TIP: Tabular-Image Pre-training for Multimodal Classification with Incomplete Data

Siyi Du†, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P. O'Regan, and Chen Qin†

Imperial College London, London, UK

{s.du23,s.zheng22,y.wang23,w.bai,declan.oregan,c.qin15}@imperial.ac.uk

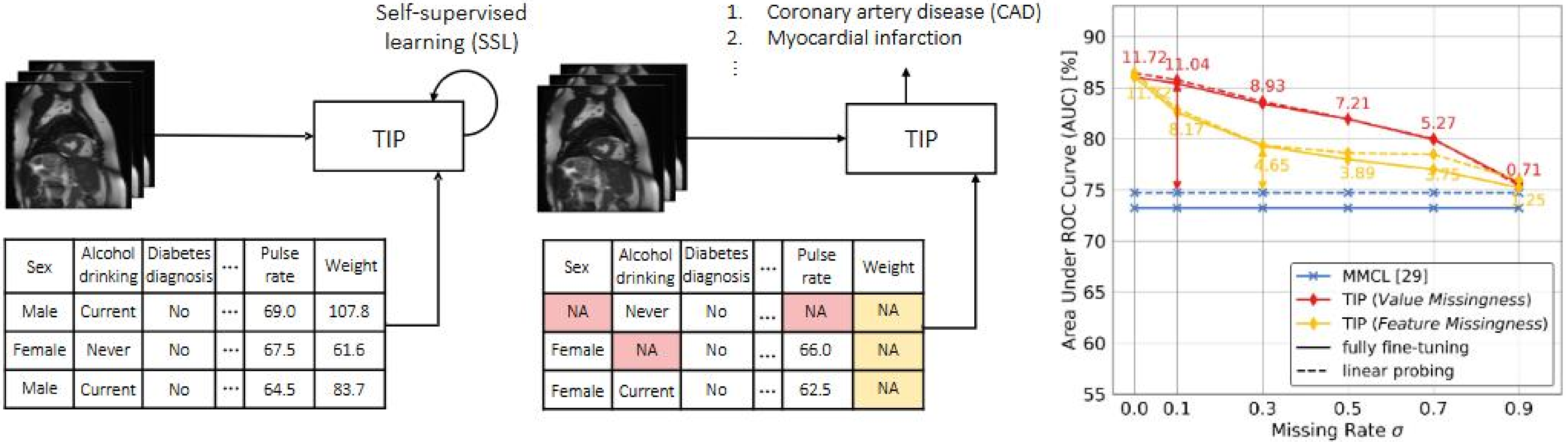
ECCV 2024

Background

In the medical field, rich tabular information (such as demographic data, lifestyle, and biochemical indicators) is often bound to image data, which is an important multimodal resource with high research value.

In order to solve the problem of **integration of image and tabular data**, especially when the tabular data is **incomplete or heterogeneous**, and improve the performance of **multimodal classification tasks**, the authors propose a new self-supervised learning method, TIP, which is a multimodal tabular image pre-training framework.

Pipeline



Sex	Alcohol drinking	Diabetes diagnosis	…	Pulse rate	Weight
Male	Current	No	…	69.0	107.8
Female	Never	No	…	67.5	61.6
Male	Current	No	…	64.5	83.7

Sex	Alcohol drinking	Diabetes diagnosis	…	Pulse rate	Weight
NA	Never	No	…	NA	NA
Female	NA	No	…	66.0	NA
Female	Current	No	…	62.5	NA

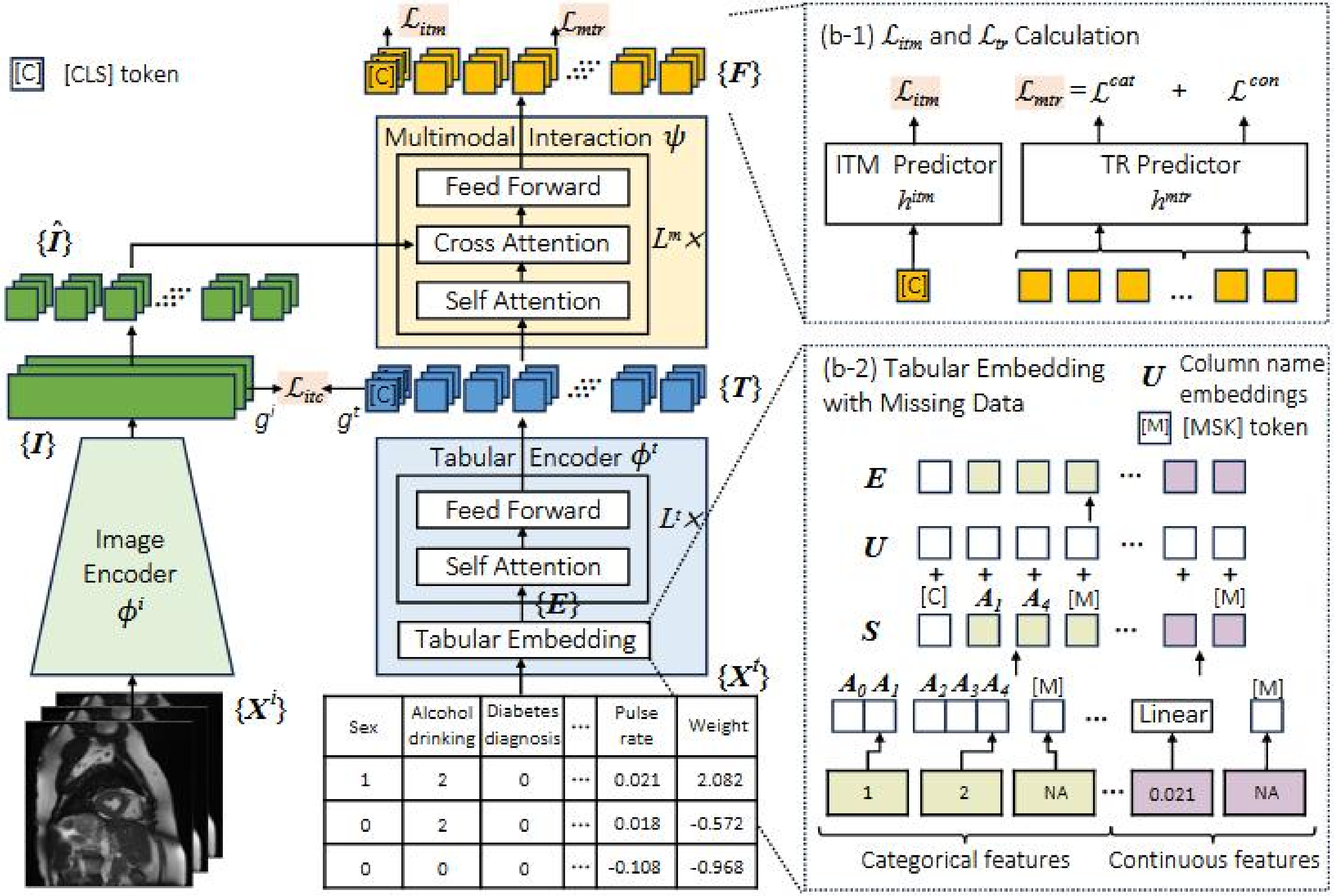
(a) Large-scale pre-training with complete image-tabular pairs

(b) Downstream task fine-tuning and inference with image and incomplete tabular data

(c) AUC results of coronary artery disease (CAD) classification in various missing rates

Architecture

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

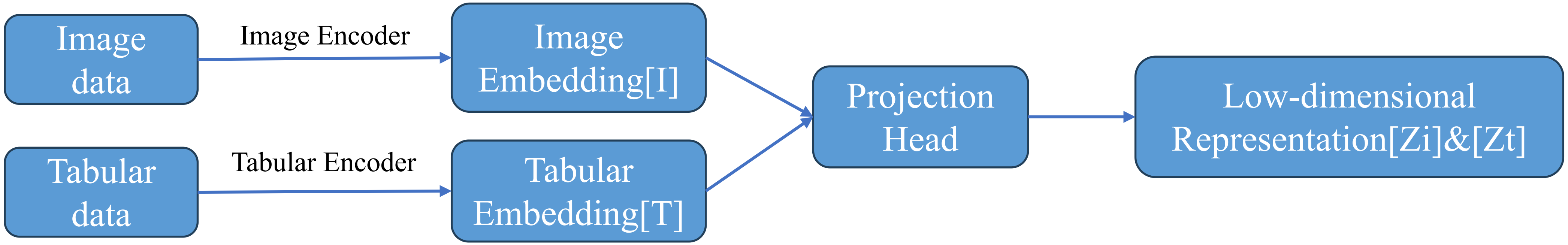


(a) Model Overview

(b) Model Details

Methods — Image-Tabular Contrastive Learning

Positive sample: Image and tabular data from the same instance form a positive sample pair.
Negative samples: Images and tables from different instances in the same batch are randomly combined to generate negative sample pairs.



Bidirectional Contrastive Loss

$s_{i\to t}$: $\quad s_{i\to t} = \mathrm{sim}(Z_i, Z_t)$

$s_{t\to i}$: $\quad s_{t\to i} = \mathrm{sim}(Z_t, Z_i)$

$$\mathcal{L}_{itc} = -\frac{1}{2}\left(\log \frac{\exp(s_{i\to t}/\tau)}{\sum_j \exp(s_{i\to t_j}/\tau)} + \log \frac{\exp(s_{t\to i}/\tau)}{\sum_j \exp(s_{t\to i_j}/\tau)}\right)$$

Methods — Image-Tabular Matching

Low-dimensional Representation[Zi]&[Zt]

cross-attention mechanism

$$CrossAttention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_k})\boldsymbol{V}$$

Joint representation[F]

[cls]

ITM predictor

$$\hat{y}_{itm} = \mathrm{Sigmoid}(\mathrm{ITM\_predictor}(F))$$

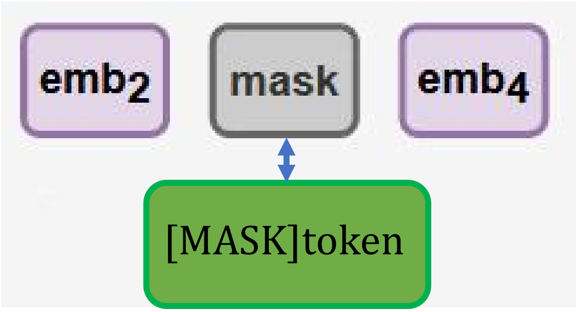
$$L_{\mathrm{itm}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$$

Methods — Masked Tabular Reconstruction

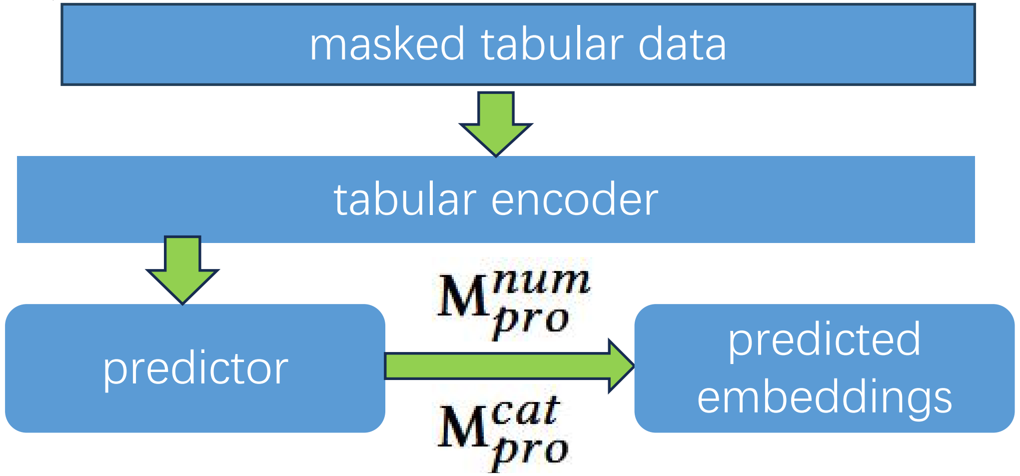
Step 1:Mask features

Sex	Alcohol drinking	Diabetes diagnosis	...	Pulse rate	Weight
NA	Never	No	...	NA	NA
Female	NA	No	...	66.0	NA
Female	Current	No	...	62.5	NA

Step 2: Replace masked features



Step 3:Reconstruct masked features



$$L_{\text{mtr, cont}} = \frac{1}{N_{\text{masked}}} \sum_{n \in \text{masked}} (x_{tn} - \hat{x}_{tn})^2$$

$$L_{\text{mtr, cat}} = \frac{1}{N_{\text{masked}}} \sum_{n \in \text{masked}} -\log\left(P(\hat{x}_{tn} = x_{tn})\right)$$

$$L_{\text{mtr}} = L_{\text{mtr, cont}} + L_{\text{mtr, cat}}$$

Experiments — classification result

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Model	DVM Accuracy (%) ↑		CAD AUC (%) ↑		Infarction AUC (%) ↑	
	❄	🔥	❄	🔥	❄	🔥
(a) Supervised Image and Multimodal Methods						
ResNet-50 [34]	87.68	87.68	63.11	63.11	59.48	59.48
Concat Fuse (CF) [68]	94.60	94.60	85.76	85.76	**85.05**	85.04
Max Fuse (MF) [73]	94.39	94.39	85.31	85.31	84.75	84.75
Interact Fuse (IF) [25]	96.24	96.24	84.89	84.89	81.91	81.91
DAFT [77]	96.60	96.60	86.21	**86.21**	56.27	56.27
(b) SSL Image Pre-training Methods						
SimCLR [19]	61.06	87.65	68.42	72.58	68.86	75.07
BYOL [29]	56.26	88.64	65.67	69.18	66.63	70.12
SimSiam [20]	23.14	78.62	57.77	67.71	53.83	64.79
BarlowTwins [84]	53.60	88.36	55.64	61.68	50.01	60.14
(c) SSL Multimodal Pre-training Methods						
MMCL [30]	91.66	93.27	74.71	73.21	76.79	76.46
TIP	**99.72**	**99.56**	**86.43**	86.03	84.46	**85.58**

Experiments — missing information prediction

Model	DVM RMSE ↓			UKBB RMSE ↓		
Missing rate σ	0.3	0.5	0.7	0.3	0.5	0.7
Mean [32]	0.9621	0.9783	0.9733	1.0162	1.0191	1.0070
MissForest [69]	0.6700	0.7653	0.8833	0.7516	0.7754	0.8177
GAIN [80]	1.0447	0.9428	2.9705	0.7920	2.0039	2.8130
MIWAE [54]	1.0105	1.0265	1.0218	1.0644	1.0680	1.0557
Hyperimpute [41]	0.6329	0.9428	0.9793	0.6803	0.7242	0.8060
TIP	**0.3899**	**0.4651**	**0.5055**	**0.6039**	**0.6460**	**0.7106**

Thanks