DM2C: Deep Mixed-Modal Clustering

Yangbangyan Jiang[1,2], Qianqian Xu[3], Zhiyong Yang[1,2], Xiaochun Cao[1,2,5], Qingming Huang[2,3,4,5*]

[1]Institute of Information Engineering, CAS [2]University of Chinese Academy of Sciences

[3]Institute of Computing Technology, CAS [4]BDKM, CAS [5]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...

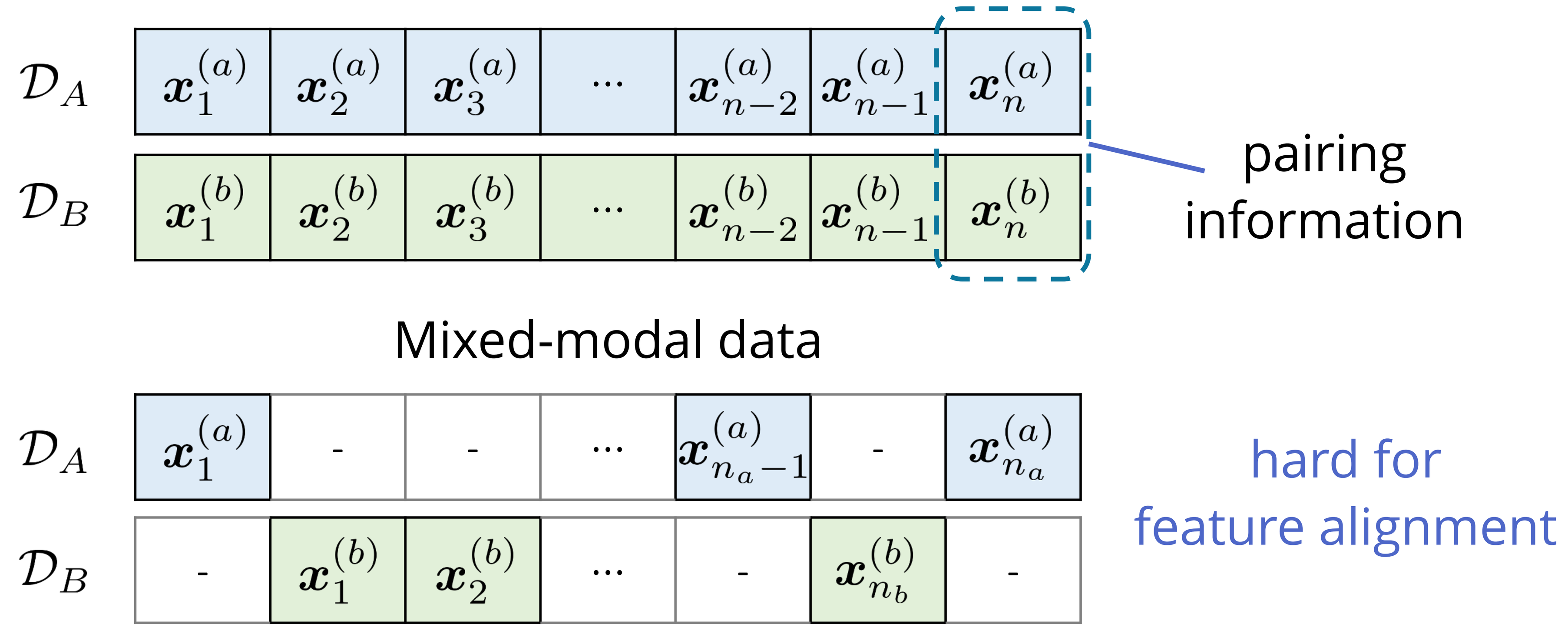
Table 1: Types of learning under multiple modalities

Type	Supervision: Class Label	Supervision: Modality Pairing
Supervised Multi-modal Learning	✓	✓
Unsupervised Multi-modal Learning	✗	✓
Unsupervised Mixed-modal Learning	✗	✗

Mixed-modal Clustering

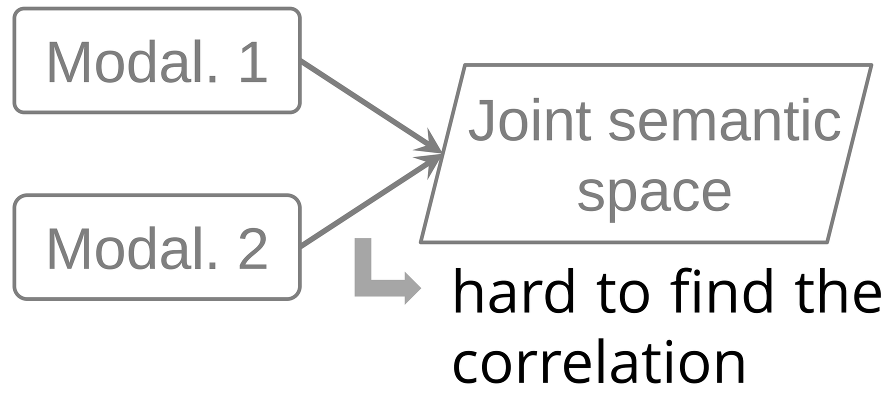
Mixed-modal: Each instance is represented in only one modality.

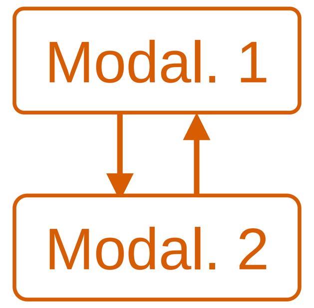
Dataset $\mathcal{D}$ ⟶ $\mathcal{D}_A = \left\{\boldsymbol{x}_i^{(a)}\right\}_{i=1}^{n_a}$ and $\mathcal{D}_B = \left\{\boldsymbol{x}_i^{(b)}\right\}_{i=1}^{n_b}$



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

Modality unifying

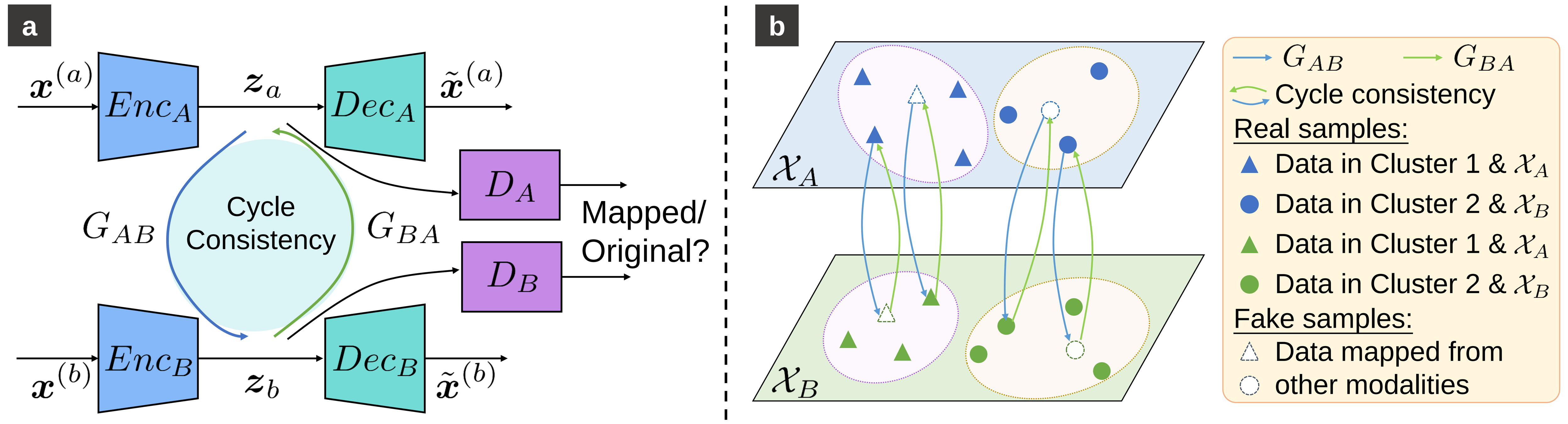




learn the cross-modal translation

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

Framework



$$\min_{\substack{\Theta_{G_{AB}},\Theta_{G_{BA}}\\ \Theta_{\mathrm{AE_A}},\Theta_{\mathrm{AE_B}}}} \max_{\Theta_{D_A},\Theta_{D_B}} \mathcal{L}_{\mathrm{adv}}^{\mathrm{A}} + \mathcal{L}_{\mathrm{adv}}^{\mathrm{B}} + \lambda_1(\mathcal{L}_{\mathrm{cyc}}^{\mathrm{A}} + \mathcal{L}_{\mathrm{cyc}}^{\mathrm{B}}) + \lambda_2(\mathcal{L}_{\mathrm{rec}}^{\mathrm{A}} + \mathcal{L}_{\mathrm{rec}}^{\mathrm{B}})$$

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

$$\mathcal{L}_{\mathrm{rec}}^{\mathrm{A}}(\Theta_{\mathrm{AE_A}}) = \|\boldsymbol{x}_i^{(a)} - \mathrm{Dec_A}(\mathrm{Enc_A}(\boldsymbol{x}_i^{(a)}))\|_2^2,$$

$$\mathcal{L}_{\mathrm{rec}}^{\mathrm{B}}(\Theta_{\mathrm{AE_B}}) = \|\boldsymbol{x}_i^{(b)} - \mathrm{Dec_B}(\mathrm{Enc_B}(\boldsymbol{x}_i^{(b)}))\|_2^2,$$

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

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

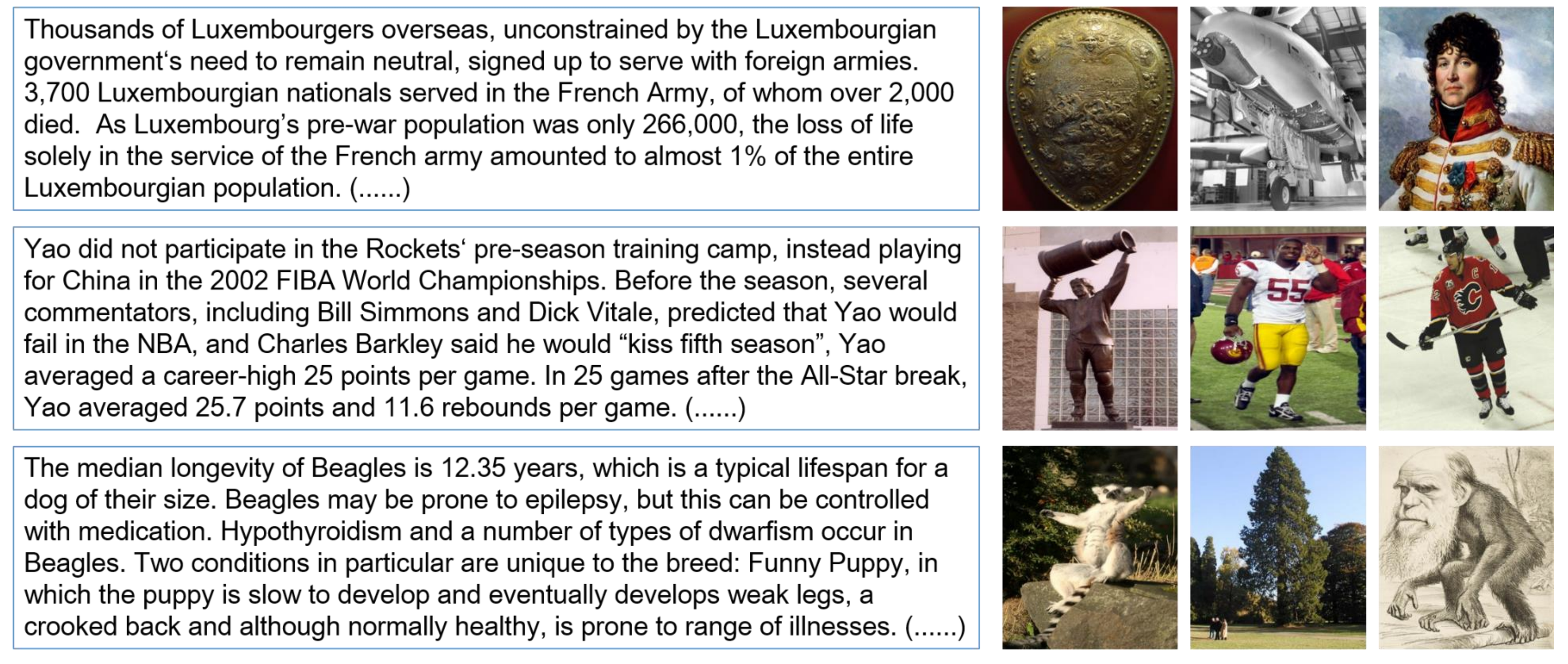
$$\mathcal{L}_{\mathrm{cyc}}^{\mathrm{B}}(\Theta_{G_{AB}},\Theta_{G_{BA}}) = \mathbb{E}_{\boldsymbol{z}_b\sim\mathcal{X}_B}\left[\|\boldsymbol{z}_b - G_{AB}(G_{BA}(\boldsymbol{z}_b))\|_1\right].$$

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

$$\mathcal{L}_{\mathrm{adv}}^{\mathrm{A}}(\Theta_{G_{BA}},\Theta_{D_A}) = \mathbb{E}_{\boldsymbol{z}_a\sim\mathcal{X}_A}[D_A(\boldsymbol{z}_a)] - \mathbb{E}_{\boldsymbol{z}_b\sim\mathcal{X}_B}[D_A(G_{BA}(\boldsymbol{z}_b))],$$

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

Results



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	0.1857	0.2301
DKM	0.2173	0.0108	0.1170	0.1729	0.2429
DCN	0.2215	0.0137	0.1172	0.1688	0.2465
IDEC	0.2153	0.0375	0.0849	0.1654	0.2606
IMSAT	0.2521	0.0573	0.1093	0.1738	0.2720
Ours	0.2720	0.0558	0.1543	0.1878	0.3075

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	0.0144	0.0512	0.2959	0.5296
IDEC	0.3045	0.0006	0.0082	0.3048	0.5036
IMSAT	0.3080	0.0038	0.0064	0.3422	0.5036
Ours	0.3300	0.0710	0.0951	0.3043	0.5492

Personal homepage: https://jiangyangby.github.io (with QR code) **Email:** jiangyangbangyan@iie.ac.cn **Group homepage:** https://qmhuang-ucas.github.io