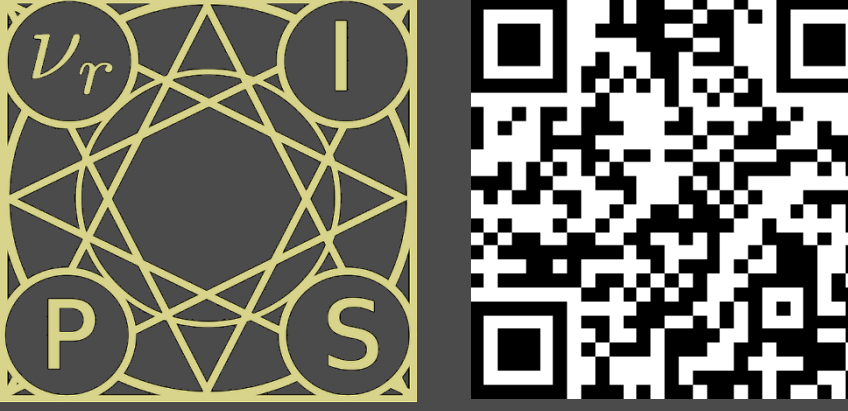


# DM2C: Deep Mixed-Modal Clustering

Yangbangyan Jiang<sup>1,2</sup>, Qianqian Xu<sup>3</sup>, Zhiyong Yang<sup>1,2</sup>, Xiaochun Cao<sup>1,2,5</sup>, Qingming Huang<sup>2,3,4,5\*</sup>

<sup>1</sup>Institute of Information Engineering, CAS   <sup>2</sup>University of Chinese Academy of Sciences

<sup>3</sup>Institute of Computing Technology, CAS   <sup>4</sup>BDKM, CAS   <sup>5</sup>Peng Cheng Laboratory



## Motivation

Traditional multi-modal learning requires **extra pairing information** among modalities for feature alignment.

- e.g., full/partial modality pairing, 'must/cannot link' constraints, co-occurrence frequency...

Table 1: Types of learning under multiple modalities

Type	Supervision	
	Class Label	Modality Pairing
Supervised Multi-modal Learning	✓	✓
Unsupervised Multi-modal Learning	✗	✓
<b>Unsupervised Mixed-modal Learning</b>	✗	✗

## Mixed-modal Clustering

**Mixed-modal:** Each instance is represented in only one modality.

Dataset  $\mathcal{D} \rightarrow \mathcal{D}_A = \{x_i^{(a)}\}_{i=1}^{n_a}$  and  $\mathcal{D}_B = \{x_i^{(b)}\}_{i=1}^{n_b}$

Multi-modal data

$\mathcal{D}_A$	$x_1^{(a)}$	$x_2^{(a)}$	$x_3^{(a)}$	...	$x_{n-2}^{(a)}$	$x_{n-1}^{(a)}$	$x_n^{(a)}$
$\mathcal{D}_B$	$x_1^{(b)}$	$x_2^{(b)}$	$x_3^{(b)}$	...	$x_{n-2}^{(b)}$	$x_{n-1}^{(b)}$	$x_n^{(b)}$

pairing information

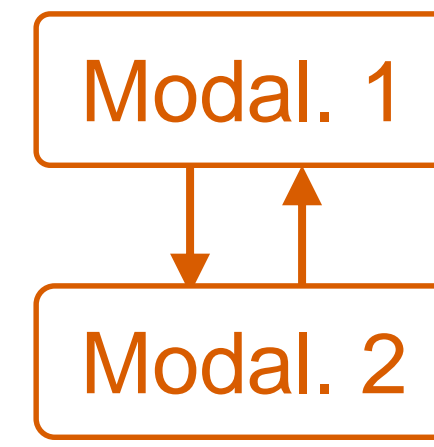
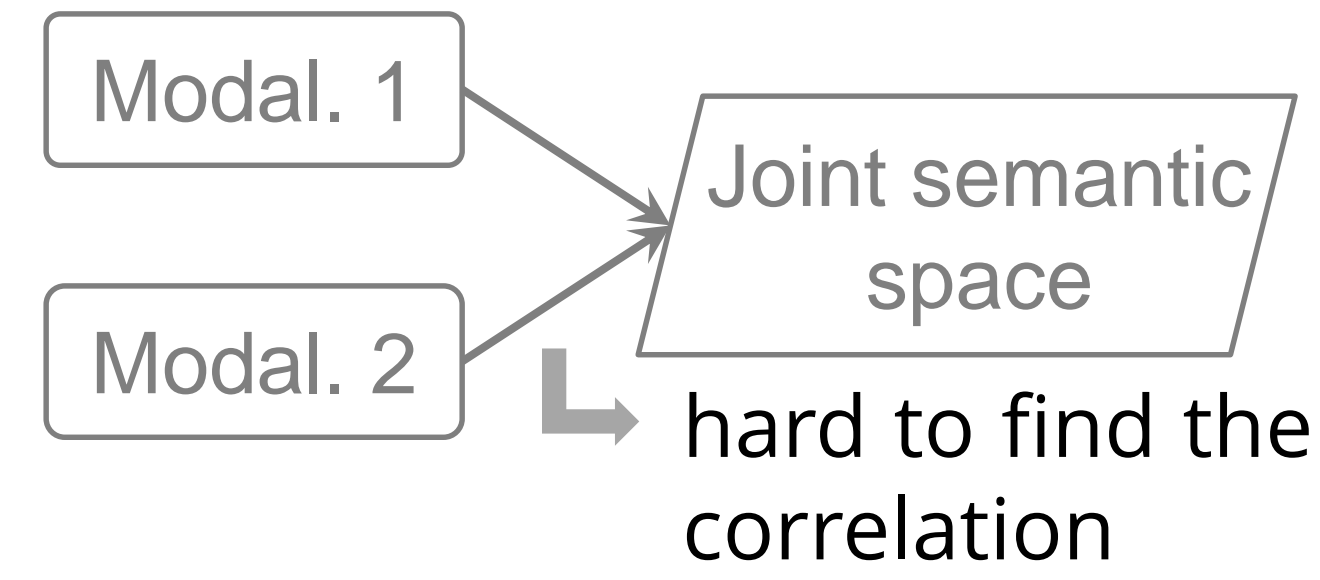
Mixed-modal data

$\mathcal{D}_A$	$x_1^{(a)}$	-	-	...	$x_{n_a-1}^{(a)}$	-	$x_{n_a}^{(a)}$
$\mathcal{D}_B$	-	$x_1^{(b)}$	$x_2^{(b)}$	...	-	$x_{n_b}^{(b)}$	-

hard for feature alignment

**Goal:** learning unified representations for the modalities, then grouping the samples into  $k$  categories.

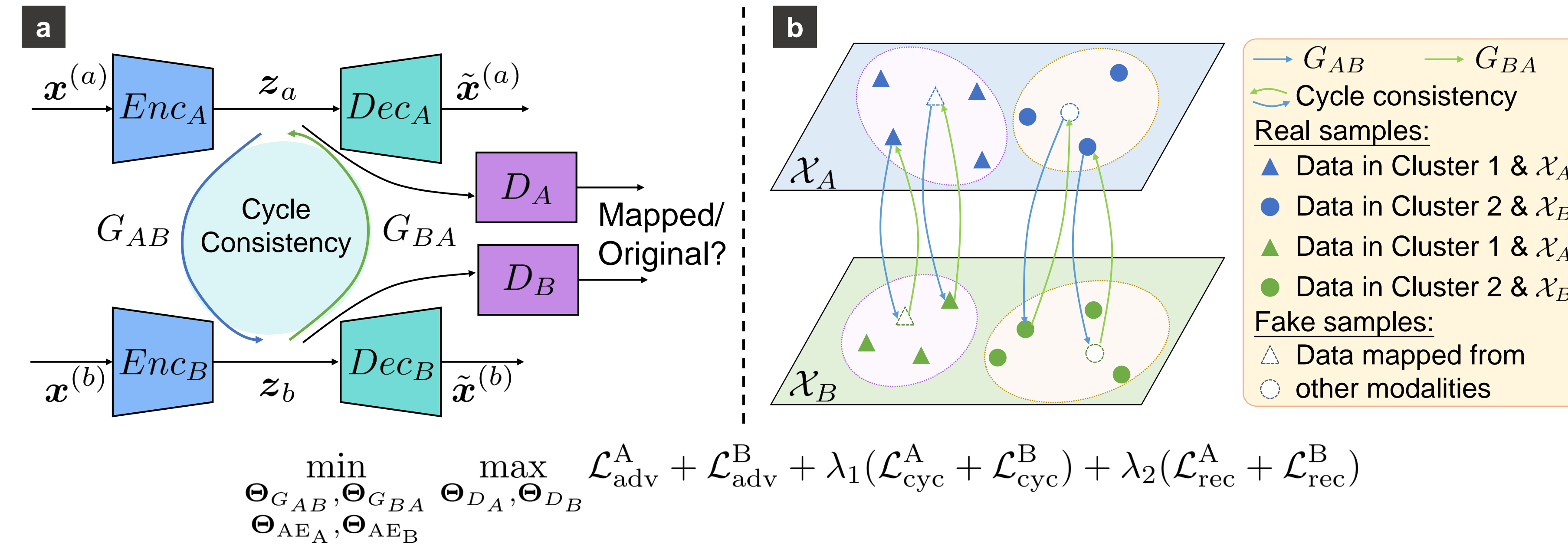
## Modality unifying



### learn the cross-modal translation

- easy to obtain via *cycle-consistency*
- unifying:** transforming all the samples into a modality specific space

## Framework



- Modality-specific Auto-Encoders: latent representations for each modality

$$\mathcal{L}_{rec}^A(\Theta_{AE_A}) = \|x_i^{(a)} - \text{Dec}_A(\text{Enc}_A(x_i^{(a)}))\|_2^2,$$

$$\mathcal{L}_{rec}^B(\Theta_{AE_B}) = \|x_i^{(b)} - \text{Dec}_B(\text{Enc}_B(x_i^{(b)}))\|_2^2,$$

- Unpaired Cross-Modal Mappings via *cycle-consistency*

$$\mathcal{L}_{cyc}^A(\Theta_{G_{AB}}, \Theta_{G_{BA}}) = \mathbb{E}_{z_a \sim \mathcal{X}_A} [\|z_a - G_{BA}(G_{AB}(z_a))\|_1],$$

$$\mathcal{L}_{cyc}^B(\Theta_{G_{AB}}, \Theta_{G_{BA}}) = \mathbb{E}_{z_b \sim \mathcal{X}_B} [\|z_b - G_{AB}(G_{BA}(z_b))\|_1].$$

- Adversarial learning between Cross-modal Mappings (Generators) and Discriminators

$$\mathcal{L}_{adv}^A(\Theta_{G_{BA}}, \Theta_{D_A}) = \mathbb{E}_{z_a \sim \mathcal{X}_A} [D_A(z_a)] - \mathbb{E}_{z_b \sim \mathcal{X}_B} [D_A(G_{BA}(z_b))],$$

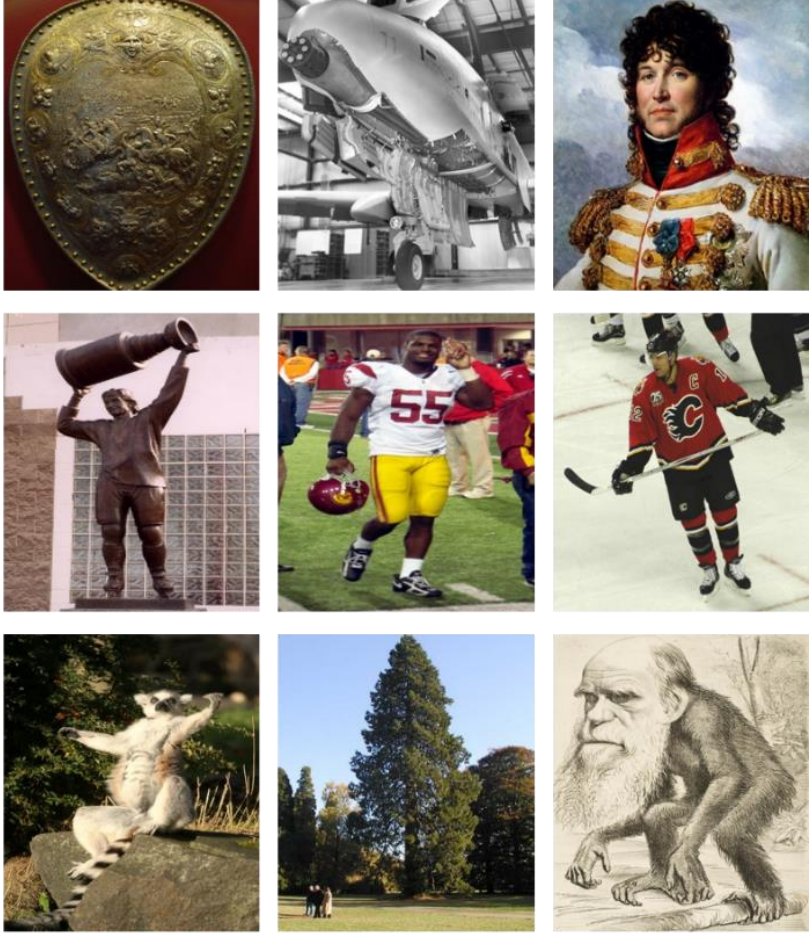
$$\mathcal{L}_{adv}^B(\Theta_{G_{AB}}, \Theta_{D_B}) = \mathbb{E}_{z_b \sim \mathcal{X}_B} [D_B(z_b)] - \mathbb{E}_{z_a \sim \mathcal{X}_A} [D_B(G_{AB}(z_a))].$$

## Results

Thousands of Luxembourgers overseas, unconstrained by the Luxembourgian government's need to remain neutral, signed up to serve with foreign armies. 3,700 Luxembourgian nationals served in the French Army, of whom over 2,000 died. As Luxembourg's pre-war population was only 266,000, the loss of life solely in the service of the French army amounted to almost 1% of the entire Luxembourgian population. (.....)

Yao did not participate in the Rockets' pre-season training camp, instead playing for China in the 2002 FIBA World Championships. Before the season, several commentators, including Bill Simmons and Dick Vitale, predicted that Yao would fail in the NBA, and Charles Barkley said he would "kiss fifth season". Yao averaged a career-high 25 points per game. In 25 games after the All-Star break, Yao averaged 25.7 points and 11.6 rebounds per game. (.....)

The median longevity of Beagles is 12.35 years, which is a typical lifespan for a dog of their size. Beagles may be prone to epilepsy, but this can be controlled with medication. Hypothyroidism and a number of types of dwarfism occur in Beagles. Two conditions in particular are unique to the breed: Funny Puppy, in which the puppy is slow to develop and eventually develops weak legs, a crooked back and although normally healthy, is prone to range of illnesses. (.....)



Samples on the Wikipedia dataset

Table 2: Performance comparisons on Wikipedia.

Algorithm	Accuracy	ARI	NMI	F-score	Purity
$k$ -means	0.2291	0.0166	0.1003	<b>0.1857</b>	0.2301
DKM	0.2173	0.0108	0.1170	0.1729	0.2429
DCN	0.2215	0.0137	<b>0.1172</b>	0.1688	0.2465
IDEC	0.2153	0.0375	0.0849	0.1654	0.2606
IMSAT	<b>0.2521</b>	<b>0.0573</b>	0.1093	0.1738	<b>0.2720</b>
Ours	<b>0.2720</b>	<b>0.0558</b>	<b>0.1543</b>	<b>0.1878</b>	<b>0.3075</b>

Table 3: Performance comparisons on NUS-WIDE-10K.

Algorithm	Accuracy	ARI	NMI	F-score	Purity
$k$ -means	0.2744	0.0044	0.0469	0.3008	0.5208
DKM	0.2932	0.0130	0.0116	0.2901	0.5036
DCN	0.3036	<b>0.0144</b>	<b>0.0512</b>	0.2959	<b>0.5296</b>
IDEC	0.3045	0.0006	0.0082	0.3048	0.5036
IMSAT	<b>0.3080</b>	0.0038	0.0064	<b>0.3422</b>	0.5036
Ours	<b>0.3300</b>	<b>0.0710</b>	<b>0.0951</b>	<b>0.3043</b>	<b>0.5492</b>