CardiacCLIP: Video-based CLIP Adaptation for LVEF Prediction in a Few-shot Manner

Yao DU[1], Jiarong GUO[1], and Xiaomeng LI[1]*

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
eexmli@ust.hk

Abstract. Echocardiography is a vital non-invasive modality for cardiac assessment, with left ventricular ejection fraction (LVEF) serving as a key indicator of heart function. Existing LVEF estimation methods depend on large-scale annotated video datasets, which are costly and limit adaptability across various clinical settings. Recent vision-language models for echocardiography, such as EchoCLIP, apply image-to-text pretraining but fail to capture crucial temporal dynamics and localized cardiac structures essential for accurate diagnosis. To address these challenges, we propose **CardiacCLIP**, a video-based framework that enhances LVEF prediction through attention-based frame aggregation and multi-resolution input scaling. Specifically, we introduce MFL (Multi Frame Learning), a novel attention-based mechanism for selectively fusing informative frames, and EchoZoom, a multi-scale feature extraction strategy that refines spatial representations of cardiac structures. As a novel adaptation of CLIP models for few-shot echocardiogram video analysis, our approach significantly improves diagnostic accuracy, reducing MAE by 2.07 on the EchoNet-Dynamic dataset under 1-shot setting. The code is available at https://github.com/xmed-lab/CardiacCLIP.

Keywords: Vision Language Model · Echocardiogram · Ejection Fraction.

1 Introduction

Left ventricular ejection fraction (LVEF) is a fundamental measure of cardiac function, widely used for diagnosing and monitoring cardiac conditions such as heart failure and cardiomyopathy [8,14,20]. Echocardiography, as a non-invasive and cost-effective imaging modality, is the primary tool for assessing LVEF in clinical practice [18,31,32]. However, estimating LVEF remains a challenging task due to its dependence on expert interpretation, inter-operator variability, and the complex temporal dynamics of cardiac motion [21,33]. Manual assessment is time-consuming and prone to subjectivity, highlighting the need for automated solutions that can improve efficiency and accuracy in LVEF estimation [7,19].

* Corresponding Author

Recent deep learning approaches have shown promise in automating LVEF prediction by leveraging large-scale echocardiographic video datasets [6,19]. These models typically rely on supervised learning paradigms that require massive labeled videos, making them highly dependent on extensive manual annotations and creating significant bottlenecks for model scalability and adaptability across different clinical settings. Furthermore, domain shifts due to variations in acquisition protocols and ultrasound manufacturers often degrade model performance when deployed in real-world clinical environments [2,7,30]. To address these challenges, it is crucial to develop data-efficient adaptation strategies that can generalize across diverse conditions with minimal labeled supervision [5,16].

Vision-language models (VLMs), particularly CLIP-based architectures, have recently emerged as powerful tools in medical image analysis [3,4,9]. By aligning visual and textual representations, CLIP enables models to learn rich semantic features from large-scale image-text pairs without requiring extensive task-specific annotations [23]. EchoCLIP [4] represents the first attempt to apply CLIP to echocardiography, achieving promising results in image-text retrieval. However, EchoCLIP [4] extracts a random frame from each video for image-text matching pretraining, and averages predictions across frames without modeling the temporal dynamics of the cardiac cycle. In addition, the diagnosis of various cardiac diseases and important indicators heavily rely on the accurate perception of the temporal changes in specific regions of the heart [11,14]. However, CLIP-based models are known to have limited fine-grained feature understanding, making them less effective in identifying subtle cardiac abnormalities. Since LVEF prediction is inherently a video-based task, existing CLIP models fail to fully exploit the temporal and anatomical information embedded in echocardiographic sequences.

To overcome these limitations, we propose **CardiacCLIP**, a novel video-based CLIP adaptation designed for few-shot LVEF prediction from echocardiogram videos. Our method introduces two key components: MFL (Multi Frame Learning), an attention-based frame aggregation module that selectively fuses frame-level information for temporal modeling; and EchoZoom, a multi-resolution input scaling strategy tailored for capturing fine-grained anatomical features from the apical four-chamber views. MFL mitigates the redundancy of frame-wise features by learning an optimal weighting mechanism, while EchoZoom ensures that the model attends to diagnostically relevant cardiac regions by fusing multi-scale representations. These two components enhance model robustness and generalization, allowing CardiacCLIP to achieve superior performance in data-limited settings. Our contributions are summarized as follows:

1. We introduce CardiacCLIP, a novel CLIP-based framework specifically designed for video-based echocardiography analysis, addressing the limitations of image-based CLIP models.
2. We develop MFL, an attention-based frame fusion mechanism that effectively captures temporal dependencies in LVEF estimation.
3. We propose EchoZoom, a multi-resolution scaling strategy that enhances the model's ability to capture fine-grained structural details.

4. We demonstrate that CardiacCLIP significantly outperforms existing methods in few-shot settings, achieving state-of-the-art performance.

2 CardiacCLIP for LVEF Prediction

2.1 Preliminary: Coarse-to-Fine Ordinal Regression

LVEF estimation can be reformulated as an ordinal regression problem, where we first convert it as a classification task by discretizing the labels as different bins and treating each bin as an independent class, and then regress the specific values based on the classification results [9,26]. The motivation for this is based on the fact that learning from a staged classification process is more effective and easier than directly learning from multiple precise values, especially in the imperfect data scenario [22]. This reformulation allows training with cross-entropy loss while maintaining numerical continuity via an MAE-based regression refinement [10,29]:

$$L_{OR} = L_{CE} + L_{MAE} \quad . \tag{1}$$

For the coarse-to-fine framework, the coarse stage maps LVEF values into discrete bins, leveraging CLIP's pretrained visual-textual alignment. This transforms the problem into a classification task, where text embeddings serve as classifier weights. The fine stage refines predictions via a lightweight MLP regressor, making the final estimation:

$$y^* = \sum_{i=1:k} p_i * \frac{b_i}{1+\delta_i} \quad , \tag{2}$$

where k is the number of classes, p_i is the class probability, b_i is the centre of i_{th} mapped numerical group, and δ_i is the estimated shift from the regressor to make the bin interval learnable. These two stages are trained end-to-end.

2.2 Video-based LVEF Prediction

Leveraging its robust representation capabilities from pretraining on extensive image-text pairs, CLIP serves as a foundational model for various downstream tasks, including video recognition [24]. Given an echocardiogram video $x_i \in \mathbb{R}^{T\times C\times H\times W}$ with T frames and each frame is with the spatial dimension of $H \times W$, we process the video through the CLIP visual encoder $f_v(\cdot)$ to extract features:

$$z^{v_i} = f_v(x_i) \quad , \tag{3}$$

where the visual feature $z^{v_i} \in \mathbb{R}^{T\times C}$ and C represents the feature dimension of the [CLS] token. Unlike previous CLIP methods for video adaptation that average frame-level features to obtain video representations [12,24], we introduce MFL (Multi Frame Learning), attention-based feature aggregation to capture critical cardiac dynamics.

For text features, we tokenize clinically relevant LVEF descriptions (e.g., *"The left ventricular ejection fraction is estimated to be mildly reduced LVEF (45-54%)"*) and embed them via the CLIP text encoder $f_t(\cdot)$:

$$z^{t_j} = f_t(t_j) \quad , \tag{4}$$

where $z^{t_j} \in \mathbb{R}^C$. To enrich text representations, we leverage GPT-4 [1] to generate diverse descriptions corresponding to LVEF intervals, enhancing data efficiency and serving as a form of text data augmentation during training. Given the ground-truth category label y_i, the model is trained via a cross-entropy loss:

$$L_{CE} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{K}\hat{y}_{i,j}\log y_{i,j} \quad , \tag{5}$$

where K denotes the total number of classes. Thus for video-based adaptation, the L_{CE} in Loss 1 should be updated with Loss 5.

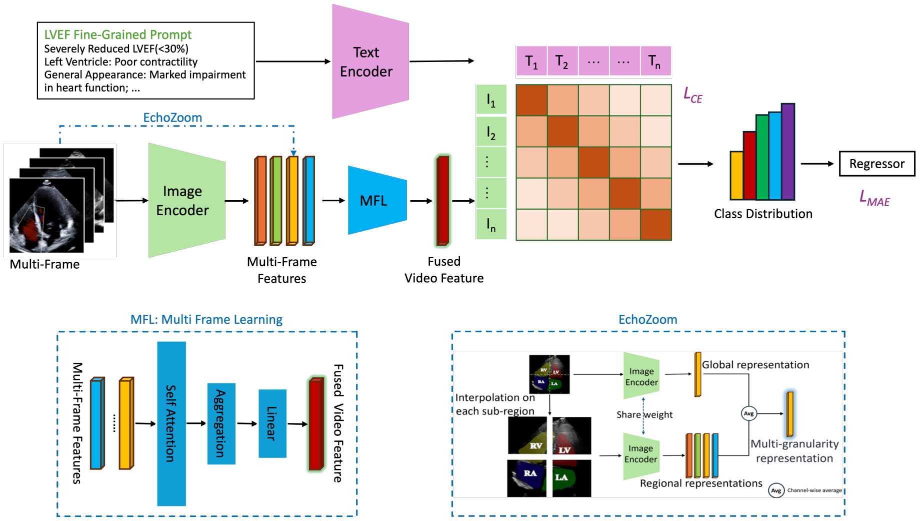

Fig. 1. CardiacCLIP: Video-based CLIP adaptation for few-shot LVEF prediction integrating multi frame learning and multi-scale representations.

2.3 Video Representation via Multi Frame Learning

Typical CLIP-based video recognition models aggregate frame features via simple averaging [12,24], overlooking temporal variability. In LVEF estimation, different frames contribute unequally due to varying cardiac contraction phases. Inspired by Multiple Instance Learning (MIL) [15,27,28], we introduce MFL, an attention-based fusion mechanism that prioritizes diagnostically relevant frames.

Given an input video sequence with B frames, we extract a set of frame-level features:

$$F = [f_1, f_2, \ldots, f_B] \in \mathbb{R}^{B\times C} \quad , \tag{6}$$

where $f_i \in \mathbb{R}^C$ represents the feature vector of the i-th frame, and C is the feature dimension. Instead of using average pooling, we introduce an attention mechanism to learn the relative importance of each frame dynamically.

Frame Importance Estimation. we compute per-frame importance scores using a multi-layer attention network:

$$s_i = W_3 \tanh(W_2 \tanh(W_1 f_i)) \quad , \tag{7}$$

where W_1, W_2, W_3 are learnable weight matrices. These scores are normalized using softmax to ensure that the sum of the weights is equal to 1:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{B} \exp(s_j)} \quad , \tag{8}$$

where α_i represents the learned weight for frame i, ensuring higher weights for informative frames while suppressing redundancy.

Dynamic Feature Aggregation. The aggregated video representation is computed as:

$$F_{\text{agg}} = \sum_{i=1}^{B} \alpha_i f_i \quad , \tag{9}$$

where $F_{\text{agg}} \in \mathbb{R}^C$ is the aggregated feature vector representing the entire video. This formulation ensures that the model prioritizes the most relevant frames. Finally, the aggregated feature vector is passed through a linear projection layer:

$$F_{\text{final}} = W_{\text{proj}} F_{\text{agg}} \quad , \tag{10}$$

where $W_{\text{proj}} \in \mathbb{R}^{C \times C}$ is a learnable projection matrix that refines the video representation.

By integrating MIL-inspired attention-based aggregation, our model learns to emphasize diagnostically-relevant frames, offering a more robust, adaptive, and interpretable approach to video-based LVEF prediction compared to the conventional average pooling strategy. In practice, we find that the input frame length plays a crucial role in the performance of feature aggregation and we discuss it in the ablation study.

2.4 EchoZoom: Multi-Scale Cardiac Representation

Echocardiographic diagnosis relies on analyzing regional cardiac dynamics, particularly within the left and right ventricles and atria [7,11]. Standard vision models process images at a fixed resolution, which limits the model's ability to capture multi-scale anatomical variations. EchoZoom enhances regional cardiac representation by applying multi-resolution input scaling. As shown in the lower right corner of Figure 1, it processes images at multiple scales (e.g., 112^2, 224^2), enabling fine-grained structural analysis. Specifically, EchoZoom splits the 224^2 image into four 112^2 sub-images. These sub-images, along with the original 112^2 image, are fed through the same pretrained model. The features extracted from these sub-images are then combined into a larger feature map corresponding to the 112^2 image. This map is subsequently average-pooled to match the feature map size of the original 112^2 image. The final output is the fused feature map generated across all scales. This process enriches feature extraction without requiring additional parameters, reinforcing the model's ability to recognize subtle morphological changes across varying resolutions.

3 Experiments

3.1 Datasets and Experiment Settings

Dataset. We evaluate our method on EchoNet-Dynamic [18], a widely used benchmark dataset in echocardiography. It contains 10,036 apical four-chamber echocardiogram videos collected from Stanford University Hospital using five different ultrasound machines. Each video, averaging 175 frames, is resized to 112×112 and annotated with its corresponding LVEF label. The dataset is pre-split into 7,465 training, 1,288 validation, and 1,277 test samples. For few-shot evaluation, we extract subsets from the training set following the 1/2/4/8-shot settings (Table 1).

Experiment Settings. Following EchoCLIP [4], we adopt ConvNext-Base CLIP as the backbone for fair comparison, and the model's pretraining dataset does not overlap with our current dataset. Our model is optimized using RAdam [17] for 100 epochs, starting with a learning rate of 5e-5, which is cosine-decayed to zero. Each input clip consists of 48 frames, sampled at a stride of 2, with a batch size of 2. To construct a typical few-shot dataset, we discretize LVEF values into integer classes (1-100) and sample training examples accordingly, skipping any missing classes, as detailed in Table 1. We adopt Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as evaluation metrics to assess the model performance.

Table 1. Sample counts under few-shot settings for EchoNet-Dynamic dataset.

Dataset	1-shot	2-shot	4-shot	8-shot
EchoNet-Dynamic	84	162	307	570

3.2 Results under Few-shot Setting

We mainly compare CardiacCLIP against two categories of methods: 1) Traditional LVEF prediction methods (video-based models trained end-to-end); 2) CLIP-based methods (pretrained VLMs adapted to echocardiography).

Table 2 presents the results. CardiacCLIP consistently outperforms existing methods, achieving a 2.07 MAE reduction over EchoNet [19] in the 1-shot setting. Similar performance gains can be observed across other shot settings, highlighting the effectiveness of our method. The performance improvement diminishes as training data increases, a typical phenomenon in few-shot learning.

Table 2. Comparison of different SOTA methods on EchoNet-Dynamic dataset under few-shot setting. (Num) indicates the performance improvement compared to EchoNet [19].

Method	MAE ↓				RMSE ↓			
	1-shot	2-shot	4-shot	8-shot	1-shot	2-shot	4-shot	8-shot
Traditional								
EchoNet [19]	9.32	9.17	7.49	6.81	12.11	11.89	9.92	9.19
AdaCon [6]	9.52	8.83	7.35	7.02	12.17	11.56	9.91	9.46
C-Mixup [34]	9.23	9.22	9.13	7.59	12.23	12.14	11.87	9.77
BalancedMSE [25]	8.50	7.87	7.55	7.05	11.22	10.18	9.65	9.35
CLIP-based								
EchoCLIP [4]	10.54	9.74	9.22	9.12	13.18	12.18	11.53	11.40
NumCLIP [9]	7.91	7.56	7.68	6.96	9.89	9.45	9.60	8.70
Our Method								
CardiacCLIP	7.25	7.11	6.79	6.42	9.06	8.89	8.49	8.02
Δ	(2.07)	(2.06)	(0.70)	(0.39)	(3.05)	(3.00)	(1.43)	(1.17)

3.3 Ablation Study

We conduct detailed ablation experiments to analyze the contributions of model components, input frame length, loss functions, and aggregation methods, under 1-shot setting.

Effect of Model Components. Table 3 shows the impact of EchoZoom and MFL, demonstrating that both modules contribute to enhanced model performance, with their joint combination achieving the best result.

Table 3. Ablation study of CardiacCLIP on EchoNet-Dynamic dataset.

Ablation Study	EchoZoom	MFL	MAE ↓
Base	✗	✗	7.91
w/o EchoZoom	✗	✓	7.42
w/o MFL	✓	✗	7.50
Ours	✓	✓	**7.25**

Effect of Frame Length. Table 4 presents the impact of input frame length on model performance. While shorter frame length (e.g., 16 frames) result in higher MAE (7.89), increasing the frame length initially improves performance, with the best MAE achieved at 48 frames (7.25). Beyond this, performance fluctuates slightly, indicating that longer sequences do not necessarily enhance feature extraction, likely due to increased redundancy in the input.

Effect of MFL Modules. Table 5 examines various design choices within MFL. Our proposed MFL achieves an MAE of 7.25, while removing the final projector increases the error to 7.53, highlighting the importance of feature transformation. Introducing a nonlinear projector yields a slight performance drop to 7.45, likely because the aggregated features are already well-structured. Incorporating a gated recurrent unit (GRU) further degrades performance, increasing the MAE to 8.26, suggesting excessive temporal dependencies may lead to overfitting.

Table 4. Ablation on frame length.

Frame Length	MAE ↓
16	7.89
36	7.64
48	**7.25**
54	7.72
64	7.38
96	7.92
128	7.94

Table 5. Ablation on MFL modules.

Aggregation	MAE ↓
MFL	**7.25**
w/o Projector	7.53
w/ Nonlinear Projector	7.45
w/ GRU	8.26

Table 6. Ablation on regression loss.

Regression Loss	MAE ↓
MAE	**7.25**
SmoothL1	7.36
Huber [13]	7.57
MSE	7.75

Table 7. Ablation on aggregation methods.

Aggregation	MAE ↓
MFL	**7.25**
Multi-Head	10.47
Multi-Head+GRU	8.3

Effect of Regression Loss. Table 6 investigates how different regression loss functions impact model performance. The standard MAE loss achieves the lowest error (7.25), while SmoothL1 (7.36) and Huber (7.57) introduce slight performance degradation. MSE loss performs the worst (7.75), likely due to its sensitivity to large errors, which may disproportionately penalize outliers.

Effect of Aggregation Methods. Table 7 evaluates the impact of different aggregation methods. Our MFL achieves the best performance, whereas Multi-Head Attention significantly degrades accuracy, increasing the MAE to 10.47, likely due to feature distortion caused by excessive attention heads. In video recognition tasks like LVEF prediction, only a subset of key frames holds critical diagnostic information. Introducing a GRU into the Multi-Head approach improves performance to 8.3, suggesting that temporal modeling can partially counteract attention-related issues. These findings are consistent with the observations in Table 5.

4 Conclusion

In this paper, we introduce **CardiacCLIP**, a novel framework for LVEF estimation from echocardiogram videos, extending CLIP-based models to effectively capture both spatial and temporal cardiac features. Our method addresses the limitations of prior approaches by incorporating Multi-frame Learning (MFL) for adaptive temporal feature aggregation and EchoZoom, a multi-scale input strategy that enhances the representation of key anatomical structures. Through a few-shot learning paradigm, CardiacCLIP demonstrates strong generalization with limited labeled data, making it well-suited for clinical applications. Extensive experiments on the EchoNet-Dynamic dataset validate the effectiveness of our method, achieving state-of-the-art performance in few-shot settings. These results highlight the potential of CardiacCLIP as a robust and data-efficient solution for automated echocardiographic analysis, paving the way for improved cardiac disease diagnosis in real-world scenarios.

Acknowledgments. This work was supported by a research grant from the Joint Research Scheme (JRS) under The National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24) as well as a grant from the National Natural Science Foundation of China (Grant No. 62306254).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Chao, C.J., Gu, Y.R., Xiang, T., Appari, L., Wu, J., Farina, J.M., Wraith, R., Jeong, J., Arsanjani, R., Kane, G.C., et al.: Comparative eminence: Foundation versus domain-specific model for cardiac ultrasound segmentation. medRxiv pp. 2023–09 (2023)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
4. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine pp. 1–8 (2024)
5. Dai, W., Du, Y., Bai, H., Cheng, K.T., Li, X.: Semi-supervised contrastive learning for deep regression with ordinal rankings from spectral seriation. NeurIPS **36**, 57087–57098 (2023)
6. Dai, W., Li, X., Chiu, W.H.K., Kuo, M.D., Cheng, K.T.: Adaptive contrast for image regression in computer-aided disease assessment. TMI **41**(5), 1255–1268 (2021)
7. Dai, W., Li, X., Ding, X., Cheng, K.T.: Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. TMI **42**(5), 1446–1461 (2022)
8. Douglas, P.S., Garcia, M.J., Haines, D.E., Lai, W.W., Manning, W.J., Patel, A.R., Picard, M.H., Polk, D.M., Ragosta, M., Ward, R.P., et al.: 2011 appropriate use criteria for echocardiography: a report of the american college of cardiology foundation appropriate use criteria task force. Journal of the American College of Cardiology **57**(9), 1126–1166 (2011)

9. Du, Y., Zhai, Q., Dai, W., Li, X.: Teach clip to develop a number sense for ordinal regression. In: ECCV. pp. 1–17. Springer (2024)
10. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR. pp. 2002–2011 (2018)
11. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. NPJ digital medicine **3**(1), 10 (2020)
12. Huang, X., Zhou, H., Yao, K., Han, K.: Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. arXiv preprint arXiv:2402.03241 (2024)
13. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution, pp. 492–518. Springer (1992)
14. Hughes, J.W., Yuan, N., He, B., Ouyang, J., Ebinger, J., Botting, P., Lee, J., Theurer, J., Tooley, J.E., Nieman, K., et al.: Deep learning evaluation of biomarkers from echocardiogram videos. EBioMedicine **73** (2021)
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136. PMLR (2018)
16. Li, Y., Zhang, Q., Xiang, T., Lin, Y., Zhang, Q., Li, X.: Few-shot lymph node metastasis classification meets high performance on whole slide images via the informative non-parametric classifier. In: MICCAI. pp. 109–119. Springer (2024)
17. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 (2019)
18. Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., Ashley, E.A., Liang, D.H., Zou, J.Y.: Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada (2019)
19. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
20. Papolos, A., Narula, J., Bavishi, C., Chaudhry, F.A., Sengupta, P.P.: Us hospital use of echocardiography: insights from the nationwide inpatient sample. Journal of the American College of Cardiology **67**(5), 502–511 (2016)
21. Pellikka, P.A., She, L., Holly, T.A., Lin, G., Varadarajan, P., Pai, R.G., Bonow, R.O., Pohost, G.M., Panza, J.A., Berman, D.S., et al.: Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. JAMA network open **1**(4), e181456–e181456 (2018)
22. Pintea, S.L., Lin, Y., Dijkstra, J., van Gemert, J.C.: A step towards understanding why classification helps regression. In: ICCV. pp. 19972–19981 (2023)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
24. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: CVPR. pp. 6545–6554 (2023)
25. Ren, J., Zhang, M., Yu, C., Liu, Z.: Balanced mse for imbalanced visual regression. In: CVPR. pp. 7926–7935 (2022)
26. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: ICCVW. pp. 10–15 (2015)
27. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NeurIPS **34**, 2136–2147 (2021)

28. Sudharshan, P., Petitjean, C., Spanhol, F., Oliveira, L.E., Heutte, L., Honeine, P.: Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications **117**, 103–111 (2019)
29. Wang, C., Song, Q., Zhang, B., Wang, Y., Tai, Y., Hu, X., Wang, C., Li, J., Ma, J., Wu, Y.: Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In: ICCV. pp. 3234–3242 (2021)
30. Yan, W., Huang, L., Xia, L., Gu, S., Yan, F., Wang, Y., Tao, Q.: Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. Radiology: Artificial Intelligence **2**(4), e190195 (2020)
31. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: ICCV. pp. 11878–11887 (2023)
32. Yang, J., Lin, Y., Pu, B., Guo, J., Xu, X., Li, X.: Cardiacnet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In: ECCV. pp. 293–311. Springer (2024)
33. Yang, J., Lin, Y., Pu, B., Li, X.: Bidirectional recurrence for cardiac motion tracking with gaussian process latent coding. NeurIPS **37**, 34800–34823 (2024)
34. Yao, H., Wang, Y., Zhang, L., Zou, J.Y., Finn, C.: C-mixup: Improving generalization in regression. NeurIPS **35**, 3361–3376 (2022)