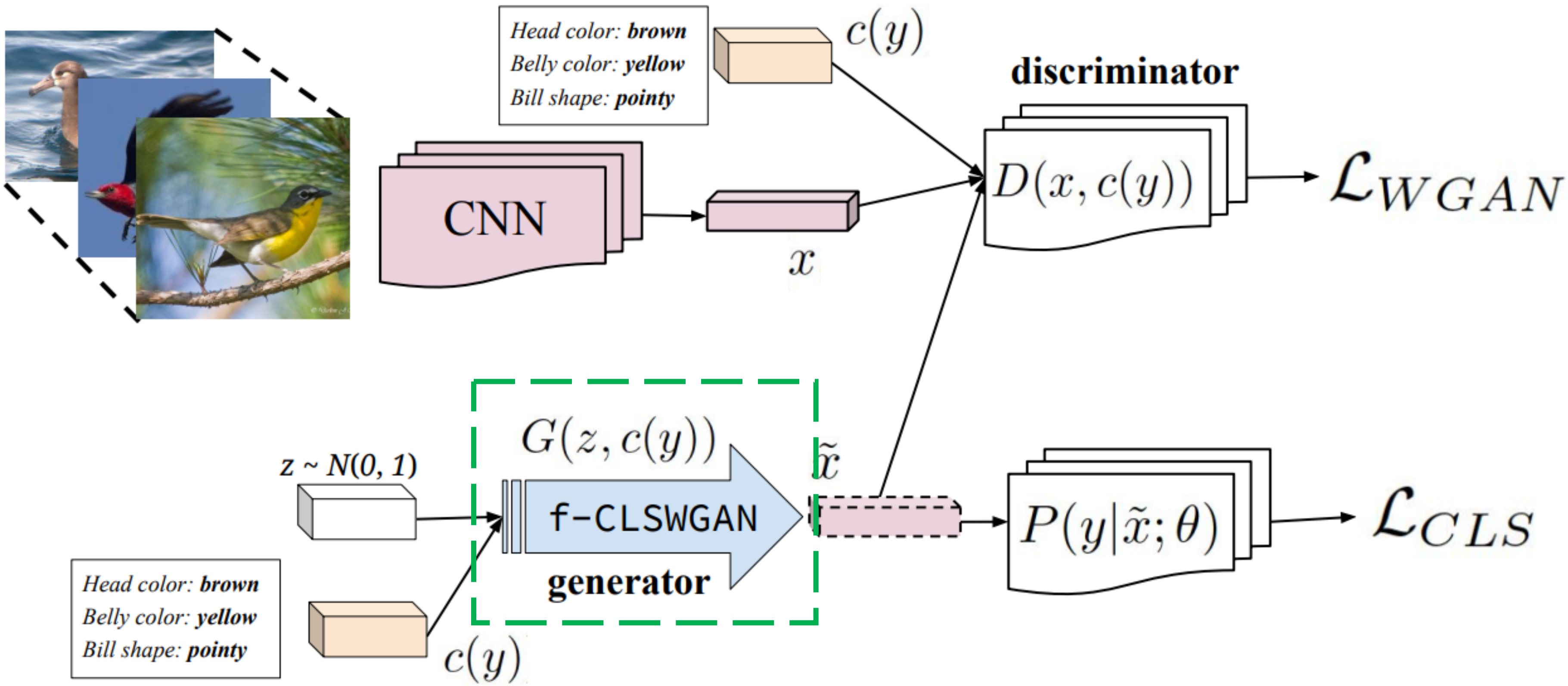
$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}^\top \mathbf{S}\|_F^2 \quad s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{S}$$

$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}^\top \mathbf{S}\|_F^2 + \lambda \|\mathbf{W}\mathbf{X} - \mathbf{S}\|_F^2$$



Kodirov, Elyor, Tao Xiang, and Shaogang Gong. "Semantic autoencoder for zero-shot learning." CVPR, 2017

Zero-Shot Learning



minimize the classification loss over the generated features and the Wasserstein distance with gradient penalty

Xian, Yongqin, et al. "Feature Generating Networks for Zero-Shot Learning." CVPR (2018).

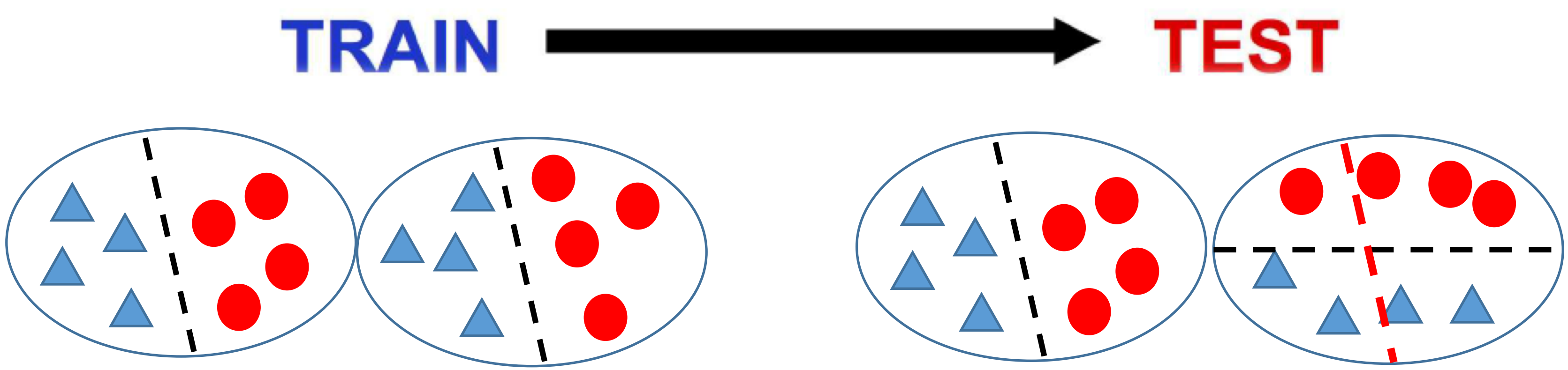
Outline

Northeastern University

Synergetic Media Learning Lab

- ❑ **Introduction & Background**
  - Multi-view Visual Data
  - Multi-view Learning Problems
  - Multi-view Learning Taxonomy
- ❑ **Multi-view Learning**
  - Projection and Embedding
  - Knowledge Fusion
  - Multi-view Clustering
  - Supervised Multi-view Learning → Zero-shot Learning
- ❑ **Domain Adaptation**
  - Transfer Learning → Domain Adaptation
  - Multi-Source Domain Adaptation & Domain Generalization
- ❑ **Conclusion**

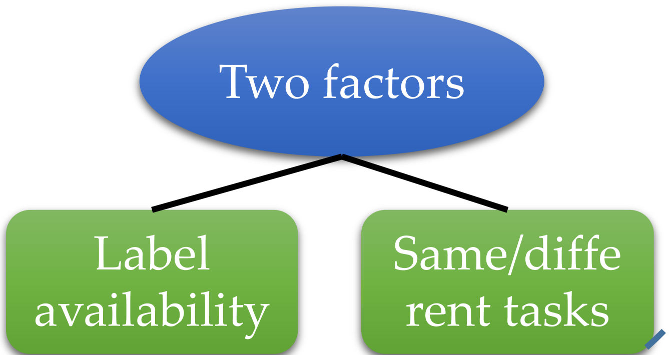
Examples

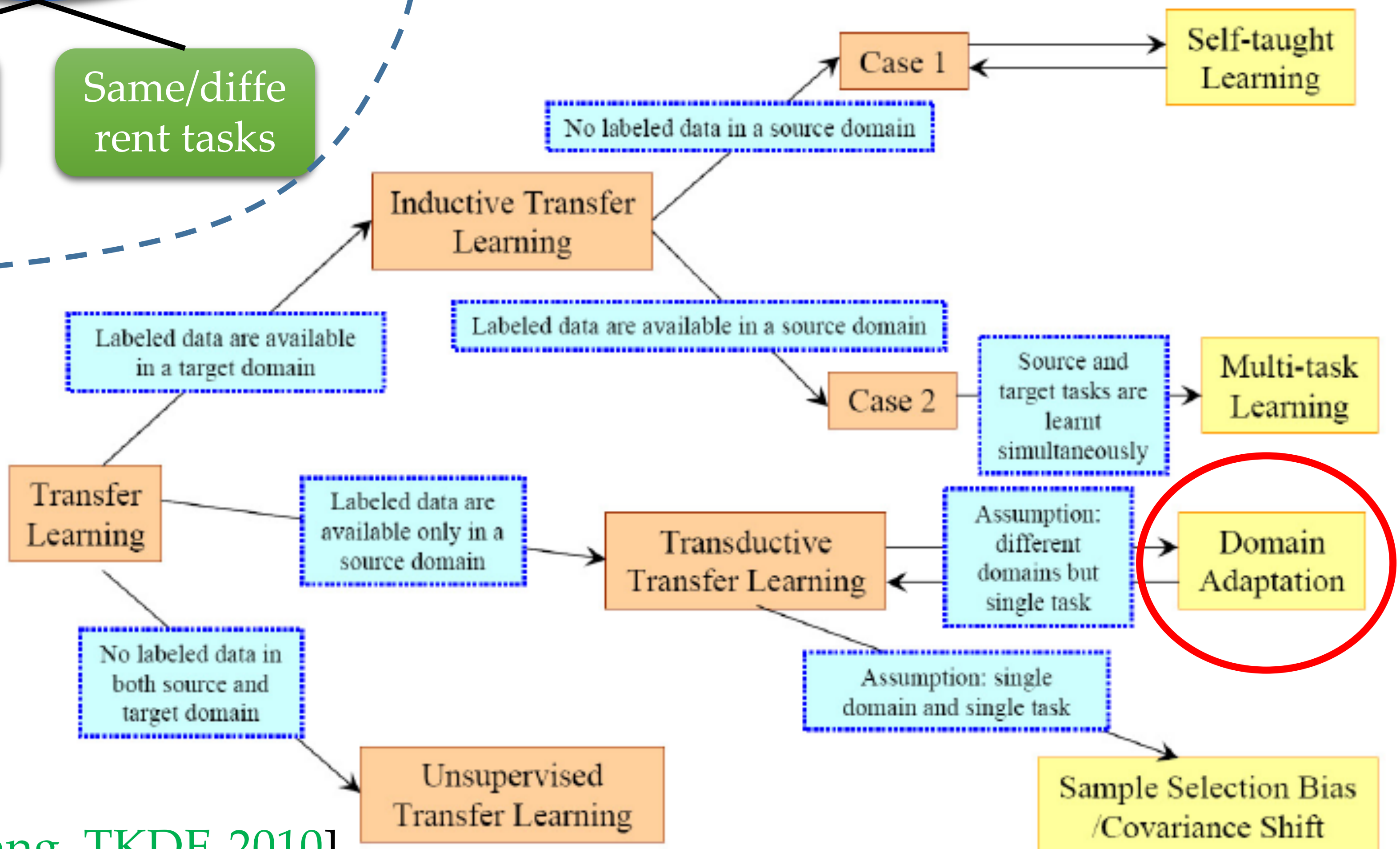


Training and test are from the same domain

Training and test are from different domains

Taxonomy





[Pan and Yang, TKDE-2010]

Domain Adaptation

- Source Views (labeled)

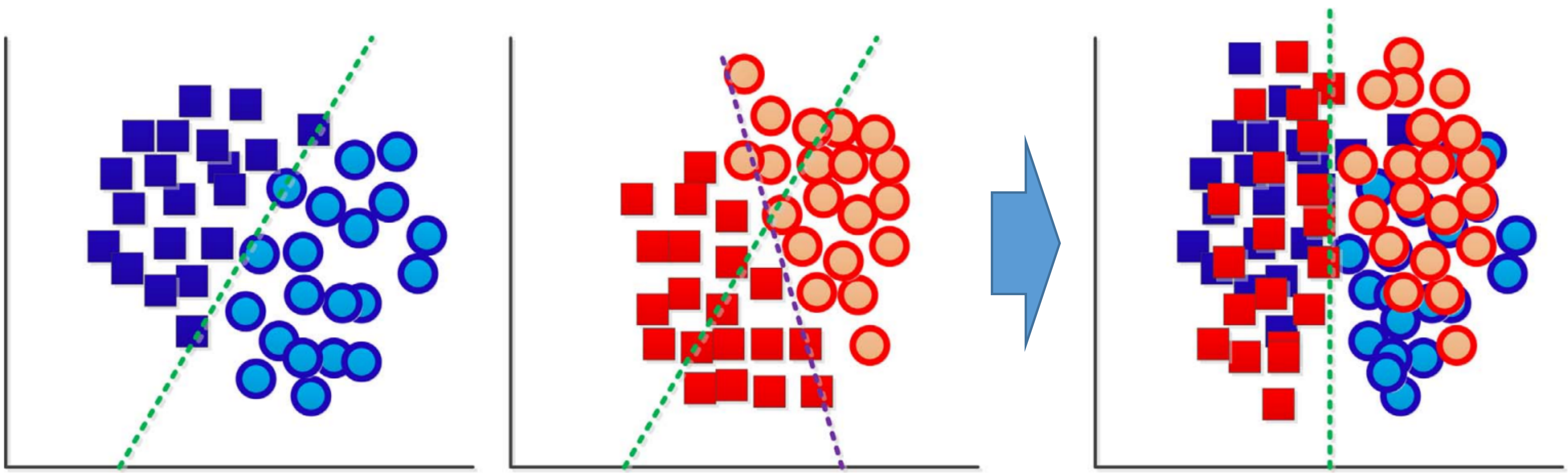
$$D_{s,M} = \{(x_{i,M}, y_{i,M}), i = 1,2 \cdots, N, \boxed{P_s(X,Y)}\}$$

mismatch

- Target Views ([un]labeled)

$$D_T = \{(x_i, ?), i = 1,2 \cdots, N, \boxed{P_T(X,Y)}\}$$

Performance degrades significantly!



Domain Adaptation Approaches

Northeastern University
Smile lab
Synergetic Media Learning Lab

Instance based approach | **Parameter based approach** | **Relational knowledge** | **Feature based approach**

❖ ***Instance:*** partial source data are reusable
- Instance selection approach: **TrAdaBoost**

❖ ***Parameter*:** individual models for related tasks should share some parameters or priors
- **Multi-task learning**

❖ ***Relational*:** no i.i.d. assumptions, transfer relationship among data between domains

❖ ***Feature*:** good representations to mitigate domain divergence
- **Subspace learning → Linear Projection**
- **Dictionary learning → New Representation**
- **Deep learning → Convolutional Neural Network**

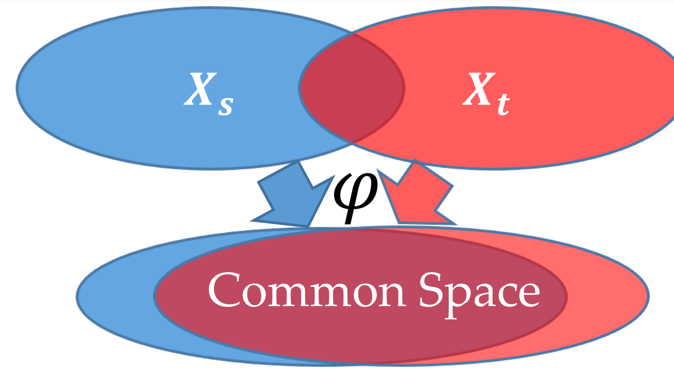
We focus on Representation + Domain Alignment

Representation Learning

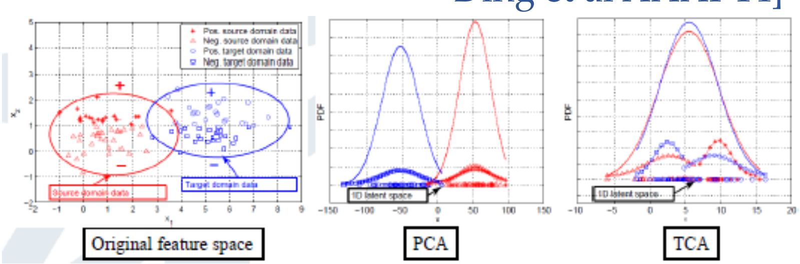
Northeastern University

Smile lab — Synergetic Media Learning Lab

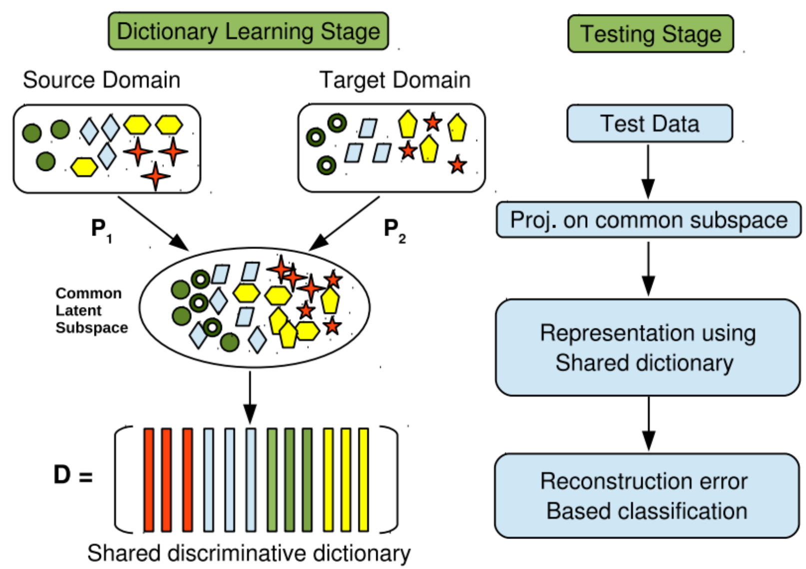
- Source and target have some overlapping features
- Have support in either source or target domain
- Find the transform φ



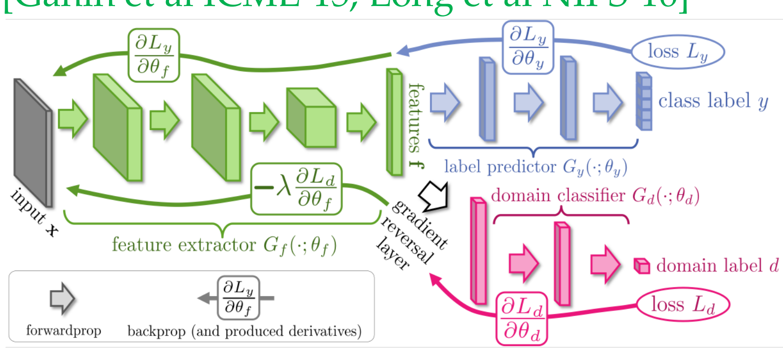
Projection learning [Pan et al AAAI-08, Saenko et al ECCV-2010, Ding et al AAAI-14]



Dictionary learning [Shekhar et al CVPR-13, Nguyen et al TIP-15]



Deep learning [Ganin et al ICML-15, Long et al NIPS-16]



Domain Alignment

Northeastern University

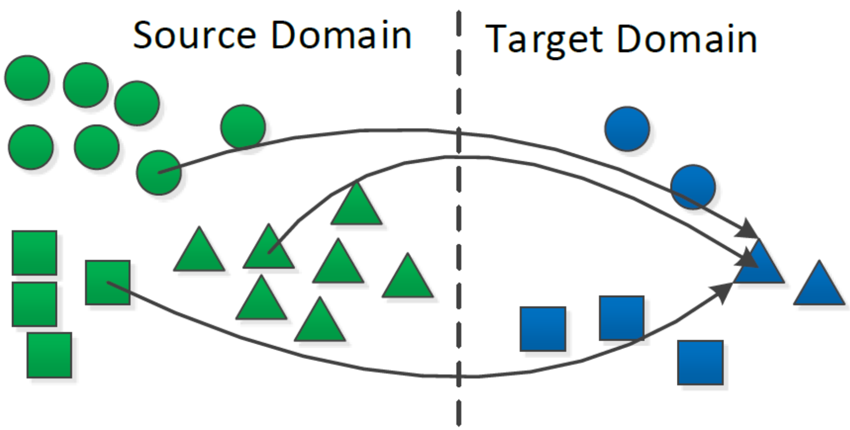
Smile lab Synergetic Media Learning Lab

➢ **Marginal distribution** $P_s(X_s), P_t(X_t)$

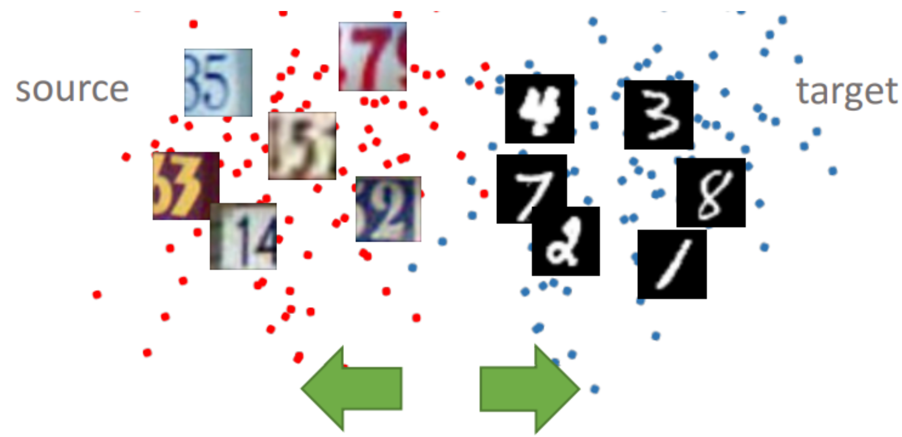
➢ **Conditional distribution** $Q_s(y_s|X_s), Q_t(y_t|X_t)$

- **MMD loss** $$MMD^2(s,t) = \sup_{\|\phi\|_{\mathcal{H}} \leq 1} \left\| E_{x^s \sim s}[\phi(x^s)] - E_{x^t \sim s}[\phi(x^t)] \right\|_{\mathcal{H}}^2$$

- **Reconstruction loss**



- **Adversarial loss**

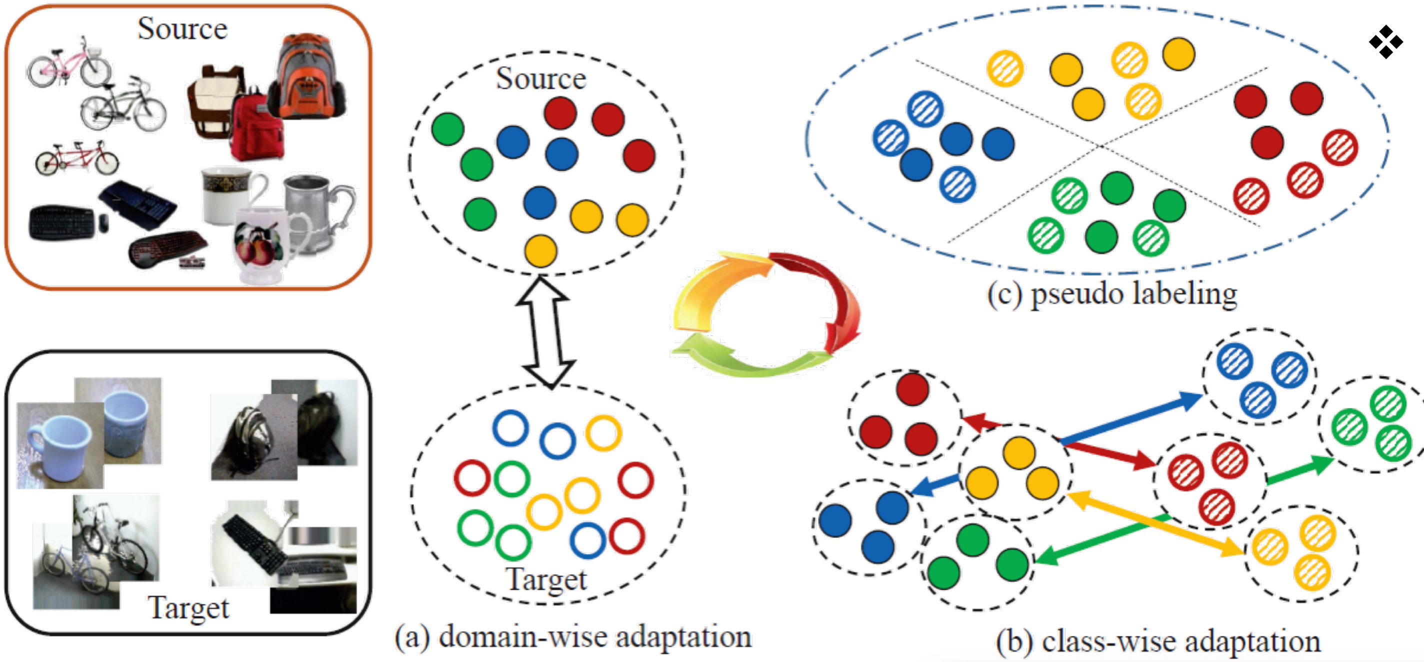


Source ? Target?

- **Parameters Sharing**, *e.g., weights/dictionary*

Conditional MMD loss +Projection learning

Northeastern University



- ❖ **Class-wise MMD**
  - Conditional distribution alignment
  - Iteratively optimize target labels

- ❖ **Domain-wise MMD**
  only aligns marginal distribution

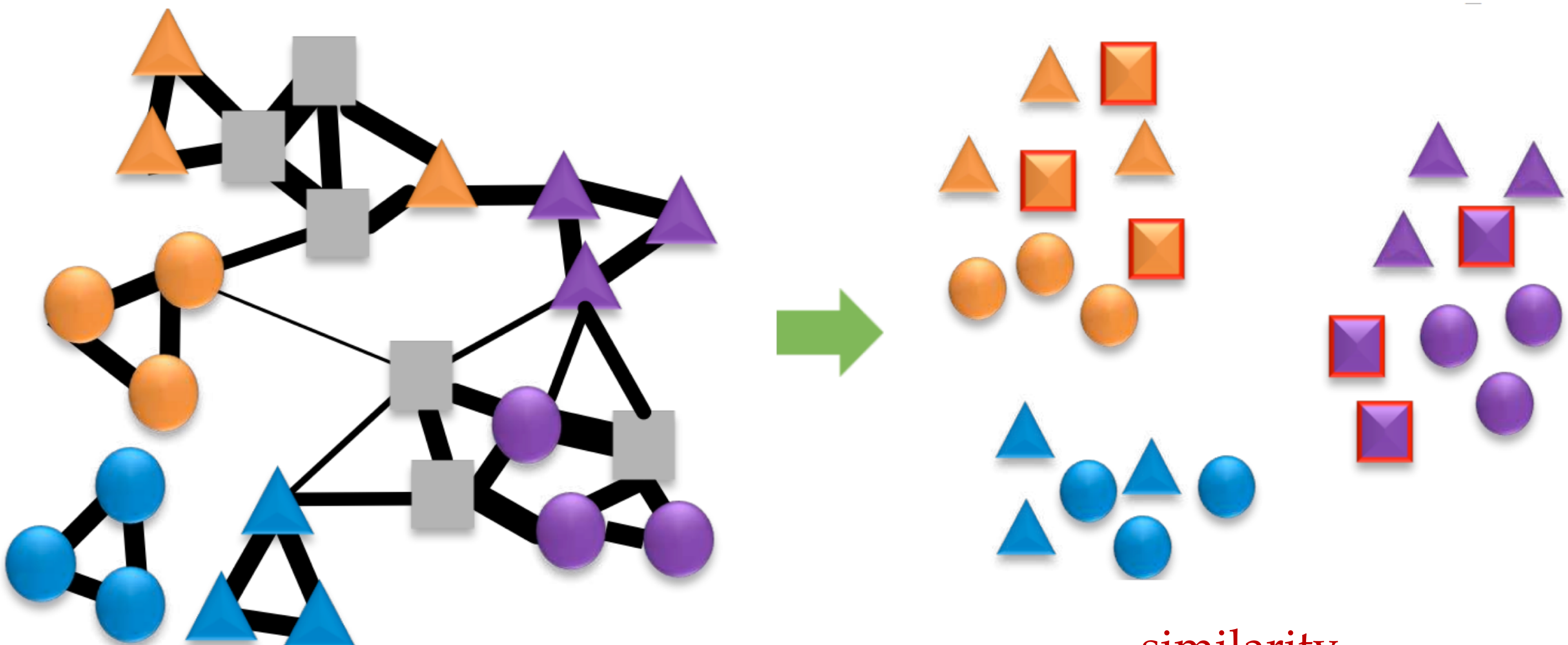
$$\left\| \frac{1}{n_s^{(c)}} \sum_{\mathbf{x}_i \in \mathcal{D}_s^{(c)}} \mathbf{A}^{\mathrm{T}} \mathbf{x}_i - \frac{1}{n_t^{(c)}} \sum_{\mathbf{x}_j \in \mathcal{D}_t^{(c)}} \mathbf{A}^{\mathrm{T}} \mathbf{x}_j \right\|^2$$

Long et al. "Transfer feature learning with joint distribution adaptation." *ICCV*. 2013.
Ding et al. "Robust transfer metric learning for image classification." *IEEE Transactions on Image Processing* 26.2 (2017): 660-670.

Modified MMD loss +Projection learning

Northeastern University

- Exploiting label and latent-domain information within and across domains

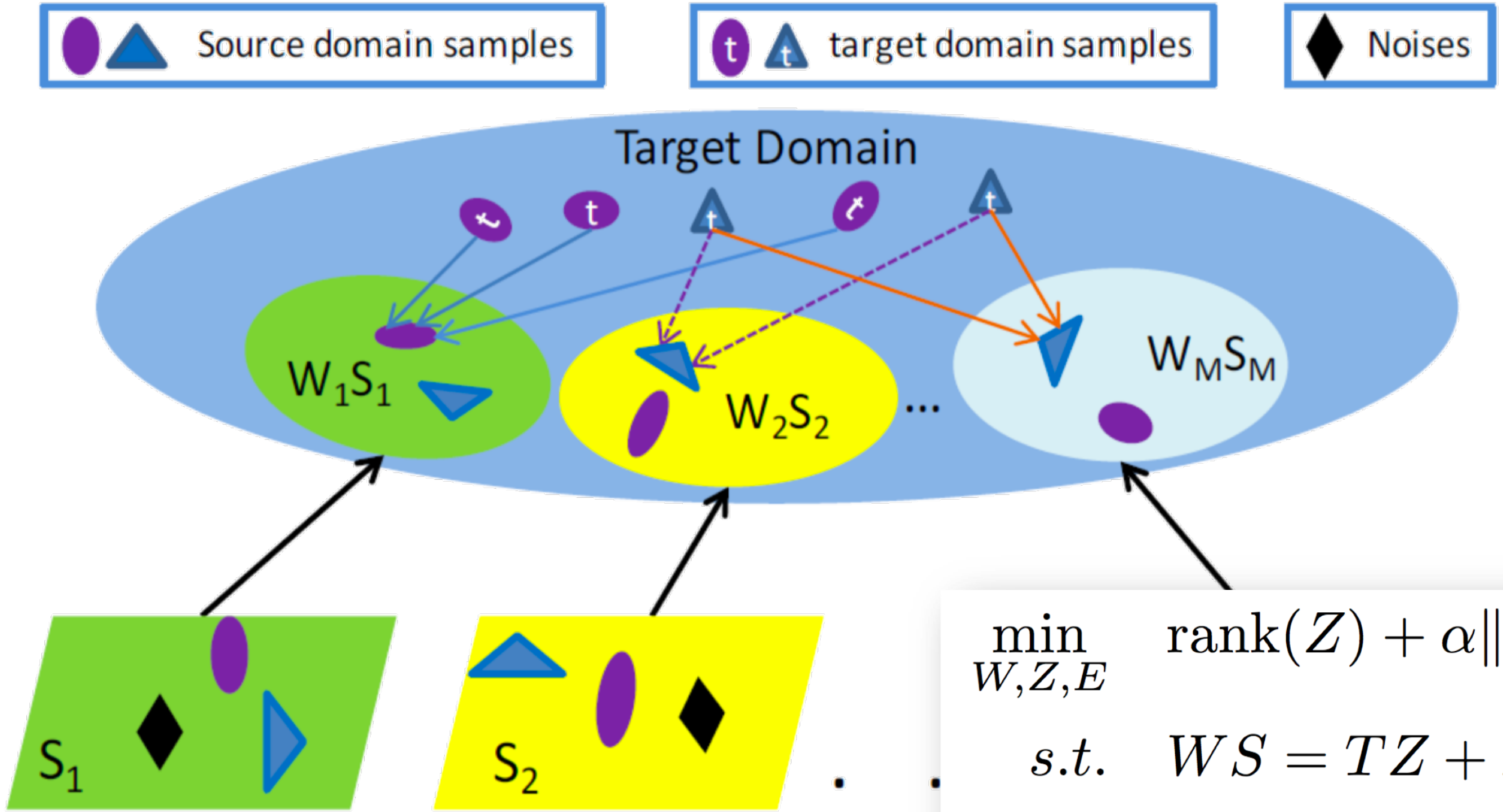


similarity

$$\mathcal{M}_{\phi,d}\left(\mathcal{P}_S(\mathbf{X}_S|\mathbf{y}_S),\mathcal{P}_T(\mathbf{X}_T|\mathbf{y}_T)\right)=\sum_{i,j}\frac{m_{ij}^{ST}}{\sum_k m_{ki}^{SS}\sum_l m_{lj}^{TT}}\left\|\hat{\phi}(\mathbf{x}_i^S)-\hat{\phi}(\mathbf{x}_j^T)\right\|^2$$

Ming Harry Hsu, Tzu, et al. "Unsupervised domain adaptation with imbalanced cross-domain data." *ICCV*. 2015.

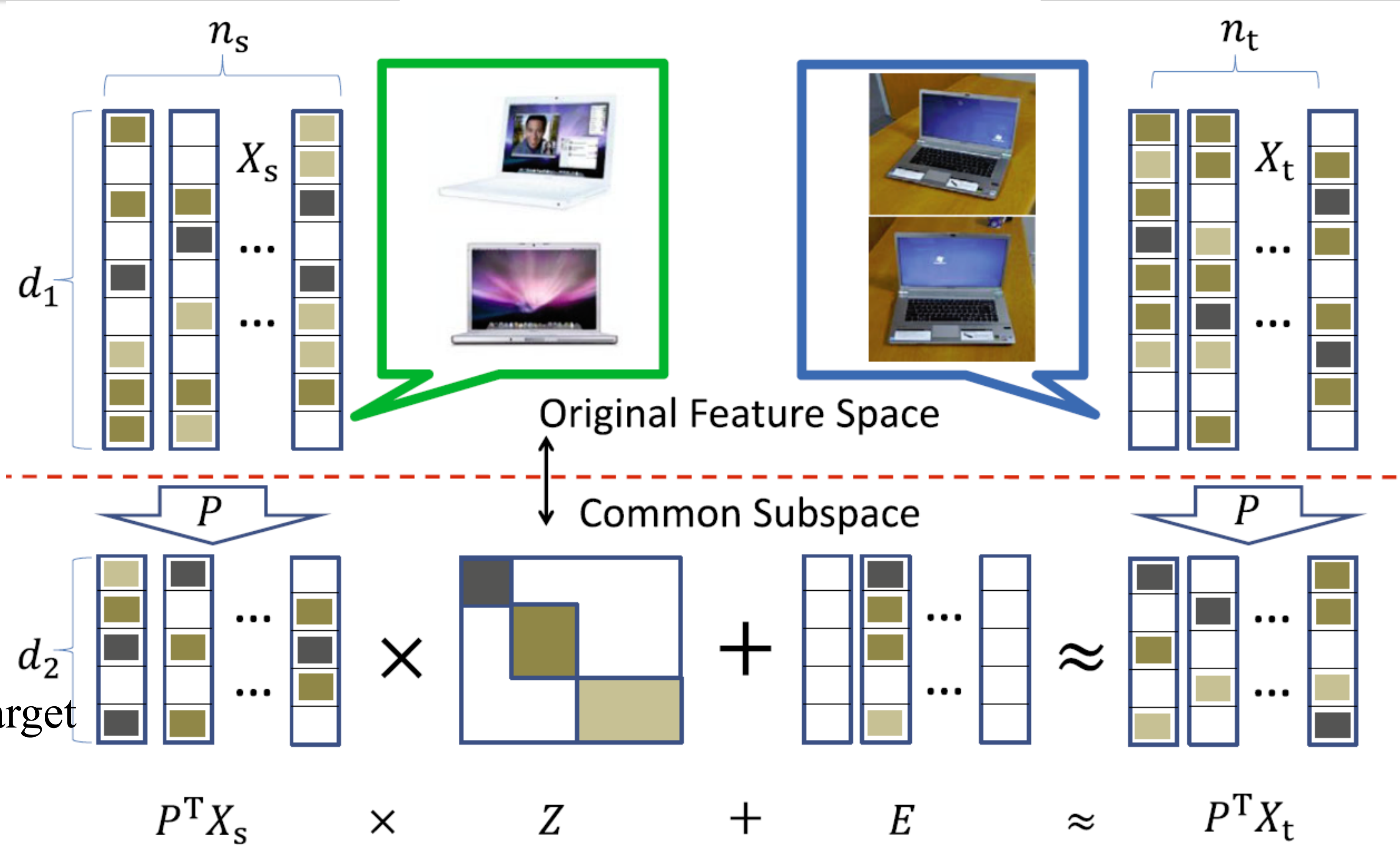
Reconstruction loss +Projection learning



$$\min_{W,Z,E} \quad \text{rank}(Z) + \alpha \|E\|_{2,1}$$
$$s.t. \quad WS = TZ + E,$$
$$WW^\top = I,$$

Jhuo, I-Hong, et al. "Robust visual domain adaptation with low-rank reconstruction." *CVPR*, 2012.

Reconstruction loss
+Projection learning



Low-Rank Transfer Subspace Learning – ICDM 2012, IJCV 2014

Ming Shao and Yun Fu

Dictionary Learning + MMD loss

Northeastern University

Smile lab Synergetic Media Learning Lab

Common Dictionary + Graph Regularizer [1]

$$\min_{\mathbf{B},\mathbf{S}} \|\mathbf{X}-\mathbf{B}\mathbf{S}\|_F^2 + \gamma \mathrm{tr}\left(\mathbf{S}\mathbf{L}\mathbf{S}^{\mathrm{T}}\right) + \lambda \sum_{i=1}^{n} |\mathbf{s}_i|$$

$$s.t. \quad \|\mathbf{b}_i\|^2 \leq c, i = 1, \ldots, k$$

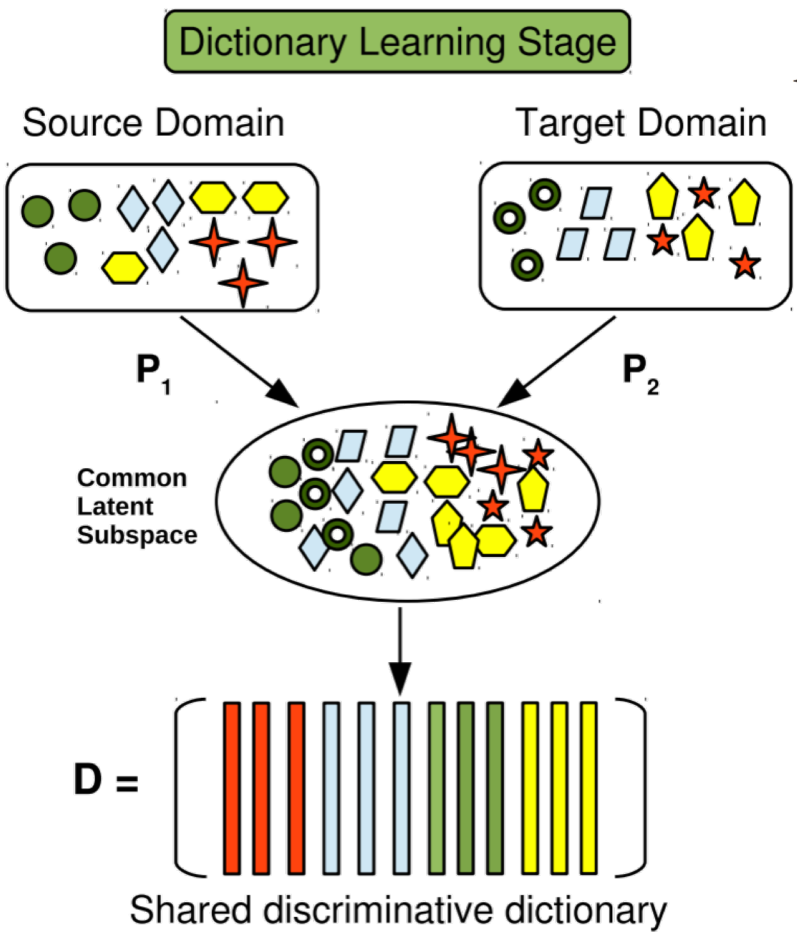
MMD loss

$$\left\| \frac{1}{n_l} \sum_{i=1}^{n_l} \mathbf{s}_i - \frac{1}{n_u} \sum_{j=n_l+1}^{n_l+n_u} \mathbf{s}_j \right\|^2 = \sum_{i,j=1}^{n} \mathbf{s}_i^{\mathrm{T}} \mathbf{s}_j M_{ij} = \mathrm{tr}\left(\mathbf{S}\mathbf{M}\mathbf{S}^{\mathrm{T}}\right)$$

[1]. Gao et al. Local features are not lonely – Laplacian sparse coding for image classification, CVPR, 2010

[2]. Long et al. Transfer sparse coding for robust image representation. *CVPR*. 2013.

Common Dictionary +Projection learning



$$\{\mathbf{D}^*, \tilde{\mathbf{P}}^*, \tilde{\mathbf{X}}^*\} = \underset{\mathbf{D},\tilde{\mathbf{P}},\tilde{\mathbf{X}}}{\arg\min}\ \mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}) + \lambda \mathcal{C}_2(\tilde{\mathbf{P}})$$

$$\text{s.t. } \mathbf{P_i P_i^T} = \mathbf{I},\ i = 1, 2 \text{ and } \|\tilde{\mathbf{x}}_j\|_0 \leq T_0, \forall j$$

- **Common dictionary**

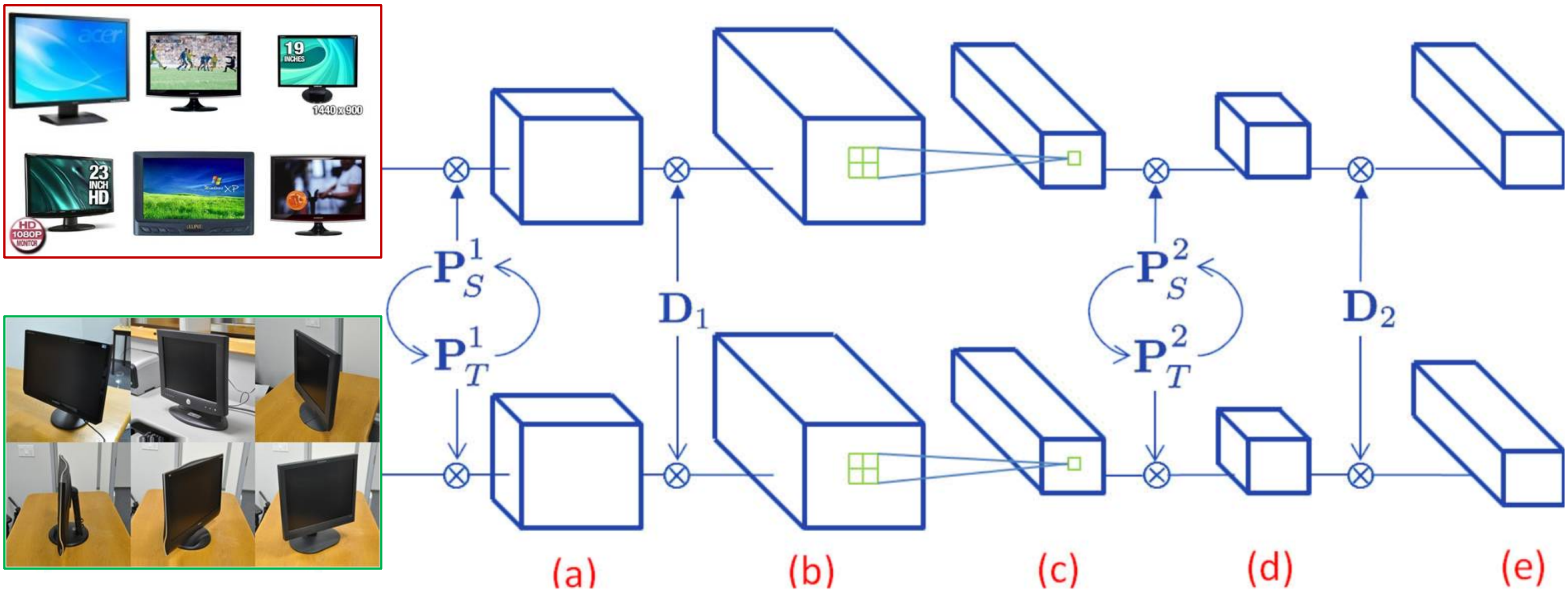
$$\mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}) = \|\mathbf{P_1 Y_1} - \mathbf{D X_1}\|_F^2 + \|\mathbf{P_2 Y_2} - \mathbf{D X_2}\|_F^2$$

- **PCA Reconstruction**

$$\mathcal{C}_2(\tilde{\mathbf{P}}) = \|\mathbf{Y_1} - \mathbf{P_1^T P_1 Y_1}\|_F^2 + \|\mathbf{Y_2} - \mathbf{P_2^T P_2 Y_2}\|_F^2.$$

Shekhar, Sumit, et al. "Generalized domain-adaptive dictionaries." CVPR, 2013

Common Dictionaries +Projection learning



$$\mathcal{L}(\mathbf{Y}_S, \mathbf{P}_S, \mathbf{D}, \mathbf{X}_S, \alpha, \beta) + \lambda \mathcal{L}(\mathbf{Y}_T, \mathbf{P}_T, \mathbf{D}, \mathbf{X}_T, \alpha, \beta)$$
$$\text{s.t. } \mathbf{P}_S \mathbf{P}_S^T = \mathbf{P}_T \mathbf{P}_T^T = \mathbf{I}, \quad \|\mathbf{d}_i\|_2 = 1, \quad \forall i \in [1, K],$$

$$\boxed{\begin{aligned}&\mathcal{L}(\mathbf{Y}, \mathbf{P}, \mathbf{D}, \mathbf{X}, \alpha, \beta) \\ &\qquad = \|\mathbf{PY} - \mathbf{DX}\|_F^2 + \alpha \|\mathbf{Y} - \mathbf{P}^T \mathbf{PY}\|_F^2 + \beta \|\mathbf{X}\|_1\end{aligned}}$$

[1] Nguyen et al. DASH-N: Joint hierarchical domain adaptation and feature learning. *IEEE TIP* 2015

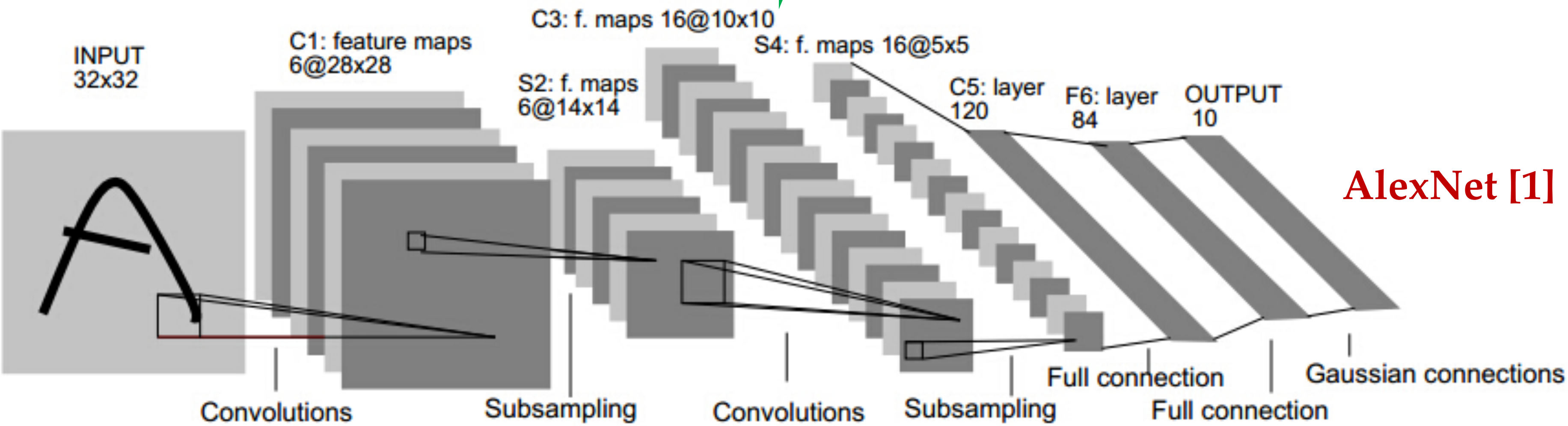
[2]. Ding et al. Deep Low-rank Coding for Transfer Learning, IJCAI 2015

Deep Learning

Northeastern University

Synergetic Media Learning Lab

Learner: $f : x \rightarrow y$ ⇨ Conduct knowledge transfer



AlexNet [1]

Pre-trained model on large-scale datasets, e.g., ImageNet

Top layers will be more task-specific [2] ⇨ **Align different domains**

[1]. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*. 2012.

[2]. Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *NIPS*, 2014

Unified Deep Domain Adaptation

Northeastern University
Smile lab
Synergetic Media Learning Lab

- *Two-Step Domain Adaptation*

Deep Features + Shallow DA methods

$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \mathcal{A}(f_i(X_i), f_j(X_j)) + \lambda \sum_{k=1}^{v} \mathcal{R}(f_k(X_k))$$

- *Deep Domain Adaptation*

Feature Learning + Domain Alignment

- MMD loss
- Adversarial loss
- Parameters Sharing

Labeled source domain

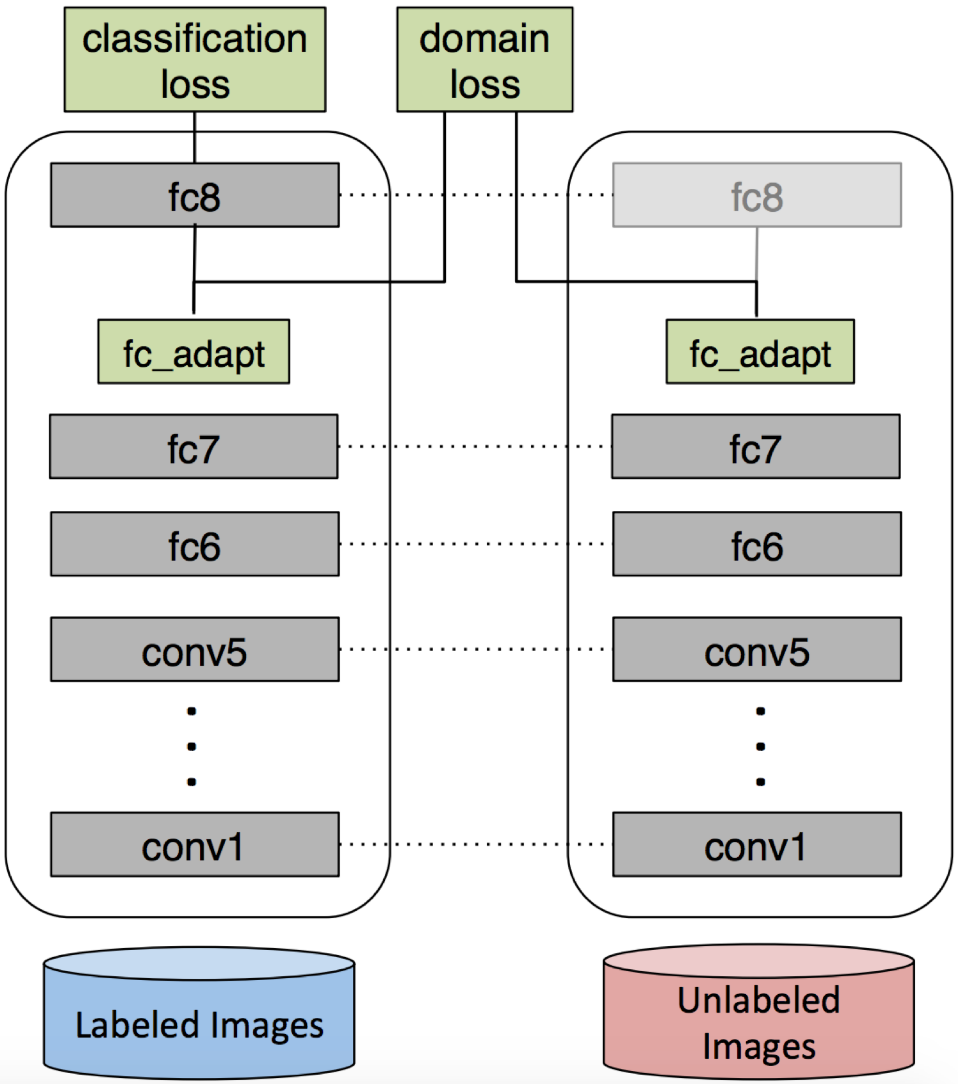
$L_c(X_s, y_s)$ +

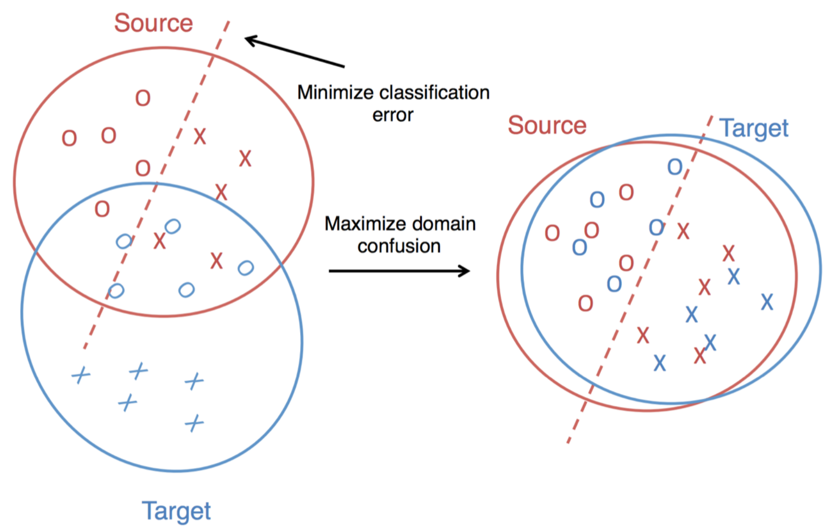
Soft labels of target domain

Zhengming Ding, Ming Shao, and Yun Fu. Robust Multi-view Representation: A Unified Perspective from Multi-view Learning to Domain Adaption. **IJCAI**, 2018

Deep Learning+ MMD Loss

Northeastern University



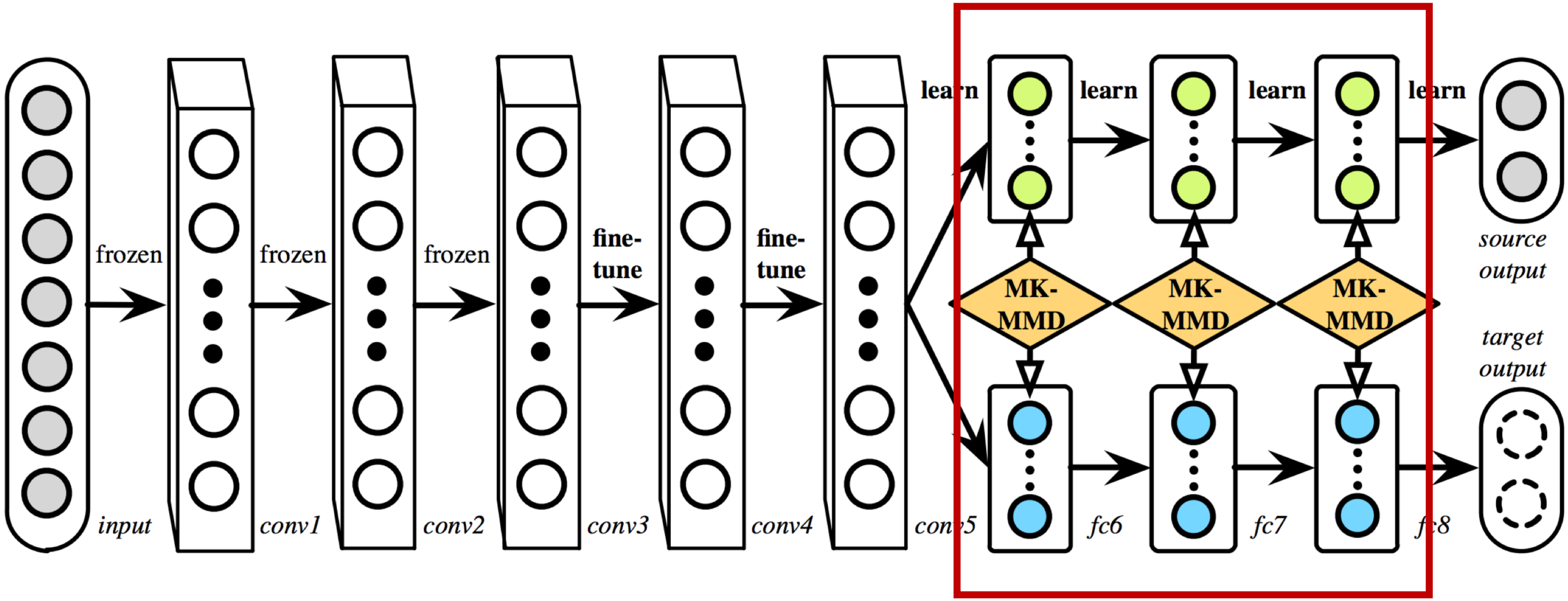


$$\mathcal{L} = \mathcal{L}_C(X_L, y) + \lambda \mathrm{MMD}^2(X_S, X_T)$$

Tzeng, Eric, et al. "Deep domain confusion: Maximizing for domain invariance." *arXiv preprint arXiv:1412.3474* (2014).

Deep Learning+ MK-MMD Loss

Northeastern University



multi-layer adaptation

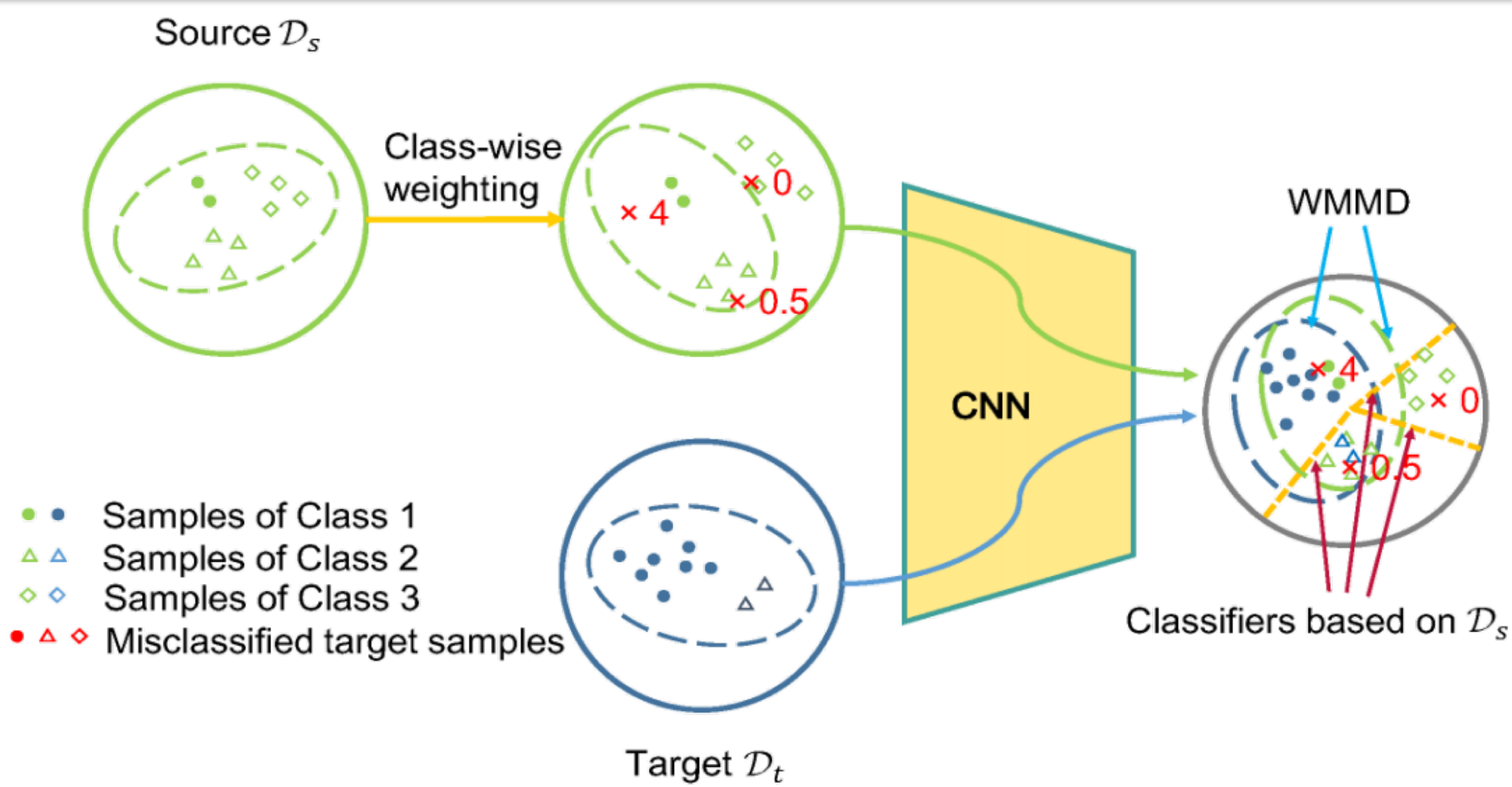
$$\min_{\theta \in \Theta} \max_{k \in \mathcal{K}} \frac{1}{n_a} \sum_{i=1}^{n_a} J\left(\theta\left(\mathbf{x}_i^a\right), y_i^a\right) + \lambda \sum_{\ell=l_1}^{l_2} d_k^2\left(\mathcal{D}_s^\ell, \mathcal{D}_t^\ell\right) \sigma_k^{-2}$$

Long, Mingsheng, et al. “Learning transferable features with deep adaptation networks.” *ICML, 2015*.

Deep Learning+ Weighted MMD loss

Northeastern University

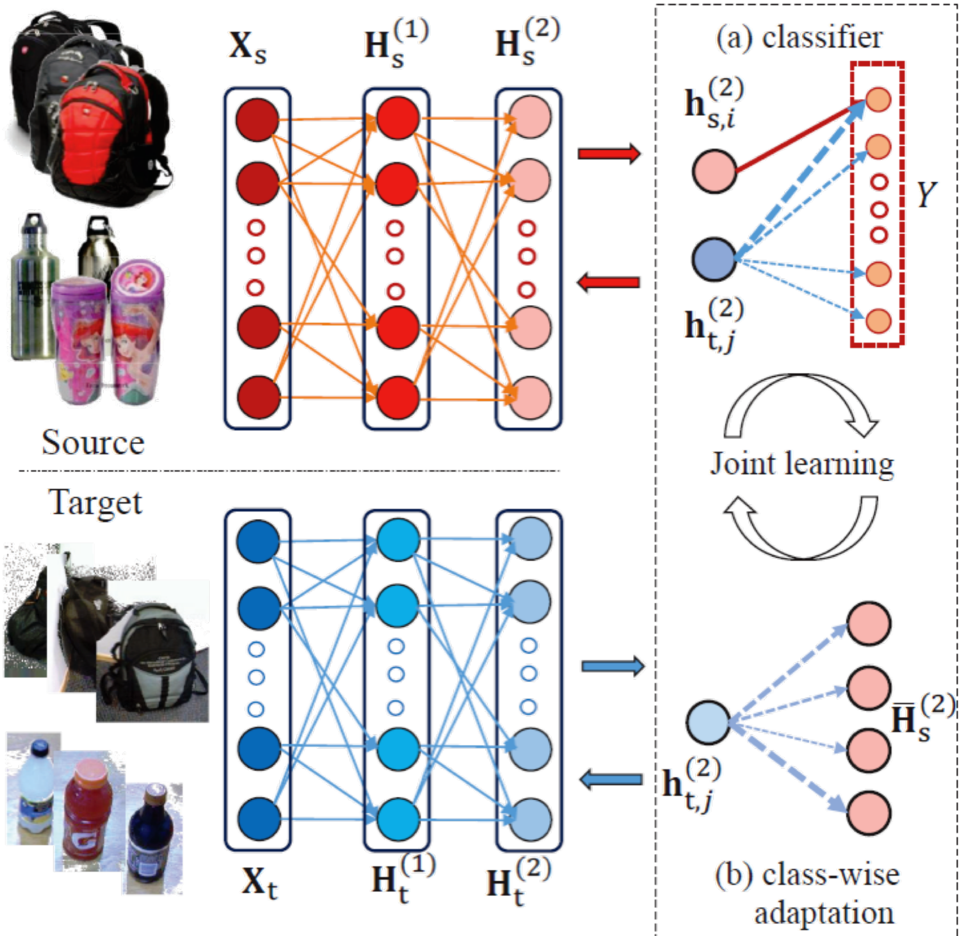
Smile lab
Synergetic Media Learning Lab



$$\mathrm{MMD}^2(\mathcal{D}_s, \mathcal{D}_t) = \left\| \frac{1}{M} \sum_{i=1}^{M} \phi(\mathbf{x}_i^s) - \frac{1}{N} \sum_{j=1}^{N} \phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2 \Rightarrow \mathrm{MMD}_w^2(\mathcal{D}_s, \mathcal{D}_t) = \left\| \frac{1}{\sum_{i=1}^{M} \alpha_{y_i^s}} \sum_{i=1}^{M} \alpha_{y_i^s} \phi(\mathbf{x}_i^s) - \frac{1}{N} \sum_{j=1}^{N} \phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2$$

Yan, Hongliang, et al. "Mind the Class Weight Bias: Weighted Maximum Mean Discrepancy for Unsupervised Domain Adaptation." *CVPR*. 2017.

Deep Learning+ Probabilistic MMD loss



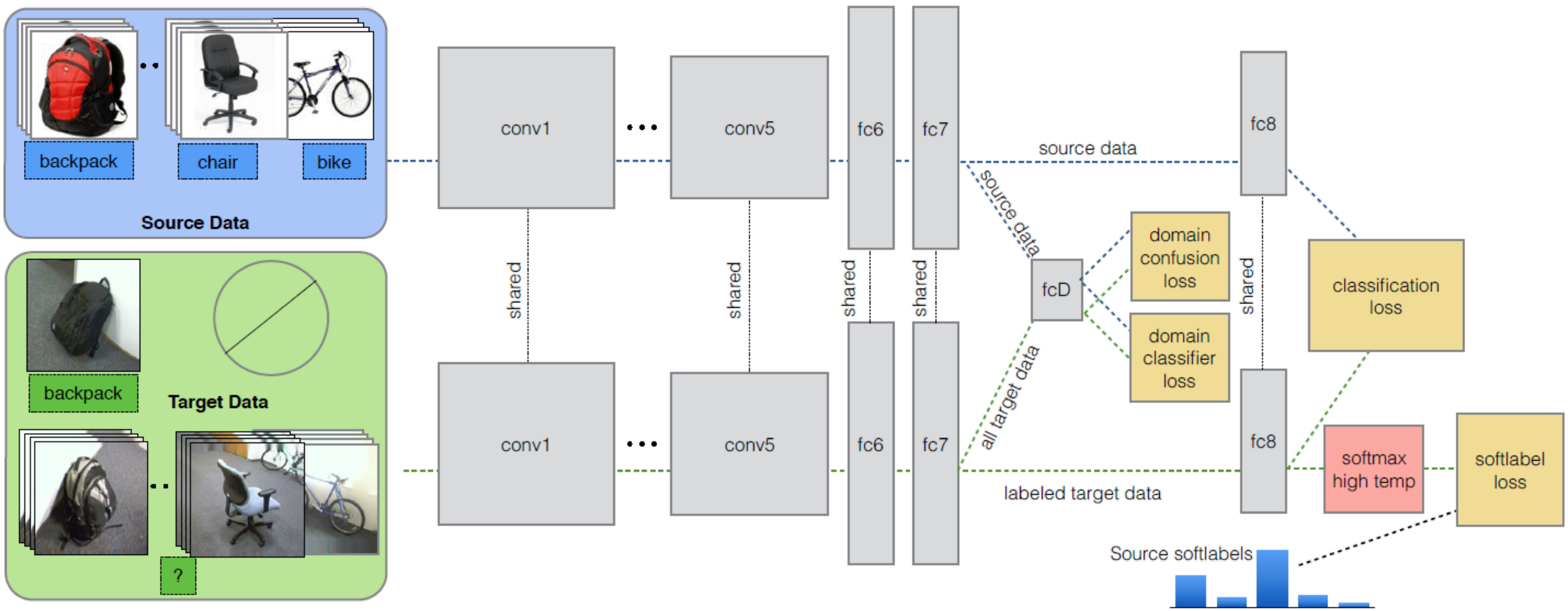
$$\mathrm{MMD}^2(\mathcal{D}_s, \mathcal{D}_t) = \left\| \frac{1}{M}\sum_{i=1}^{M}\phi(\mathbf{x}_i^s) - \frac{1}{N}\sum_{j=1}^{N}\phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2$$

$$\frac{1}{n_t}\sum_{j=1}^{n_t}\|\mathbf{h}_{t,j}^{(L)} - \sum_{c=1}^{C} p_{t,c}^{j}\bar{\mathbf{h}}_{s,c}^{(L)}\|_2^2$$

- the center of the c-th class source data;
- the probability of the j-th target point to be assigned to the label of the c-th class (***Softmax output of target data***)

Zhengming Ding, Nasser Nasrabadi, and Yun Fu. Semi-supervised Deep Domain Adaptation via Coupled Neural Networks, IEEE Transactions on Image Processing (TIP), 2018

Deep Learning+ Adversarial loss



$$\begin{aligned}\mathcal{L}(x_S, y_S, x_T, y_T, \theta_D; \theta_{\text{repr}}, \theta_C) = \; & \mathcal{L}_C(x_S, y_S, x_T, y_T; \theta_{\text{repr}}, \theta_C) \\ & + \lambda \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) \\ & + \nu \mathcal{L}_{\text{soft}}(x_T, y_T; \theta_{\text{repr}}, \theta_C).\end{aligned}$$

Classifier loss

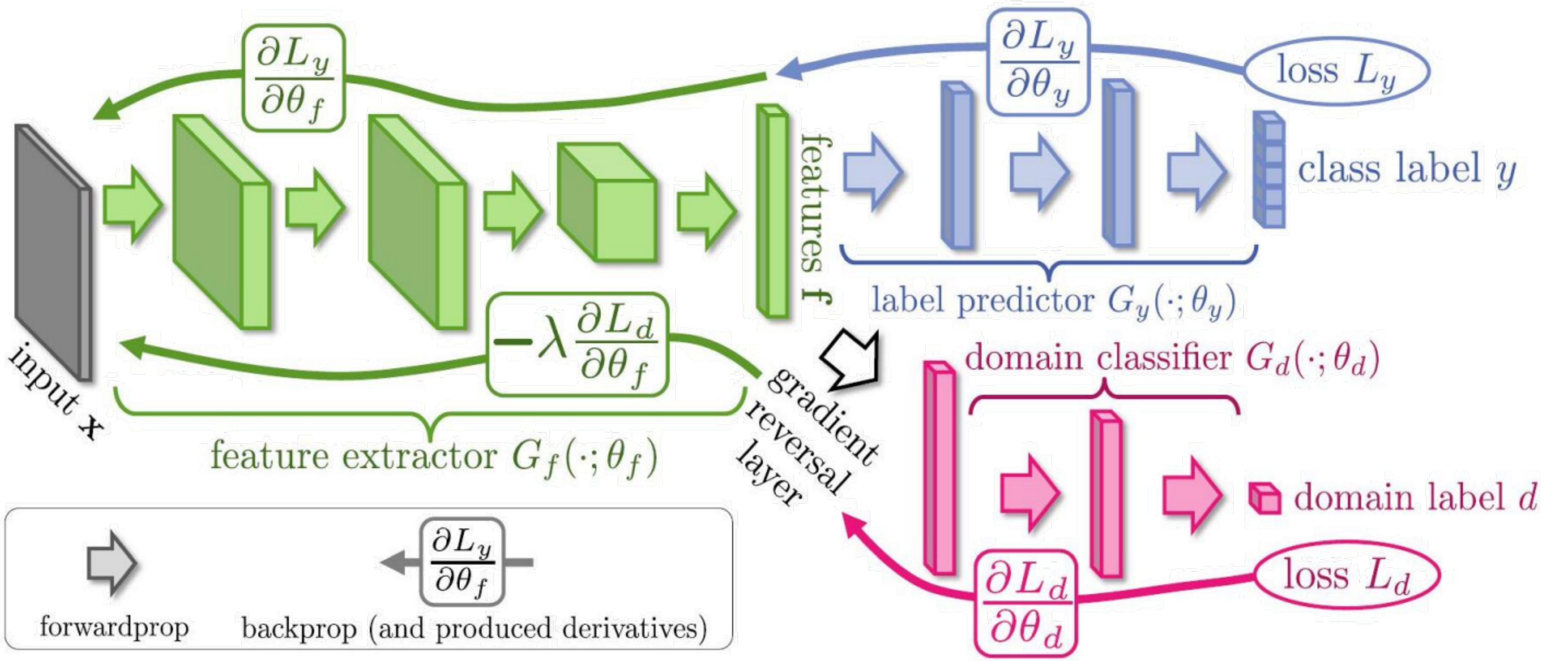
Domain confusion loss

Soft label loss

Tzeng, Eric, et al. "Simultaneous deep transfer across domains and tasks." *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015.

Deep Learning+ Adversarial loss

Northeastern University
Smile lab
Synergetic Media Learning Lab



$$E(\theta_f,\theta_y,\theta_d) = \sum_{\substack{i=1..N\\ d_i=0}} L_y\left(G_y(G_f(\mathbf{x}_i;\theta_f);\theta_y),y_i\right) - \lambda \sum_{i=1..N} L_d\left(G_d(G_f(\mathbf{x}_i;\theta_f);\theta_d),y_i\right)$$

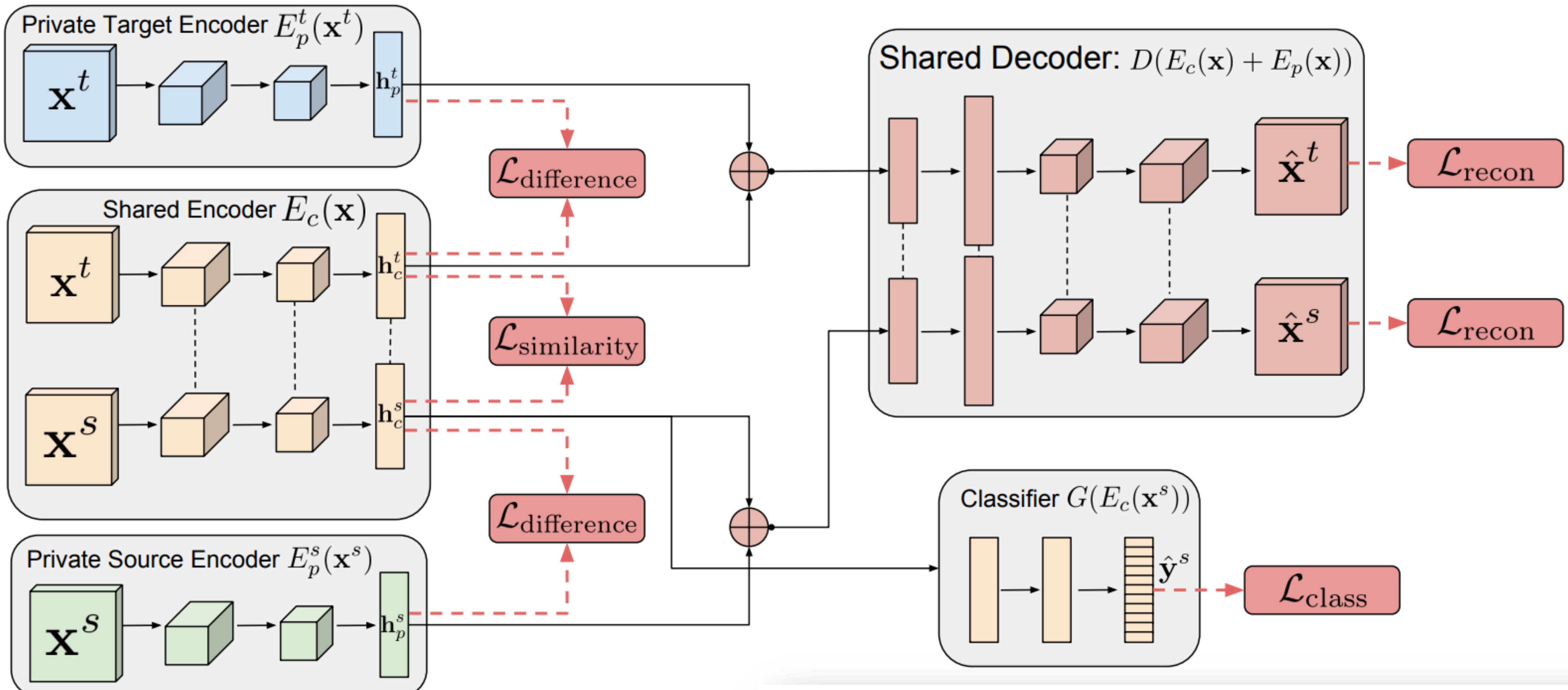
Ganin, Yaroslav, and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *ICML, 2015*

Deep Learning+ Adversarial loss

Northeastern University

Smile lab

Synergetic Media Learning Lab



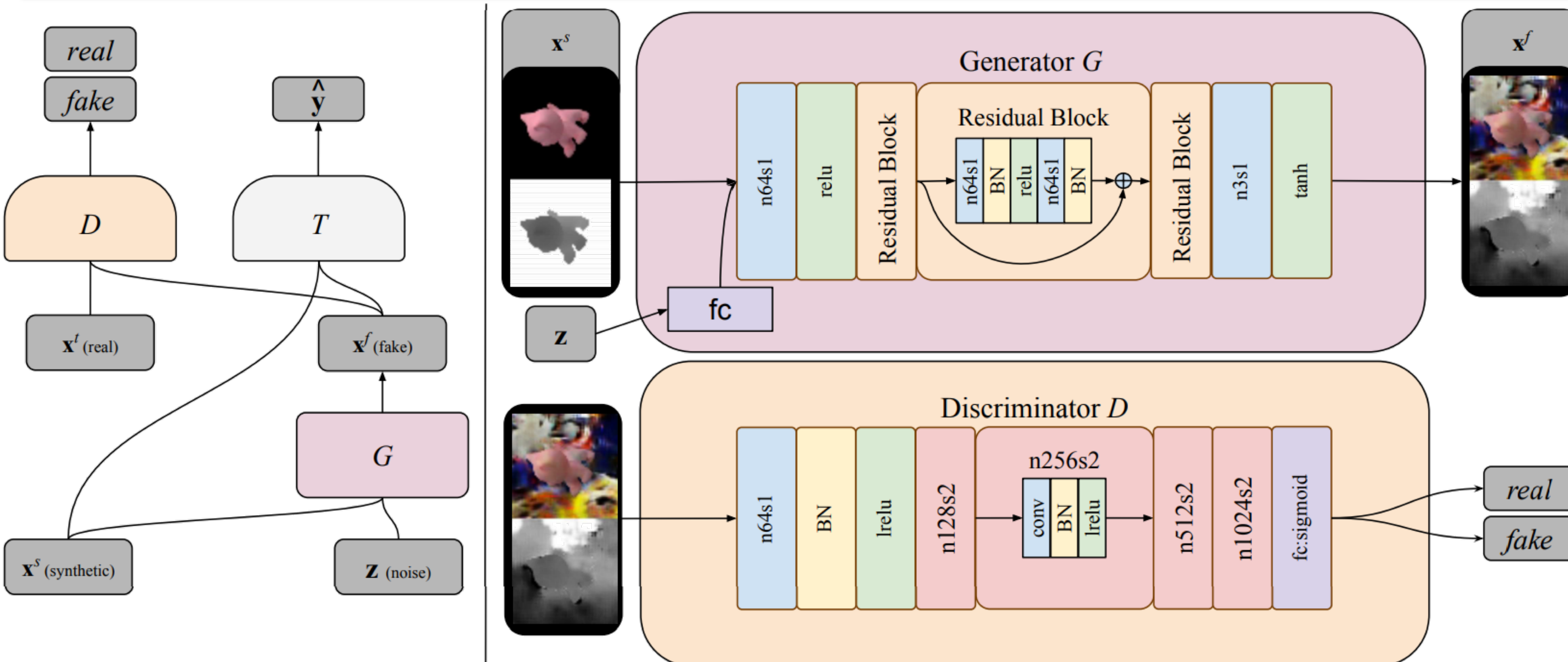
$$\mathcal{L}_{\text{difference}} = \left\| {\mathbf{H}_c^s}^\top \mathbf{H}_p^s \right\|_F^2 + \left\| {\mathbf{H}_c^t}^\top \mathbf{H}_p^t \right\|_F^2$$

$$\mathcal{L}_{\text{similarity}}^{\text{DANN}} = \sum_{i=0}^{N_s+N_t} \left\{ d_i \log \hat{d}_i + (1-d_i)\log(1-\hat{d}_i) \right\}$$

Bousmalis, Konstantinos, et al. "Domain separation networks." *Advances in Neural Information Processing Systems*. 2016.

Deep Learning+ Adversarial loss

Northeastern University
Smile lab
Synergetic Media Learning Lab



masked-PMSE loss

$$\min_{\boldsymbol{\theta}_G, \boldsymbol{\theta}_T} \max_{\boldsymbol{\theta}_D} \alpha \mathcal{L}_d(D, G) + \beta \mathcal{L}_t(T, G) + \gamma \boxed{\mathcal{L}_c(G)}$$

$$\mathcal{L}_c(G) = \mathbb{E}_{\mathbf{x}^s, \mathbf{z}} \Big[\frac{1}{k} \left\| (\mathbf{x}^s - G(\mathbf{x}^s, \mathbf{z}; \boldsymbol{\theta}_G)) \circ \mathbf{m} \right\|_2^2 - \frac{1}{k^2} \left((\mathbf{x}^s - G(\mathbf{x}^s, \mathbf{z}; \boldsymbol{\theta}_G))^\top \mathbf{m} \right)^2 \Big]$$

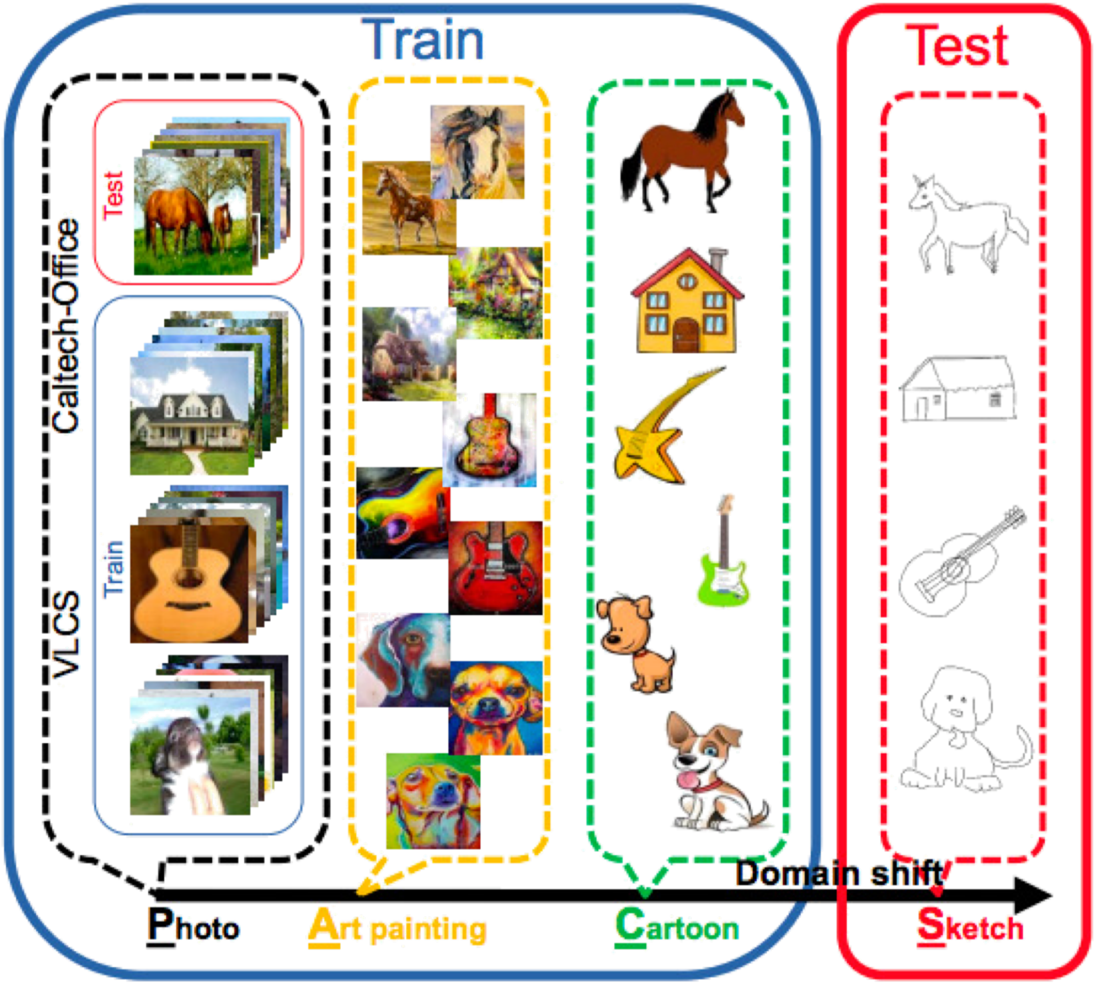
Bousmalis, Konstantinos, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CVPR*. 2017.

Outline

Northeastern University

- Introduction & Background
  - Multi-view Visual Data
  - Multi-view Learning Problems
  - Multi-view Learning Taxonomy
- Multi-view Learning
  - Projection and Embedding
  - Knowledge Fusion
  - Multi-view Clustering
  - Supervised Multi-view Learning → Zero-shot Learning
- Domain Adaptation
  - Transfer Learning → Domain Adaptation
  - Multi-Source Domain Adaptation & Domain Generalization
- Conclusion

Multi-Source Domains



Domain Adaptation

- Labeled Source
- Unlabeled or Limited Labeled Target

Multi-Source Domain Adaptation

- Multiple Labeled Sources
- Unlabeled or Limited Labeled Target

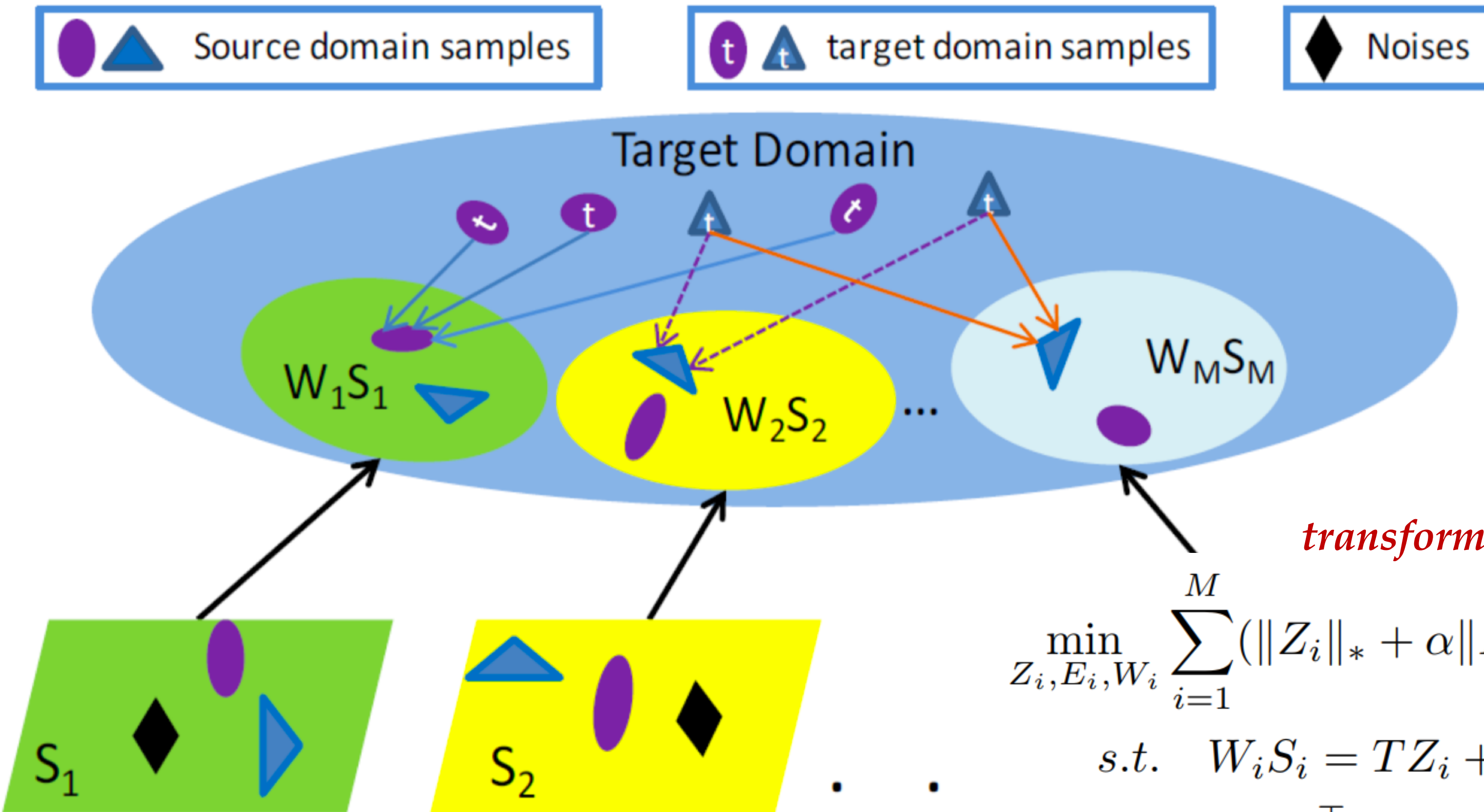
Domain Generalization

- Multiple Labeled Sources
- Missing Targets

http://homepages.inf.ed.ac.uk/thospeda/papers/li2017dg.pdf

Reconstruction loss

Northeastern University
Smile lab
Synergetic Media Learning Lab



transformed source domains

$$\min_{Z_i, E_i, W_i} \sum_{i=1}^{M} (\|Z_i\|_* + \alpha \|E_i\|_{2,1}) + \beta \boxed{\|Q\|_*}$$

$$s.t. \quad W_i S_i = T Z_i + E_i,$$

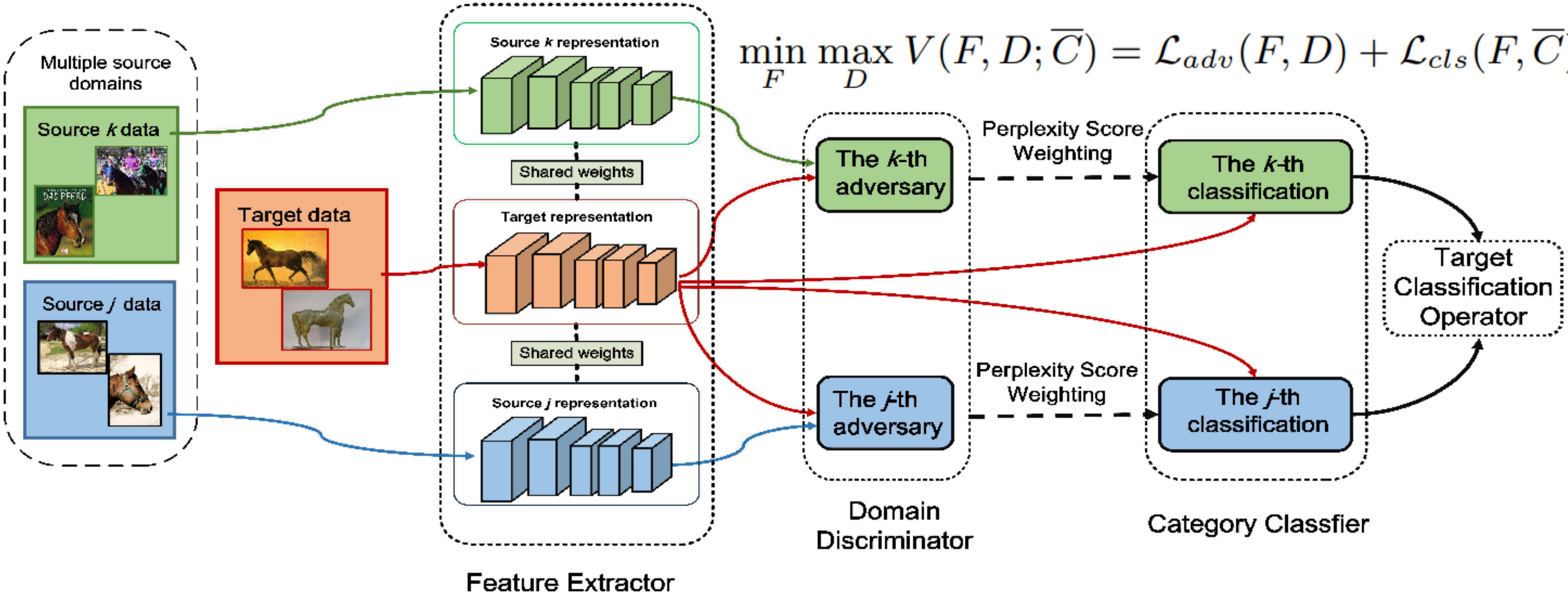
$$W_i W_i^\top = I, i = 1, ..., M,$$

Jhuo, I-Hong, et al. "Robust visual domain adaptation with low-rank reconstruction." *CVPR*, 2012.

Multi-Source Adaptation

Northeastern University
Smile lab
Synergetic Media Learning Lab

$$\min_{F}\max_{D} V(F, D; \overline{C}) = \mathcal{L}_{adv}(F, D) + \mathcal{L}_{cls}(F, \overline{C})$$



$$\mathcal{L}_{adv}(F, D) = \frac{1}{N}\sum_{j}^{N} \mathbb{E}_{x \sim X_{s_j}}[\log D_{s_j}(F(x))] + \mathbb{E}_{x^t \sim X_t}[\log(1 - D_{s_j}(F(x^t)))]$$

$$\min_{F,\, C} \mathcal{L}_{cls}(F, C) = \sum_{j}^{N} \mathbb{E}_{(x,y)\sim(X_{s_j}, Y_{s_j})}[\mathcal{L}(C_{s_j}(F(x)), y)] \boxed{+ \mathbb{E}_{(x^t, \hat{y})\sim(X_t^p, Y_t^p)}[\sum_{\hat{y}\in\mathcal{C}_{\hat{s}}} \mathcal{L}(C_{\hat{s}}(F(x^t)), \hat{y})]}$$

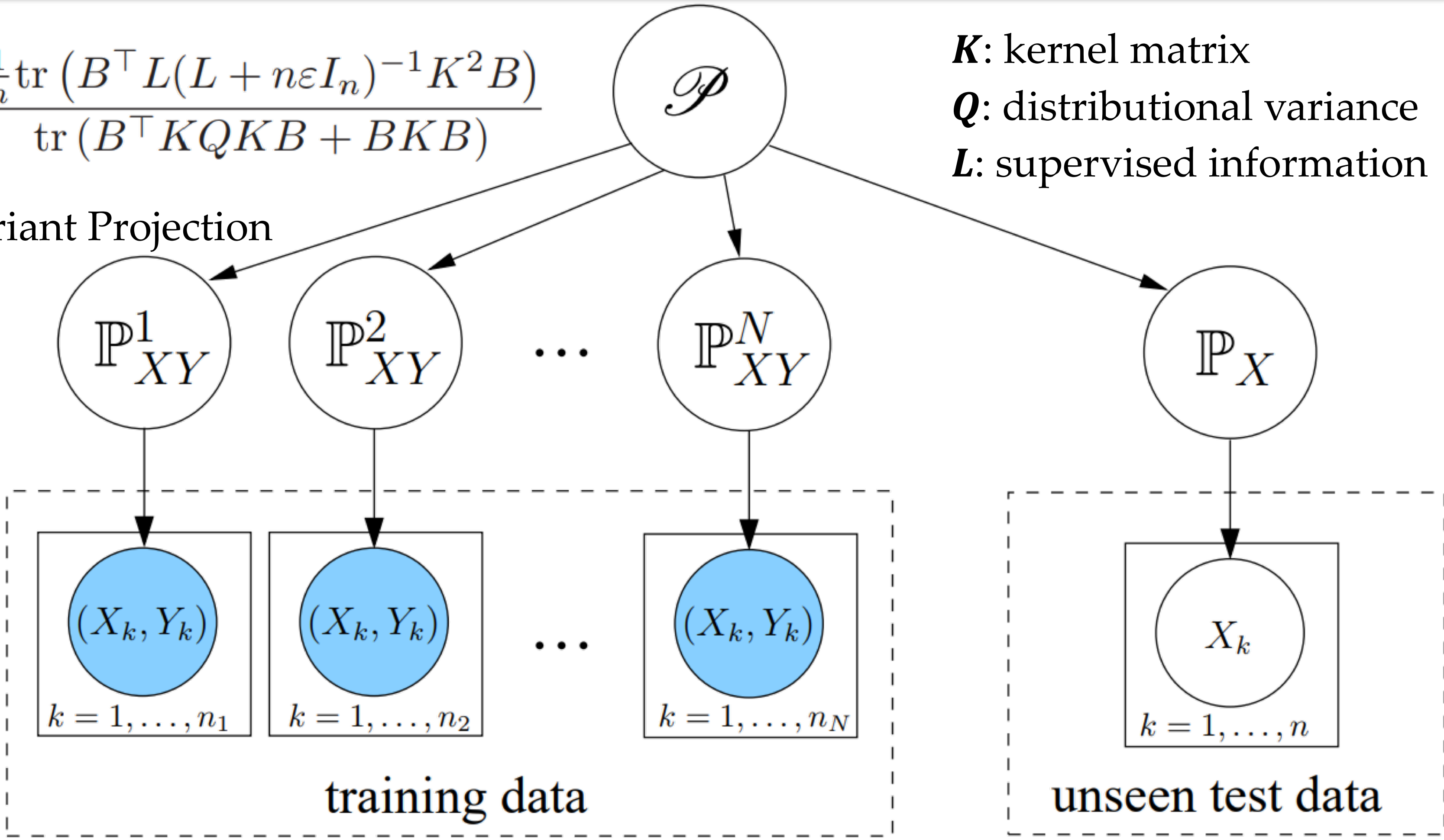
Xu, Ruijia, et al. "Deep Cocktail Network: Multi-source Unsupervised Domain Adaptation with Category Shift." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Domain Generalization

$$\max_{B \in \mathbb{R}^{n\times m}} \frac{\frac{1}{n}\mathrm{tr}\left(B^\top L(L + n\varepsilon I_n)^{-1}K^2 B\right)}{\mathrm{tr}\left(B^\top KQKB + BKB\right)}$$

K: kernel matrix
Q: distributional variance
L: supervised information

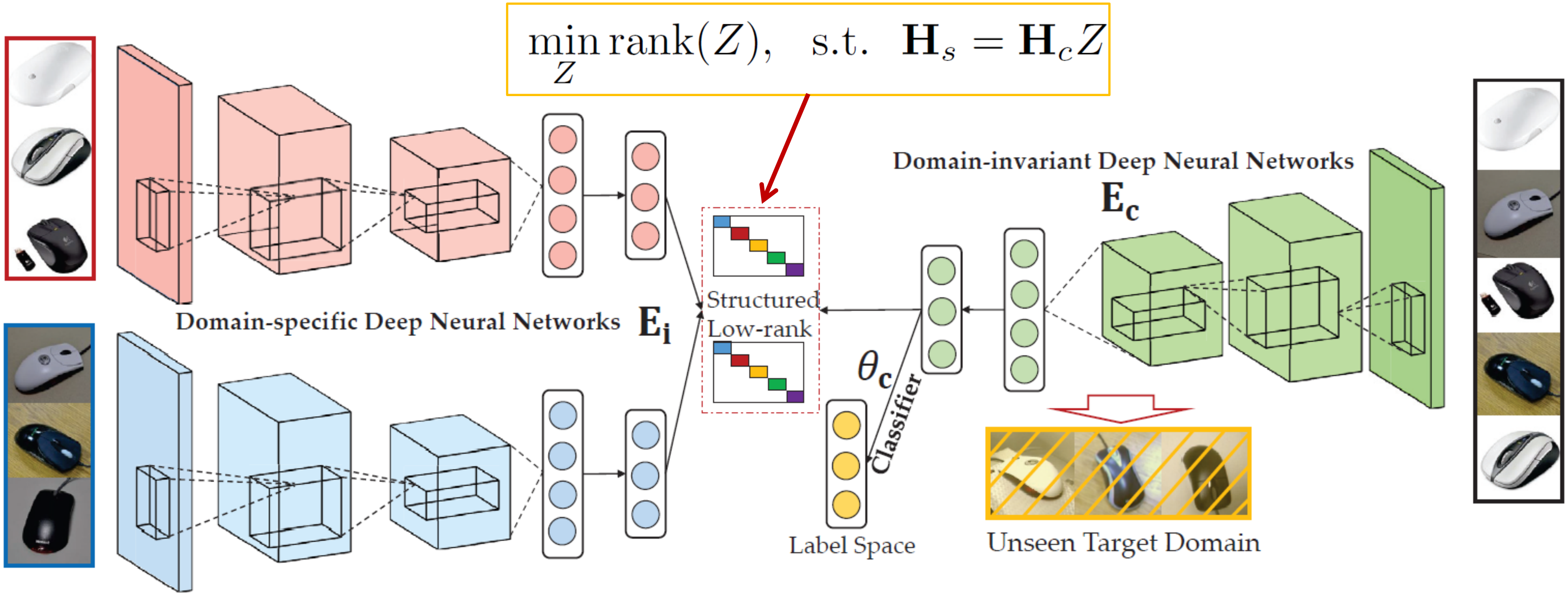
B: Invariant Projection



Muandet, Krikamol, David Balduzzi, and Bernhard Schölkopf. "Domain generalization via invariant feature representation." *International Conference on Machine Learning*. 2013.

Domain Generalization

Northeastern University
Smile lab
Synergetic Media Learning Lab

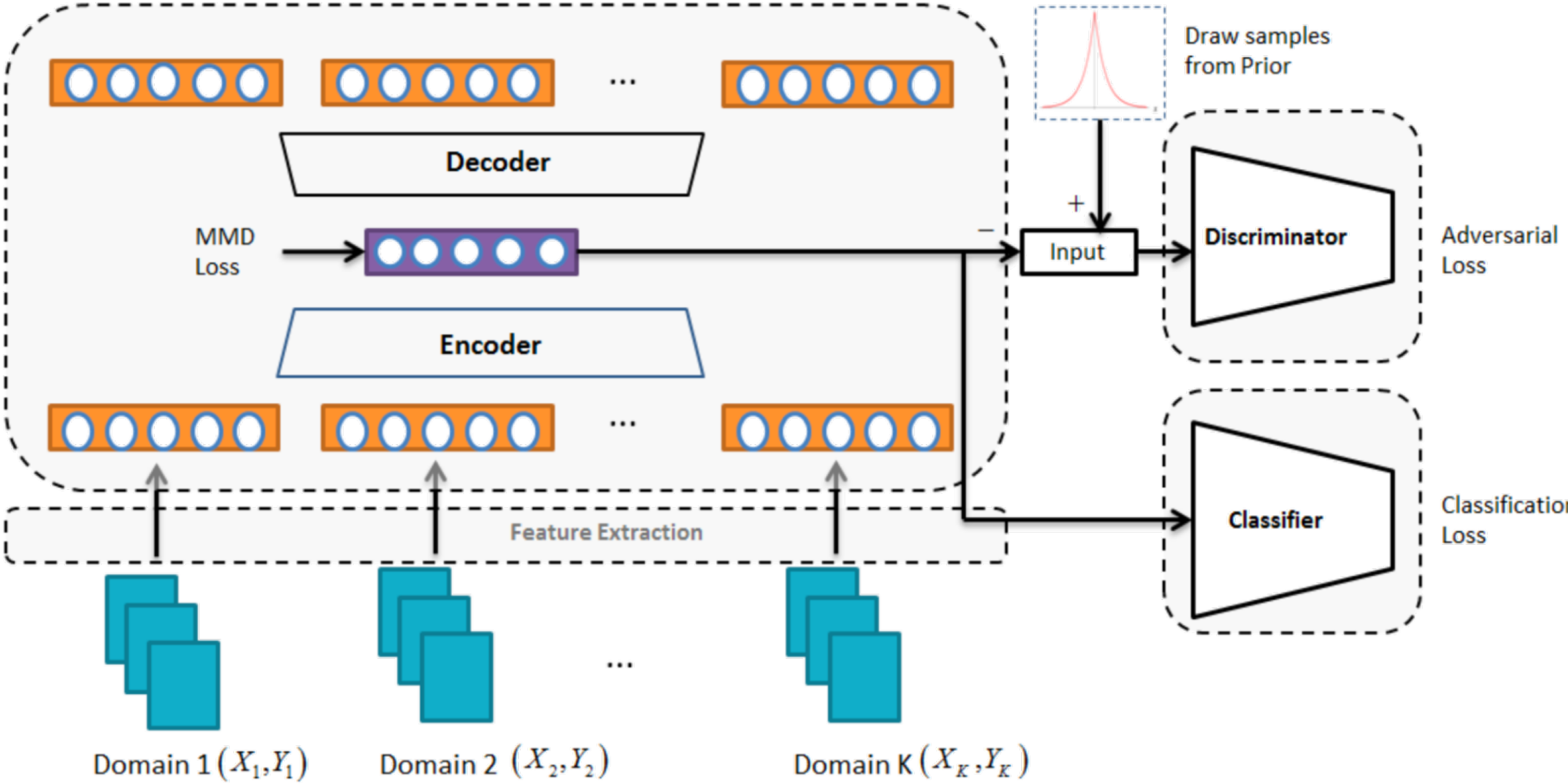


(a) Multiple domain-specific deep structures tend to be learned to capture the rich information from each source.
(b) A domain-invariant deep structure is built for all the domains, and further generalize to the unseen domain in the test stage with learned classifier .
(c) Low-rank reconstruction is adopted to align two types of networks in structured low-rank fashion.

Deep Domain Generalization with Structured Low-Rank Constraint, IEEE TIP 2018
Zhengming Ding and Yun Fu

Domain Generalization

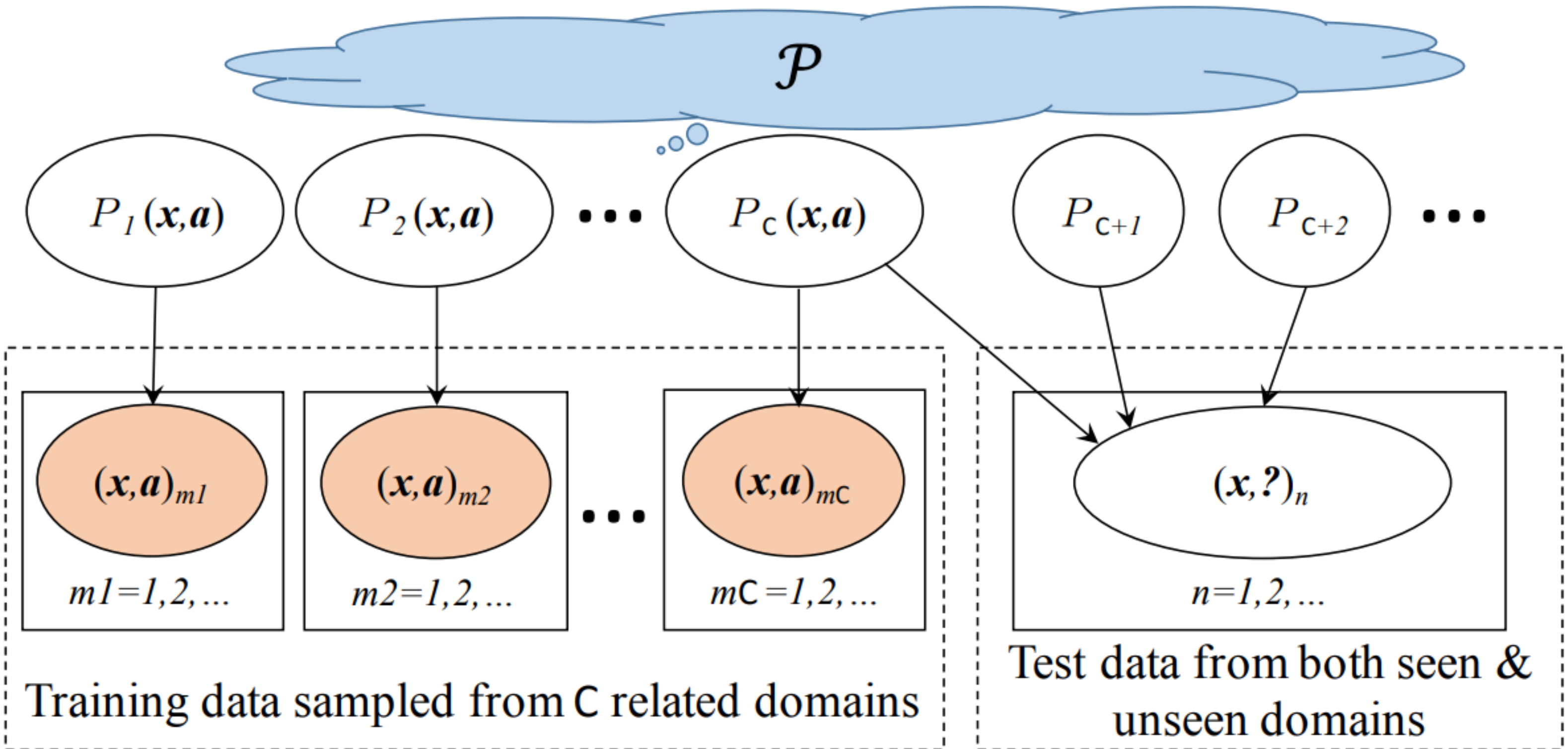
$$\min_{C,Q,P}\ \max_{D}\ \mathcal{L}_{\text{err}} + \lambda_0\mathcal{L}_{\text{ae}} + \lambda_1\mathcal{R}_{\text{mmd}} + \lambda_2\mathcal{J}_{\text{gan}}$$



Li, Haoliang, et al. "Domain generalization with adversarial feature learning." *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*. 2018.

Domain Generalization & Zero-Shot Learning

Northeastern University



$$\max_{B} \quad \frac{\operatorname{tr}(\gamma B^T K^2 B/\mathrm{M} + (1-\gamma) B^T K L K B)}{\operatorname{tr}(B^T K Q K B + B^T K B)}$$

Gan, Chuang, Tianbao Yang, and Boqing Gong. "Learning attributes equals multi-source domain generalization." CVPR. 2016.

Outline

Northeastern University

- ❑ **Introduction & Background**
  - Multi-view Visual Data
  - Multi-view Learning Problems
  - Multi-view Learning Taxonomy
- ❑ **Multi-view Learning**
  - Projection and Embedding
  - Knowledge Fusion
  - Multi-view Clustering
  - Supervised Multi-view Learning → Zero-shot Learning
- ❑ **Domain Adaptation**
  - Transfer Learning → Domain Adaptation
  - Multi-Source Domain Adaptation & Domain Generalization
- ❑ **Conclusion**

Conclusion

Northeastern University

Synergetic Media Learning Lab

Taxonomy

- **Multi-view Data** — sample-wise correspondence; class-wise correspondence
- **Multi-view Problems** — projection/embedding; clustering; classification
  zero-shot learning, domain adaptation/generalization
- **Multi-view Knowledge** — Knowledge Integration; Knowledge Transfer

Unified model

From Shallow To Deep

Our focus

Feature Learning + View Alignment

$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \mathcal{A}(f_i(X_i), f_j(X_j)) + \lambda \sum_{k=1}^{v} \mathcal{R}(f_k(X_k))$$

Northeastern University

Thank you!

Q& A