

Global and Local Contrastive Learning for Joint Representations from Cardiac MRI and ECG

Alexander Selivanov¹, Philip Müller^{1,2}, Özgün Turgut^{1,2}, Nil Stolt-Ansó^{1,4}, and Daniel Rueckert^{1,2,3,4}

¹ Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany

² School of Medicine, Klinikum rechts der Isar, TUM, Germany

³ Department of Computing, Imperial College London, UK

⁴ Munich Center for Machine Learning (MCML), Munich, Germany
alexander.selivanov@tum.de

Abstract. An electrocardiogram (ECG) is a widely used, cost-effective tool for detecting electrical abnormalities in the heart. However, it cannot directly measure functional parameters, such as ventricular volumes and ejection fraction, which are crucial for assessing cardiac function. Cardiac magnetic resonance (CMR) is the gold standard for these measurements, providing detailed structural and functional insights, but is expensive and less accessible. To bridge this gap, we propose PTACL (Patient and Temporal Alignment Contrastive Learning), a multimodal contrastive learning framework that enhances ECG representations by integrating spatio-temporal information from CMR. PTACL uses global patient-level contrastive loss and local temporal-level contrastive loss. The global loss aligns patient-level representations by pulling ECG and CMR embeddings from the same patient closer together, while pushing apart embeddings from different patients. Local loss enforces fine-grained temporal alignment within each patient by contrasting encoded ECG segments with corresponding encoded CMR frames. This approach enriches ECG representations with diagnostic information beyond electrical activity and transfers more insights between modalities than global alignment alone, all without introducing new learnable weights. We evaluate PTACL on paired ECG-CMR data from 27,951 subjects in the UK Biobank. Compared to baseline approaches, PTACL achieves better performance in two clinically relevant tasks: (1) retrieving patients with similar cardiac phenotypes and (2) predicting CMR-derived cardiac function parameters, such as ventricular volumes and ejection fraction. Our results highlight the potential of PTACL to enhance non-invasive cardiac diagnostics using ECG. The code is available at: <https://github.com/alsalivan/ecgcmr>

Keywords: Contrastive Learning · Time-Alignment · ECG · MRI

1 Introduction

Cardiovascular diseases (CVD) remain a leading cause of mortality worldwide [28]. Electrocardiograms (ECGs) are a non-invasive and cost-effective tool for

detecting electrical heart abnormalities (e.g., arrhythmias [1]). However, ECGs can’t directly assess structural and functional cardiac properties like ventricular volumes, myocardial mass, or ejection fraction. In contrast, cardiac magnetic resonance (CMR) is the gold standard for these assessments [11], but it is expensive, time-consuming, and requires specialized expertise, limiting its accessibility.

To integrate information from both modalities, researchers have applied contrastive learning to align ECG and CMR embeddings globally [5, 18, 25], pulling same-patient representations closer while pushing apart different patients. However, these approaches treat ECG and CMR as holistic representations, overlooking finer temporal relationships. Since CMR captures dynamic sequences across the cardiac cycle while ECG records continuous electrical activity, global alignment alone lacks the temporal precision needed to effectively integrate both modalities.

To address this limitation, we introduce PTACL (**P**atient and **T**emporal **A**lignment **C**ontrastive **L**earning), a contrastive learning framework that combines global and local alignment. The global loss aligns ECG and CMR embeddings at the patient level, while the local loss enforces fine-grained temporal alignment by contrasting ECG-derived segments—encoded tokens from a single heartbeat representation—with their corresponding CMR frame representations at the same cardiac phase (Fig. 1). Notably, the local alignment is completely parameter-free, introducing no additional learnable parameters. Our method is fully self-supervised and is trained on paired ECG-CMR data from 27,951 subjects in the UK Biobank [21]. We evaluate PTACL on patient retrieval and phenotype regression, assessing how well ECG embeddings capture CMR-derived cardiac function. Our results demonstrate that local contrastive alignment improves ECG representations, enhancing their diagnostic utility for assessing cardiac structure and function.

2 Related Works

Self-supervised learning has advanced both single modality and multimodal representation learning. For single modality tasks, masked autoencoders (MAE) [7] have proven highly effective in learning meaningful representations from unlabeled data. They have been widely applied to signals such as ECG [4, 16, 19, 26, 30, 32], speech, and audio [8], as well as medical imaging, where 3D MAEs [6, 24] learn spatial-temporal features from volumetric scans [15, 22, 33].

In multimodal learning, most contrastive approaches are inspired by CLIP [17], which aligns per-patient representations from different modalities in a shared embedding space, pulling similar samples together while pushing apart unrelated ones. This framework has been applied in medical domains, such as ECG-CMR integration [5, 18, 25] and X-ray-text association [23, 27, 34], enabling zero-shot classification and improved diagnostic insights. However, these methods primarily focus on global alignment, treating each modality as a single representation and potentially overlooking fine-grained local interactions.

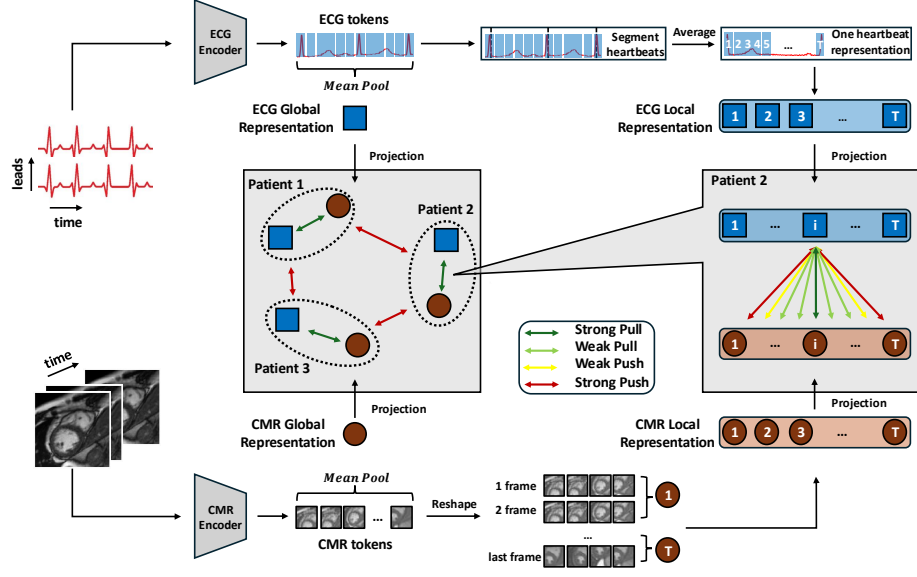


Fig. 1. Overview of PTACL. We train a multimodal model on pairs of ECG and CMR using a combination of a global contrastive loss (left square) and a local contrastive loss (right square). The global loss aligns patient-level ECG and CMR embeddings by pulling representations from the same patient closer together while pushing apart those from different patients. The local loss enforces fine-grained temporal alignment by contrasting ECG-derived segments—encoded from a single-heartbeat representation—with their corresponding CMR frames. In the figure, representations from 1 to T start at the end-diastole (ED) phase. Overlays in the tokens (CMR and ECG) are for visualization purposes only.

To address this, recent work has started exploring local contrastive learning for medical applications. For example, GLORIA [9] applies a global-local learning scheme to image-text pairs, while Seibold et al. [20] refine CLIP by contrasting local image features with radiology reports, and LoVT [14] aligns image subregions with text embeddings. These methods mirror broader trends in vision-language research [12, 29, 31].

3 Methods

Our framework consists of three phases: single modality pre-training (SMP) (Sec. 3.1), multimodal contrastive learning (Sec. 3.2), and inference. First, we pre-train separate ECG and CMR encoders using MAE [7] on unlabeled data. Then, we jointly train both encoders with substantial augmentations, applying a global contrastive loss for patient-level alignment and a local contrastive loss for fine-grained temporal alignment. The local loss contrasts ECG-derived time segments—fixed length segments from a single heartbeat representation of multi-

lead ECG—with corresponding encoded frames from CMR at the same cardiac phase. Finally, we evaluate the pre-trained ECG encoder on multiple downstream tasks.

3.1 Pre-training for ECG and CMR

To integrate ECG and CMR, each modality first learns meaningful representations independently through single modality pre-training (SMP). This step enables the encoders to capture modality-specific features before multimodal contrastive learning. Following previous works [8, 24, 32], we pre-train the ECG and CMR encoders independently on each modality, using the MAE framework [7] in both cases.

We embed the multi-lead ECG signals using a 1D convolution over the time dimension with kernel size p_t , transforming each lead into a sequence of non-overlapping time-based tokens. The convolution is applied on each lead independently, with weights shared across all leads. To learn temporal and lead structure, we add learnable positional embeddings before randomly masking a portion of tokens.

For CMR, we treat the 2D+T short-axis sequence as a multi-frame input and apply a 3D convolution with kernel size $[t, p, p]$, where t defines the temporal window and p controls the spatial patch size. This generates a sequence of spatio-temporal tokens, each representing a localized region of the heart across time. To learn spatial and temporal structure, we add learnable positional embeddings before masking.

In both cases, a transformer encoder processes only the visible tokens, while a lightweight transformer decoder reconstructs the masked tokens. The Mean Squared Error (MSE) loss is computed only on the masked tokens, comparing the reconstructed output to the original uncorrupted input.

3.2 Multimodal Contrastive Pre-Training

To learn robust cross-modal representations, we employ two contrastive losses: a global loss for patient-level alignment and a local loss for fine-grained temporal alignment. The global loss ensures that embeddings from the same patient are similar while pushing apart embeddings from different patients. The local loss further refines alignment by matching ECG-derived temporal embeddings with corresponding CMR frames. ECG local embeddings consist of T segments, each spanning p_t time steps from a single heartbeat representation of multi-lead ECG, while CMR embeddings represent encoded frames from the cardiac sequence.

Global Contrastive Loss: For a batch B , let z_E, z_C be the ECG and CMR embeddings, obtained via average pooling over token representations followed by separate projection layers. We enforce cross-modal alignment using the InfoNCE loss, following [17]. The loss is defined symmetrically for $ECG \rightarrow CMR$ and $CMR \rightarrow ECG$ directions. The $ECG \rightarrow CMR$ contrastive loss is:

$$\mathcal{L}_{\text{ECG} \rightarrow \text{CMR}} = \frac{1}{B} \sum_{i=1}^B -\log \frac{\exp\left(\cos\left(z_{\mathcal{E}}^{(i)}, z_{\mathcal{C}}^{(i)}\right) / \tau\right)}{\sum_{j=1}^B \exp\left(\cos\left(z_{\mathcal{E}}^{(i)}, z_{\mathcal{C}}^{(j)}\right) / \tau\right)}, \quad (1)$$

where τ is the temperature parameter, $z_{\mathcal{E}}$ and $z_{\mathcal{C}}$ are the projected embeddings, and $\cos(\cdot, \cdot)$ denotes cosine similarity. The final global contrastive loss $\mathcal{L}_{\text{global}}$ is a weighted combination of both directions.

Local Contrastive Loss: CMR acquisition is guided by ECG gating, where imaging data is continuously recorded over multiple heartbeats while simultaneously capturing the ECG signal. The detected R peaks are later used to align and reconstruct a single representative heartbeat cycle. This physiological alignment motivates our local contrastive loss formulation.

Local Embedding Formation: For CMR, each frame is divided into spatial tokens representing $p \times p$ image patches and temporal tokens capturing sequences of t frames. Local embeddings are obtained by averaging spatial tokens within each temporal token, producing $z_{\mathcal{C}}^{(i,t)}$, where t indexes the CMR frames. For ECG, each token represents a segment of length p_t from a single lead. To standardize heartbeat representations across patients, we extract tokens between consecutive R-peaks for each detected heartbeat. These variable-length sequences are interpolated to a fixed length of T tokens. Finally, we compute the mean across multiple heartbeats to obtain the ECG local embeddings for patient i , denoted as $z_{\mathcal{E}}^{(i,k)}$, where k indexes the ECG time segments.

Temporal Alignment and Local Loss: Following [10], we enforce temporal alignment between $z_{\mathcal{E}}^{(i,k)}$ (ECG) and $z_{\mathcal{C}}^{(i,t)}$ (CMR) local embeddings, using symmetrical alignment matrix $\mathcal{P}_{k,t}$:

$$\mathcal{P}_{k,t} \propto \begin{cases} \delta_{k,t} & \sigma = 0 \quad (\text{hard alignment}) \\ \exp\left(-\frac{1}{2}\left(\frac{\text{dist}_{k,t}}{\sigma}\right)^2\right) & \sigma > 0 \quad (\text{soft alignment}) \end{cases} \quad (2)$$

Here $\sum_{t=1}^T \mathcal{P}_{k,t} = 1$ and $\delta_{k,t}$ ensures exact matches for $\sigma = 0$, while for $\sigma > 0$, alignment is softened using a Gaussian weighting based on the normalized temporal distance $\text{dist}_{k,t}$ (Fig. 2).

The local contrastive loss is defined symmetrically for $\text{ECG} \rightarrow \text{CMR}$ and $\text{CMR} \rightarrow \text{ECG}$ directions. The local per-patient, per-segment loss for $\text{ECG} \rightarrow \text{CMR}$:

$$\ell_{\text{ECG} \rightarrow \text{CMR}}^{(i,k)} = -\sum_{t=1}^T \mathcal{P}_{k,t} \log \left[\frac{\exp\left(\cos\left(z_{\mathcal{E}}^{(i,k)}, z_{\mathcal{C}}^{(i,t)}\right) / \tau\right)}{\sum_{t'=1}^T \exp\left(\cos\left(z_{\mathcal{E}}^{(i,k)}, z_{\mathcal{C}}^{(i,t')}\right) / \tau\right)} \right]. \quad (3)$$

The total local contrastive loss for $\text{ECG} \rightarrow \text{CMR}$ is obtained by averaging $\ell_{\text{ECG} \rightarrow \text{CMR}}^{(i,k)}$ over all ECG segments k and patients in the batch. The final local contrastive loss $\mathcal{L}_{\text{local}}$ is a weighted combination of both directions.

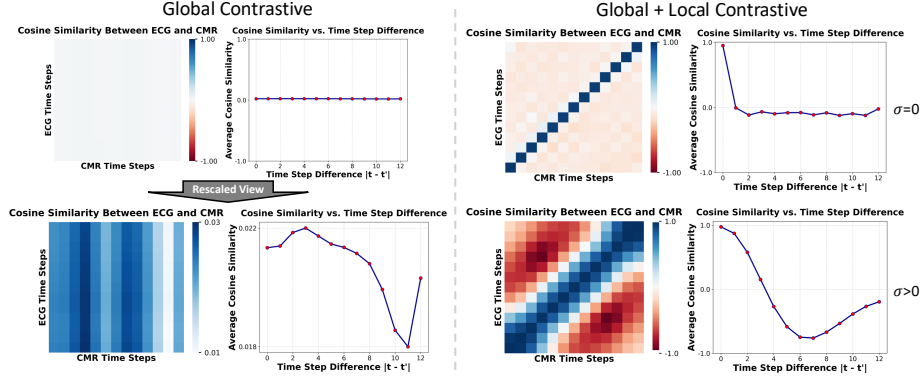


Fig. 2. Impact of local contrastive loss on temporal alignment between time steps from both modalities computed during inference. The left side, using only global contrastive learning, shows no clear temporal correspondence between modalities. The right side demonstrates the effect of adding local contrastive loss (PTACL), leading to improved alignment. For $\sigma = 0$, the model enforces strict one-to-one time step matching, while for $\sigma > 0$, Gaussian weighting enables smoother alignment. Our method achieves temporal alignment between ECG and CMR time steps, which is absent in the global-only approach.

Total Loss: The final objective combines both global and local contrastive losses to ensure alignment at both patient and temporal levels. The total loss is:

$$L = \mathcal{L}_{\text{global}} + \beta \cdot \mathcal{L}_{\text{local}} \quad (4)$$

where β controls the relative contribution of the local contrastive loss.

4 Experiments

We evaluate our model using paired 12-lead resting ECG and short-axis CMR data from the UK Biobank [21], with 27,951 patients for training and 5,991 for test. CMR sequences contain 50 frames, with the middle short-axis slice cropped around the cardiac center (84×84) and min-max normalized. ECGs are 10-second, 12-lead recordings sampled at 500 Hz. ECGs are preprocessed via an 8-level wavelet transform, following [13, 32], for baseline wander and noise removal and followed by z-score normalization. CMR augmentations include random resized cropping, time sampling, rotation, flipping, color jitter, Gaussian blurring, and noise, while ECG augmentations include Fourier surrogate transformations, random cropping (2500 points), jittering, and rescaling. We evaluate 10 CMR-derived cardiac phenotypes [2]: LV/RV end-diastolic and end-systolic volumes (EDV, ESV), stroke volume (SV), ejection fraction (EF), LV mass (M), and LV cardiac output (CO).

4.1 Retrieval of Similar Patients

To assess the clinical utility of our learned ECG embeddings, we test their ability to retrieve patients with similar CMR-derived phenotypes. For each query ECG, we rank all patients in the CMR embedding database using cosine similarity. A match is counted if the candidate’s phenotype is within $\pm 0.5\sigma$ of the query’s value, where σ is the standard deviation of that phenotype across test data. We compare two embedding strategies: a baseline model (global) and PTACL (global+local). While our global model already demonstrates strong retrieval performance, incorporating a local loss further improves precision and ranking. This enhancement is clinically significant as it enables more accurate patient stratification. Table 1 summarizes Precision@ k (P@ k), mean rank (MnR), and median rank (MdR) for key heart phenotypes. For reference, values in parentheses indicate performance when retrieving k random patients, reported only for $k = 1$ and $k = 10$ to save space.

Table 1. Retrieval results for CMR-derived heart phenotypes using ECG embeddings. We compare baseline (global $\triangle$) and PTACL (global+local $\square$) models, reporting performance as $\triangle/\square$ (random), with random retrieval in parentheses. A match is defined as a candidate whose phenotype differs from the query by no more than $\pm 0.5\sigma$. The inclusion of local loss in PTACL consistently improves patient retrieval with higher precision and lower rank scores across all phenotypes, demonstrating its potential to enhance clinical patient stratification.

	P@1 $\uparrow$	P@3 $\uparrow$	P@10 $\uparrow$	P@15 $\uparrow$	MnR $\downarrow$	MdR $\downarrow$
EDV _{LV}	.492 / .510 (.291)	.462 / .463	.426 / .427 (.287)	.416 / .415	3.1 / 2.9	2.0 / 1.0
ESV _{LV}	.501 / .524 (.303)	.469 / .480	.438 / .442 (.284)	.429 / .432	3.0 / 3.0	1.0 / 1.0
SV _{LV}	.458 / .472 (.294)	.417 / .423	.387 / .388 (.296)	.380 / .377	3.4 / 3.2	2.0 / 2.0
EF _{LV}	.431 / .436 (.278)	.392 / .396	.355 / .359 (.282)	.349 / .351	3.9 / 3.8	2.0 / 2.0
CO _{LV}	.439 / .453 (.295)	.407 / .417	.377 / .378 (.284)	.368 / .368	3.9 / 3.6	2.0 / 2.0
M _{LV}	.547 / .561 (.293)	.516 / .522	.479 / .486 (.294)	.471 / .474	2.6 / 2.6	1.0 / 1.0
EDV _{RV}	.500 / .516 (.280)	.467 / .477	.433 / .441 (.287)	.426 / .430	3.0 / 2.8	2.0 / 1.0
ESV _{RV}	.518 / .531 (.290)	.485 / .494	.455 / .460 (.281)	.446 / .451	2.8 / 2.6	1.0 / 1.0
SV _{RV}	.458 / .473 (.285)	.420 / .428	.386 / .390 (.289)	.380 / .382	3.8 / 3.7	2.0 / 2.0
EF _{RV}	.433 / .443 (.288)	.397 / .405	.370 / .373 (.290)	.364 / .365	4.3 / 4.2	2.0 / 2.0

4.2 Regression of Cardiac Phenotypes

We evaluate the regression performance of different pre-training strategies on various cardiac phenotypes using R^2 scores. Table 2 presents the results for models trained on ECG, CMR, or paired ECG-CMR data. CMR single modality pre-training (SMP) reveals the upper bound of imaging information, while the lower performance of ECG SMP highlights the challenges of using ECG

alone. Multimodal learning aims to reduce this gap. Among self-supervised approaches, MAE [7] consistently outperforms SimCLR [3] on both modalities, although fully supervised single modality training remains a robust baseline. Our Global model serves as a strong baseline, already outperforming previous works. Adding a local contrastive loss (PTACL) further improves performance across all cardiac phenotypes. Importantly, PTACL achieves these gains without introducing additional learnable parameters, making it a computationally efficient enhancement.

Table 2. Regression performance (R^2) for different pre-training methods. Models are evaluated with either linear probing (L) or full fine-tuning (F) using attention pooling (A). Our Global model outperforms previous multimodal approaches. Adding a local contrastive loss (PTACL) further improves results across all phenotypes without introducing additional learnable parameters. SMP stands for single modality pre-training.

Method	EDV _{LV}	ESV _{LV}	SV _{LV}	EF _{LV}	CO _{LV}	M _{LV}	EDV _{RV}	ESV _{RV}	SV _{RV}	EF _{RV}
CMR $R^2 \uparrow$										
Random Init ^L	.015	.017	.009	.008	.003	.021	.018	.020	.009	.007
SimCLR ^L [3]	.572	.612	.466	.513	.418	.628	.618	.681	.463	.520
MAE ^L [7]	.809	.772	.678	.489	.571	.824	.799	.784	.657	.492
Supervised ^F	.814	.789	.665	.470	.536	.823	.801	.808	.623	.473
ECG $R^2 \uparrow$										
Random Init ^{*,L}	-4.8	-3.8	-7.	-38.8	-7.6	18.3	.9	3.6	-9.6	.9
SimCLR ^L [3]	.278	.261	.196	.085	.156	.343	.308	.317	.205	.130
MAE ^L [7]	.439	.425	.332	.218	.252	.524	.466	.481	.333	.254
Supervised ^F	.458	.443	.333	.198	.252	.516	.481	.491	.340	.237
Multimodal (ECG+CMR) $R^2 \uparrow$										
ECCL ^{1,F} [5]	.372	.348	.270	.176	.212	.449	.397	.397	.281	.203
CMAE ^{2,L} [18]	.451	.380	.316	.103	.281	.536	.490	.445	.320	.129
MMCL ^{3,F,A} [25]	.498	.497	.360	.245	-	.597	.527	.534	.375	.248
PTACL ^{1,4,L}	.359	.334	.256	.114	.187	.406	.381	.386	.253	.149
PTACL ^{1,5,6,L}	.509	.496	.383	.252	.286	.614	.537	.553	.387	.295
Global ^{6,L}	.507	.499	.377	.258	.281	.612	.534	.553	.380	.297
PTACL^{1,6,L}	.514	.500	.386	.255	.289	.616	.540	.557	.389	.299
PTACL^{1,6,F,A}	.544	.537	.408	.283	.310	.645	.568	.582	.408	.315

* $\times 10^{-5}$, ¹ Single mid SAX slice, ² LAX + reconstruction loss, ³ Three SAX slices, ⁴ No SMP, ⁵ Local loss changed to maximising cosine similarity for matching time steps (no negatives), ⁶ With SMP, std ~ 0.001 – 0.002 , ^L Linear probing, ^F Full fine-tuning, ^A Attention pooling.

4.3 Implementation Details

Our approach employs transformer-based architectures for both modalities. The CMR model has 4 layers, 8 attention heads, a patch [2, 12, 12], and a 90% mask

ratio (13M parameters). The ECG model has 6 layers, 8 heads, a patch 50, and a 75% mask ratio (19M parameters). Both use a 1-layer, 4-head decoder. During multimodal pre-training, the first two layers are frozen, leaving 6.3M and 12.6M trainable parameters for CMR and ECG, respectively. The contrastive loss follows the $\mathcal{L}_{out}^{sup}$ variant from [10]. We use $T = 13$, $\beta = 1.0$. Regression is performed with $\sigma = 0.1$ and global/local temperatures of 0.1/1.0, retrieval uses $\sigma = 0.0$ with temperatures of 0.07/0.07. Linear probing is performed with ordinary least squares regression. Results are averaged over five random seeds.

5 Conclusion and Discussion

We introduced PTACL, a contrastive learning framework that enhances ECG representations through patient-level and time step alignment with CMR. Unlike prior methods focused solely on global alignment, PTACL captures finer temporal correspondences, leading to better retrieval and phenotype regression, all without adding extra learnable parameters. Our approach relies on only a single middle short-axis slice from the CMR, demonstrating that minimal imaging data can still significantly enrich ECG. However, PTACL relies on paired ECG-CMR data, and our evaluation is limited to a single cohort. Future work could explore generalization to external datasets, extend PTACL to other modalities (e.g., ultrasound), or reduce reliance on paired multimodal data.

Acknowledgments. This research has been conducted using the UK Biobank Resource under Application Number 87802. This work is funded by the European Research Council (ERC) project Deep4MI (884622). A. S. is funded via the EVUK program ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ansari, Y., Mourad, O., Qaraqe, K., Serpedin, E.: Deep learning for ECG arrhythmia detection and classification: An overview of progress for period 2017–2023. *Front. Physiol.* **14**, 1246746 (2023)
2. Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., et al.: A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.* **26**(10), 1654–1662 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *ICML*. pp. 1597–1607. PmLR (2020)
4. Chien, H.Y.S., Goh, H., Sandino, C.M., Cheng, J.Y.: MAEEG: Masked auto-encoder for EEG representation learning. In: *NeurIPS Workshop* (2022), <https://arxiv.org/abs/2211.02625>
5. Ding, Z., Hu, Y., Li, Z., Zhang, H., Wu, F., Xiang, Y., et al.: Cross-modality cardiac insight transfer: A contrastive learning approach to enrich ECG with CMR features. In: *MICCAI 2024*. vol. LNCS 15003, pp. 109–119. Springer (2024)

6. Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Koyejo, S., et al. (eds.) *Advances in NeurIPS*. vol. 35, pp. 35946–35958 (2022)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *CVPR*. pp. 16000–16009 (2022)
8. Huang, P.Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., et al.: Masked autoencoders that listen. In: *NeurIPS*. vol. 35, pp. 28708–28720 (2022)
9. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: *ICCV*. pp. 3942–3951 (2021)
10. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., et al.: Supervised contrastive learning. In: Larochelle, H., et al. (eds.) *Advances in NeurIPS*. vol. 33, pp. 18661–18673 (2020)
11. Leiner, T., Bogaert, J., Friedrich, M.G., Mohiaddin, R., et al.: SCMR position paper (2020) on clinical indications for cardiovascular magnetic resonance. *J. Cardiovasc. Magn. Reson.* **22**(1), 76 (2020)
12. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In: *ACM MM*. pp. 638–647. ACM (2022)
13. Martis, R.J., Acharya, U.R., Min, L.C.: ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. *Biomed. Signal Process. Control* **8**(5), 437–448 (2013)
14. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Joint learning of localized representations from medical images and reports. In: Avidan, S., et al. (eds.) *ECCV*. pp. 685–701. Springer (2022)
15. Munk, A., Ambsdorf, J., Llambias, S.N., Nielsen, M.: AMAES: Augmented masked autoencoder pretraining on public brain MRI data for 3D-native segmentation. *arXiv preprint arXiv:2408.00640* (2024)
16. Na, Y., Park, M., Tae, Y., Joo, S.: Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In: *ICLR* (2024)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: *ICML*. pp. 8748–8763. PmLR (2021)
18. Radhakrishnan, A., Friedman, S.F., Khurshid, S., Ng, K., Batra, P., Lubitz, S.A., et al.: Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nat. Commun.* **14**(1), 2436 (2023)
19. Sawano, S., Kodera, S., Takeuchi, H., Sukeda, I., Katsushika, S., et al.: Masked autoencoder-based self-supervised learning for electrocardiograms to detect left ventricular systolic dysfunction. In: *NeurIPS Workshop* (2022)
20. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: *MICCAI*. pp. 690–700. Springer (2022)
21. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al.: UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**(3), e1001779 (2015)
22. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., et al.: Self-supervised pre-training of Swin transformers for 3D medical image analysis. In: *CVPR*. pp. 20730–20740 (2022)
23. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**(12), 1399–1406 (2022)

24. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in NeurIPS* **35**, 10078–10093 (2022)
25. Turgut, Ö., Müller, P., Hager, P., Shit, S., Starck, S., Menten, M.J., et al.: Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging. *Med. Image Anal.* **101**, 103451 (2025)
26. Wang, G., Wang, Q., Iyer, G., Nag, A., John, D.: Unsupervised pre-training using masked autoencoders for ECG analysis. In: *BioCAS*. pp. 1–5 (2023)
27. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Goldberg, Y., et al. (eds.) *EMNLP*. p. 3876. *ACL* (2022)
28. World Health Organization: WHO Mortality Database (2025), <https://www.who.int/data/data-collection-tools/who-mortality-database>, accessed: Jan. 31, 2025
29. Yang, J., Bisk, Y., Gao, J.: TACo: Token-Aware cascade contrastive learning for video-text alignment. In: *ICCV*. pp. 11562–11572 (2021)
30. Yang, S., Lian, C., Zeng, Z., Xu, B., Su, Y., et al.: Masked self-supervised ECG representation learning via multiview information bottleneck. *Neural Comput. Appl.* **36**(14), 7625–7637 (2024)
31. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., et al.: FILIP: Fine-grained interactive language-image pre-training. In: *ICLR* (2022)
32. Zhang, H., Liu, W., Shi, J., Chang, S., Wang, H., He, J., et al.: MaeFE: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Trans. Instrum. Meas.* **72**, 1–15 (2023)
33. Zhang, Y., Chen, C., Shit, S., Starck, S., Rueckert, D., Pan, J.: Whole heart 3D+T representation learning through sparse 2D cardiac MR images. In: *MICCAI*. vol. 15001, pp. 359–369. Springer, Cham (2024)
34. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., et al.: Contrastive learning of medical visual representations from paired images and text. In: Lipton, Z., et al. (eds.) *MLHC*. vol. 182, pp. 2–25. PMLR (2022)