Learning to Unify Deformable Shape and Texture Representations for Cardiac Video Classification

Tonmoy Hossain[1] and Miaomiao Zhang[1,2]

[1] Department of Computer Science, University of Virginia, Virginia, USA
[2] Department of Electrical and Computer Engineering, University of Virginia, USA

Abstract. Deformable shape representations have proven to be robust complements to texture features in cardiac image classification, offering geometric priors that are invariant to imaging artifacts and intensity variations. However, existing deep networks perform simple concatenation to combine these distinct feature representations, which neither fully exploits their complementary nature nor learns cross-modal feature dependencies. Furthermore, this results in uniform attention across all timepoints; hence ignoring the varying diagnostic importance across the cardiac phases. In this paper, we propose a novel cardiac video classification model that, for the first time, *learns temporal features in an integrated space of deformable shape and image texture representations*. In particular, we design a bi-directional cross-attention in the latent space to fuse latent deformable shape and image features, allowing each modality to adaptively weight the other based on spatio-temporal correspondence. In contrast to current methods that apply uniform weighting across all the cardiac phases, our approach learns to dynamically adjust the contributions of shape and texture representations, derived from images, over time. We demonstrate state-of-the-art classification performance on a cine cardiac magnetic resonance (CMR) video dataset, achieving improved interpretability from attention mechanisms that identify diagnostically critical cardiac phases and modality contributions. Our code is publicly available at `https://github.com/tonmoy-hossain/ShapeFuse`.

Keywords: CMR Video Classification · Feature Fusion · Deformations.

1 Introduction

Cardiac MRI enables non-invasive assessment of myocardial function and is widely used for disease diagnosis, yet automated classification from CMR videos remains challenging [8,17,18,28]. Recent models capture temporal dependencies across frames but solely operate on raw image intensities. As a result, they mostly focus on texture-based features without understanding the underlying myocardial deformations [2,10,31]. However, pathological conditions, such as cardiomyopathy and myocarditis manifest as regional wall motion abnormalities, asymmetric contractility, and altered strain patterns [11,12]. Therefore, geometric shape features captured in motion or deformation fields often encode richer information that can complement intensity-based representations alone [21,29,30].

Recent works have explored integrating deformable shape representations with intensity features for neurodegenerative disease classification, where geometric priors have proven to be robust complements to texture-based representations [14,15,24]. Extending this paradigm to cardiac video analysis, however, is not straightforward given the temporal nature of the cardiac cycle. A critical challenge lies in how to combine these two intertwined representations, as simple operations of element-wise addition or concatenation can ignore cross-modal feature dependencies. In particular, such approaches do not model how shape and texture jointly interact across the cardiac cycle, which limits their ability to learn complementary physiological information. For instance, deformation features are most informative during dynamic transition phases of the cardiac cycle, particularly from end-diastole through peak systolic contraction and early relaxation, where myocardial motion changes are most evident. In contrast, texture-based intensity patterns are less sensitive to these functional dynamics and may become unreliable under conditions such as low contrast-to-noise ratio or motion-induced imaging artifacts [1]. Despite the clear clinical relevance, prior works have not explicitly learned to model the dynamic interaction between myocardial deformation and image texture for disease classification.

Building on the premise of integrating deformable shape and texture representations, this paper presents `ShapeFuse`, the first framework to jointly learn the fusion of deformable shape and texture representations in a shared latent space for cardiac video classification. In contrast to existing methods that rely on static concatenation or addition, `ShapeFuse` explicitly learns cross-modal dependencies and dynamically weights modality contributions across the cardiac cycle. To achieve this, we propose a three-fold contribution:

- Develop a bidirectional cross-modal temporal attention mechanism that learns mutual dependencies between latent deformable shape and texture representations across the cardiac cycle.
- Design an adaptive fusion gate that learns to weight shape and texture contributions at each cardiac phase, coupled with a learned temporal pooling that prioritizes diagnostically relevant timepoints for classification.
- Demonstrate SOTA cardiac video classification on real-world cine CMR videos, with interpretable attention maps that localize diagnostically critical cardiac phases and quantify modality-wise contributions.

2 Background: Deformation Learning from Image Pairs

Deformation learning from image pairs aims to establish spatial correspondence between a source image $\mathcal{S}$ and a target image $\mathcal{T}$ by estimating a transformation that aligns $\mathcal{S}$ to $\mathcal{T}$. By following the principles of diffeomorphic registration, the transformation ϕ can be derived by mapping from $\mathcal{S}$ to $\mathcal{T}$ through integrating a stationary velocity field (SVF [3]) v via the flow equation

$$\frac{d\phi(t)}{dt} = v \circ \phi(t), \quad \phi(0) = \mathrm{Id}, \tag{1}$$

where v remains constant over time, and Id denotes an identity transformation. The solution at $t = 1$ is denoted by $\phi = \exp(v)$, computed via scaling-and-squaring as $\phi \approx (\mathrm{Id}+v/2^K)^{2^K}$. While we adopt the SVF formulation in this work, our proposed framework generalizes to other diffeomorphic parameterizations such as time-varying velocity fields. Following this transformation to deform $\mathcal{S}$ onto $\mathcal{T}$, we minimize the energy function

$$E = \frac{1}{\sigma^2}\mathcal{D}[\mathcal{S} \circ \phi(v), \mathcal{T}] + \|\nabla v\|_1, \quad \text{s.t. Eq. (1)}, \tag{2}$$

where $\mathcal{D}[\cdot,\cdot]$ measures image dissimilarity between the warped source $\mathcal{S} \circ \phi(v)$ and target $\mathcal{T}$, σ^2 represents noise variance, and the regularization term $\|\nabla v\|_1$ enforces smoothness of the velocity field. For the dissimilarity metric $\mathcal{D}$, we can employ sum of squared differences (SSD) [7], normalized cross-correlation (NCC) [4], or mutual information (MI) [26].

3 Our Method: ShapeFuse

This section introduces ShapeFuse, a novel framework that unifies representations of deformable shape and image intensity for improved cardiac video classification. At the core of ShapeFuse is a *latent shape-aware feature fusion* module that consists of two key components: (i) a *bidirectional cross-modal temporal attention* mechanism to enable dynamic interactions between shape and image features over time, and (ii) an *adaptive gating and diagnostic importance pooling* strategy to selectively emphasize the most clinically informative features and temporal phases for classification. An overview of the proposed architecture is illustrated in Fig. 1.

Problem Definition. Given a CMR dataset, $\mathcal{I} = \{I_0, I_1, \ldots, I_T\}$ with $T+1$ frames, our objective is to predict a disease label $y \in \mathcal{Y}$. To leverage geometric priors that complement texture-based representations, we estimate frame-wise velocity fields $\mathbf{S}$ from an encoder-decoder network (θ_S^E, θ_S^D), and derive texture representations $\mathbf{X}$ from an image encoder θ_I^E. *The core challenge is to unify these two distinct yet intertwined feature representations in the latent space such that their cross-modal dependencies and temporal relationships across cardiac phases are learned jointly, rather than treating them as independent modalities.*

3.1 Latent Shape-aware Feature Fusion

Given the input sequence $\mathcal{I}$, we estimate the latent geometric deformation of a target frame I_t at each timepoint t relative to a source frame I_0 as $\mathbf{z}_{v_t} \in \mathbb{R}^{C\times H\times W}$, inferred from the shape encoder θ_S^E. In parallel, we encode each frame I_t through θ_I^E to obtain its latent texture representation $\mathbf{z}_{f_t}$ in a shared latent space, enabling cross-modal interaction with the latent shape features.

Bidirectional Cross-Modal Temporal Attention. While SOTA models extensively adopt element-wise summation or concatenation as the fusion strategy, neither exploits the complementary nature of the two modalities nor learns

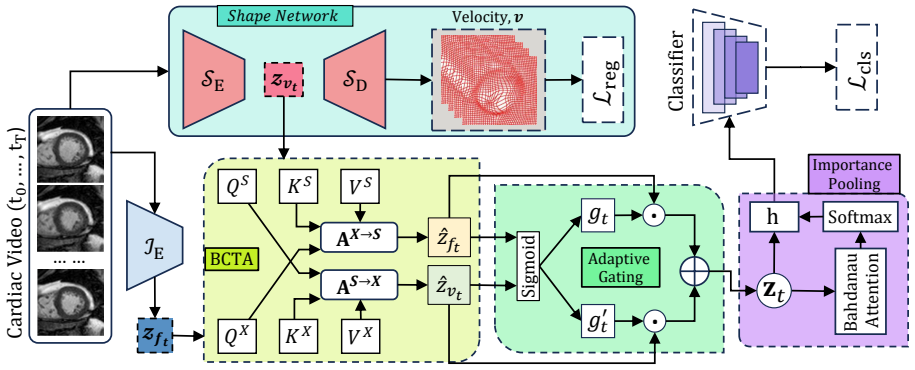

Fig. 1. An overview of our cardiac video classification model.

cross-modal feature dependencies. As a result, these approaches impose uniform weighting across all timepoints, ignoring the varying diagnostic importance across cardiac phases. Each latent map is first flattened and linearly projected to a token of dimension d, so that $\mathbf{z}_{v_t}, \mathbf{z}_{f_t} \in \mathbb{R}^d$. To this end, we formulate a bidirectional cross-modal temporal attention over the shape sequence $\mathbf{S} = [\mathbf{z}_{v_0}, \ldots, \mathbf{z}_{v_T}]$ and texture sequence $\mathbf{X} = [\mathbf{z}_{f_0}, \ldots, \mathbf{z}_{f_T}]$, where each modality has its own learnable query, key, and value projections, denoted as $\{\mathbf{W}_Q^S, \mathbf{W}_K^S, \mathbf{W}_V^S\}$ and $\{\mathbf{W}_Q^X, \mathbf{W}_K^X, \mathbf{W}_V^X\} \in \mathbb{R}^{d\times d}$, with d denoting the latent dimension. This allows each modality to attend to the other across all timepoints $i, j \in \{0, 1, \ldots, T\}$ by

$$\begin{aligned} Q_i^S &= \mathbf{W}_Q^S\, \mathbf{z}_{v_i}, \quad K_j^S = \mathbf{W}_K^S\, \mathbf{z}_{v_j}, \quad V_j^S = \mathbf{W}_V^S\, \mathbf{z}_{v_j}, \\ Q_i^X &= \mathbf{W}_Q^X\, \mathbf{z}_{f_i}, \quad K_j^X = \mathbf{W}_K^X\, \mathbf{z}_{f_j}, \quad V_j^X = \mathbf{W}_V^X\, \mathbf{z}_{f_j}. \end{aligned} \tag{3}$$

The resulting attention maps $\mathbf{A}^{S\to X}, \mathbf{A}^{X\to S} \in \mathbb{R}^{(T+1)\times(T+1)}$ explicitly parameterize cross-phase dependencies, where i and j index the query and key timepoints, each entry $\mathbf{A}_{ij}^{S\to X}$ captures the relevance of texture at timepoint j to shape at timepoint i, and vice versa, as

$$\mathbf{A}_{ij}^{S\to X} = \frac{\exp\left(Q_i^S \cdot K_j^X/\sqrt{d}\right)}{\sum_{j'} \exp\left(Q_i^S \cdot K_{j'}^X/\sqrt{d}\right)}, \quad \mathbf{A}_{ij}^{X\to S} = \frac{\exp\left(Q_i^X \cdot K_j^S/\sqrt{d}\right)}{\sum_{j'} \exp\left(Q_i^X \cdot K_{j'}^S/\sqrt{d}\right)}. \tag{4}$$

We then obtain the attended representations $\tilde{\mathbf{z}}_{v_t}$ and $\tilde{\mathbf{z}}_{f_t}$ as

$$\tilde{\mathbf{z}}_{v_t} = \sum_{j=0}^{T} \mathbf{A}_{tj}^{S\to X}\, V_j^X, \quad \tilde{\mathbf{z}}_{f_t} = \sum_{j=0}^{T} \mathbf{A}_{tj}^{X\to S}\, V_j^S. \tag{5}$$

Similar to [23], the final feature representations are stabilized using residual connections and layer normalization modules, i.e., $\hat{\mathbf{z}}_{v_t} = \mathrm{LN}(\mathbf{z}_{v_t} + \tilde{\mathbf{z}}_{v_t})$ and $\hat{\mathbf{z}}_{f_t} = \mathrm{LN}(\mathbf{z}_{f_t} + \tilde{\mathbf{z}}_{f_t})$.

Adaptive Gating and Diagnostic Importance Pooling. Since the diagnostic relevance of shape and texture is inherently timepoint-dependent, symmetric combinations of $\hat{\mathbf{z}}_{v_t}$ and $\hat{\mathbf{z}}_{f_t}$ would ignore this critical property. We therefore introduce a per-timepoint adaptive gate g_t that dynamically controls the relative contribution of each modality as

$$g_t = \sigma(\mathbf{W}_g\,[\hat{\mathbf{z}}_{v_t} \,\|\, \hat{\mathbf{z}}_{f_t}] + b_g)\,, \quad \mathbf{z}_t = g_t \odot \hat{\mathbf{z}}_{v_t} + (1 - g_t) \odot \hat{\mathbf{z}}_{f_t}, \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{W}_g \in \mathbb{R}^{d\times 2d}$ and $\mathbf{b}_g \in \mathbb{R}^d$ are the learnable gating weight and bias, $\|$ denotes concatenation, $\odot$ is element-wise multiplication, $g'_t = 1 - g_t$ is the complementary gate applied to the latent texture representation $\hat{\mathbf{z}}_{f_t}$. In contrast to previous uniform temporal pooling that treats all cardiac phases as equally informative, we further aggregate the fused sequence $\{\mathbf{z}_0, \ldots, \mathbf{z}_T\}$ via a learned diagnostic importance weighting by following Bahdanau Attention [5]

$$\alpha_t = \frac{\exp\left(\mathbf{w}^\top \tanh(\mathbf{W}_h\,\mathbf{z}_t)\right)}{\sum_{t'} \exp\left(\mathbf{w}^\top \tanh(\mathbf{W}_h\,\mathbf{z}_{t'})\right)}, \quad \mathbf{h} = \sum_{t=0}^{T} \alpha_t\,\mathbf{z}_t, \tag{7}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{W}_h \in \mathbb{R}^{d\times d}$ are learnable parameters, t' indexes the timepoints $\{0, \ldots, T\}$ over which the attention weights are normalized, and α_t reflects the learned diagnostic importance of each cardiac phase.

3.2 Training Objective

We train the classification framework in two stages. Following Eqs. (1)- (2), we first optimize the shape network (θ_S^E, θ_S^D) over the velocity fields v by minimizing the objective function defined as

$$\mathcal{L}_{\text{reg}}(\theta_S^E, \theta_S^D) = \sum_{t=1}^{T} \left[\frac{1}{\sigma^2} \mathcal{D}[I_0 \circ \phi_t(\mathbf{z}_{v_t}(\theta_S^E, \theta_S^D)), I_t] + \|\nabla v_t\|_1 \right] + \text{reg}(\theta_S^E, \theta_S^D), \tag{8}$$

where $\mathcal{D}[\cdot,\cdot]$ denotes the sum-of-squared differences between the warped source $I_0 \circ \phi_t$ and target I_t, σ^2 denotes noise variance, $\|\nabla v_t\|_1$ enforces smoothness of the velocity field, and $\text{reg}(\cdot)$ denotes network parameter regularization.

During the second stage, the latent representations $\{\mathbf{z}_{v_t}\}$ learned from Eq. (8) are held fixed. The image encoder θ_I^E, cross-modal attention $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$, adaptive gating $\{\mathbf{W}_g, \mathbf{b}_g\}$, diagnostic importance pooling $\{\mathbf{w}, \mathbf{W}_h\}$, and classifier θ_c, collectively denoted as Θ, are jointly optimized via a binary cross-entropy loss over the aggregated representation $\mathbf{h}$ as

$$\mathcal{L}_{\text{cls}}(\Theta) = -\frac{1}{N} \sum_{n=1}^{N} [y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)] + \text{reg}(\Theta), \tag{9}$$

where $\hat{y}_n = \theta_c(\mathbf{h}_n)$ is the predicted probability for sample n and $\text{reg}(\Theta)$ denotes a regularization term on the parameters Θ.

4 Experimental Evaluation

We evaluate `ShapeFuse` on a cine CMR video dataset [25]. This dataset consists of 510 cine CMR video sequences from 125 subjects, each accompanied by manually delineated left ventricular myocardium segmentation maps. Nearly 40% of subjects present with myocardial scar-induced wall motion abnormalities, providing a clinically relevant class imbalance. All sequences span a complete cardiac cycle of 24 time frames and are standardized to $224 \times 224 \times 24$.

4.1 Experiments

Effectiveness of `ShapeFuse` in Image Classification. We validate our model on classification tasks, adopting convolutional architectures (ResNet [13], EfficientNet [22], and DenseNet [16]) as image encoder backbones. Performance is evaluated using micro-averaged accuracy and F1-score. We ablate the fusion module by replacing `ShapeFuse` with addition, concatenation [24], weighted [9], bilinear [20], and temporal attention-based fusion strategies, evaluated across all encoder backbones and two registration networks, including VoxelMorph (VM) [6] and TLRN [27].
Interpretability Analysis. We visualize Grad-CAM activation maps across cardiac timepoints and compare them with those produced by models employing conventional fusion strategies to assess how different fusion mechanisms influence spatial attention and feature localization.
Cross-Modal Attention and Gating Analysis. We further analyze the learned cross-modal attention maps and adaptive gating distributions to examine the dynamic interaction between shape and texture modalities across the cardiac phases.

Implementation Details. Both shape and texture features are projected into a shared hidden space of dimension $d = 512$, fused via bidirectional multi-head cross-attention (8 heads) with residual connections and layer normalization. We pass the fused representation through a three-layer fully connected classifier with hidden dimensions 512, 256, and 64, with batch normalization, LeakyReLU activations, and dropout ($p = 0.3$). We kept the shape encoder frozen during classifier training, but joint training can also be done. We optimize the model using AdamW [19] with a learning rate of 1×10^{-5} and weight decay of 0.01, with a ReduceLROnPlateau scheduler (factor 0.5). We train the model up to 500 epochs with early stopping (patience 20, $\delta = 0.001$) based on validation loss. We utilize a 40GB NVIDIA A100 Tensor Core GPU for training and testing.

4.2 Results

Tab. 1 reports the classification performance of `ShapeFuse` against established fusion strategies across multiple image encoder backbones and two registration

Table 1. Classification performance comparison of ShapeFuse against competing fusion strategies across multiple image encoder backbones and registration networks.

	Backbone	*ResNet*			*EfficientNet*			*DenseNet*		
	Metrics	Acc	F1-sc	AUC	Acc	F1-sc	AUC	Acc	F1-sc	AUC
	Image	0.764	0.751	0.796	0.787	0.822	0.923	0.797	0.802	0.896
VM	Shape	0.809	0.825	0.906	0.809	0.825	0.906	0.809	0.825	0.906
VM	+ Add	0.807	0.884	0.959	0.787	0.822	0.928	0.820	0.849	0.949
VM	+ Concat [24]	0.764	0.800	0.926	0.809	0.841	0.923	0.775	0.818	0.921
VM	+ Weighted [9]	0.832	0.857	0.957	0.787	0.816	0.914	0.753	0.804	0.934
VM	+ Bilinear [20]	0.832	0.857	0.957	0.820	0.846	0.921	0.787	0.829	0.971
VM	+ Attention	0.787	0.808	0.877	0.820	0.852	0.949	0.843	0.865	0.977
VM	+ ShapeFuse	**0.888**	**0.902**	**0.966**	**0.843**	**0.857**	**0.937**	**0.854**	**0.871**	**0.940**
TLRN	Shape	0.797	0.836	0.922	0.798	0.830	0.925	0.775	0.811	0.866
TLRN	+ Add	0.854	0.871	0.950	0.831	0.857	0.956	0.820	0.852	0.950
TLRN	+ Concat [24]	0.820	0.840	0.918	0.843	0.860	0.935	0.831	0.860	0.968
TLRN	+ Weighted [9]	0.854	0.876	0.963	0.865	0.882	0.964	0.820	0.852	0.953
TLRN	+ Bilinear [20]	0.876	0.884	0.952	0.854	0.874	0.956	0.855	0.871	0.961
TLRN	+ ShapeFuse	**0.899**	**0.901**	**0.965**	**0.876**	**0.884**	**0.954**	**0.876**	**0.891**	**0.964**

networks. Under both VM and TLRN registration backbones, ShapeFuse consistently outperforms all competing fusion strategies, demonstrating that explicitly modeling cross-modal dependencies yields meaningful gains over naive approaches such as concatenation and element-wise addition, which fail to reliably improve over the shape-only baseline. The consistent gains across all image encoders and registration backbones confirm that ShapeFuse is an effective and generalizable cross-modal fusion module for cardiac video classification.

Fig. 2 visualizes Grad-CAM activation maps across cardiac timepoints for different fusion strategies. Unlike competing methods that produce diffuse or spatially inconsistent activations, ShapeFuse consistently highlights the myocardial region across all timepoints, which is particularly meaningful given that myocardial motion throughout the cardiac cycle is an established indicator of scar presence and regional wall motion abnormalities. This targeted localization demonstrates that ShapeFuse captures temporally evolving myocardial deformation patterns most relevant to scar diagnosis, yielding both superior classification performance and more clinically interpretable spatial attention maps.

Fig. 3 validates ShapeFuse's learned cross-modal attention and adaptive gating. The shape-to-image attention shows deformation features consistently attending to mid-systolic texture frames, suggesting peak contraction carries the most discriminative evidence for scar classification. The image-to-shape attention follows a diagonal pattern, indicating texture representations seek phase-aligned geometric context. This asymmetry confirms that BCTA captures complementary cross-modal interactions that symmetric fusion operators cannot model. The gating distribution centered slightly below $g = 0.5$ reflects that tex-

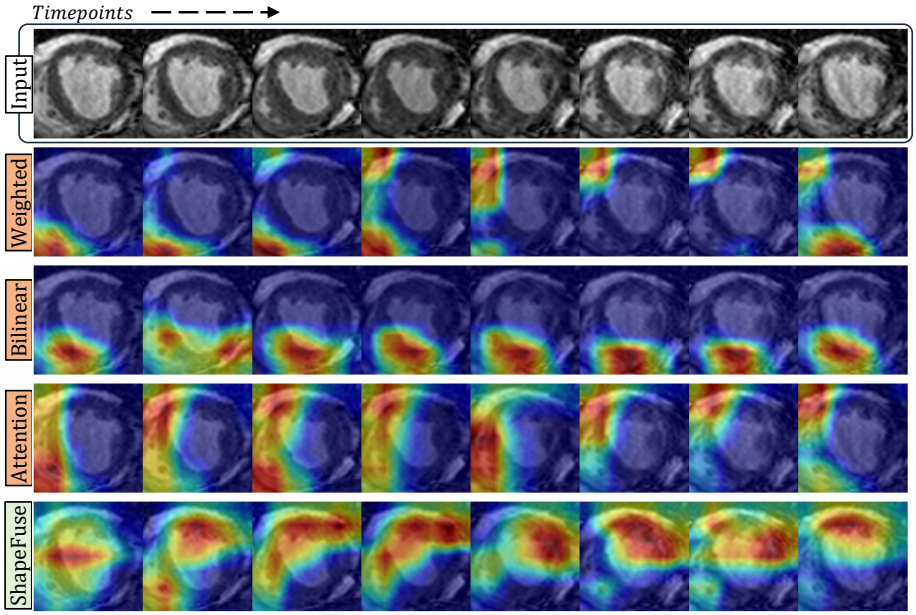

Fig. 2. Grad-CAM activation maps across the cardiac timepoints for `ShapeFuse` and baseline fusion strategies.

ture representations benefit more from geometric grounding, consistent with the clinical challenge of identifying scar-induced hypokinesia from intensity alone. The per-dimension gate weight profile further shows clean separation between shape/image-dominant dimensions, confirming the adaptive gate allocates the two modalities into complementary subspaces, providing evidence that `ShapeFuse`'s learned fusion directly translates into the classification gains observed in Tab. 1.

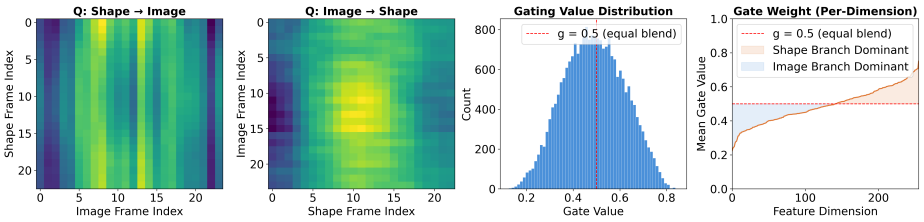

Fig. 3. Analysis of the learned cross-modal attention and adaptive gating strategy.

5 Conclusion

This paper presents ShapeFuse, a novel framework for cardiac video classification that jointly learns the fusion of deformable shape and texture representations in a shared latent space. By replacing naive concatenation with bidirectional cross-modal temporal attention and adaptive gating, ShapeFuse explicitly models cross-modal dependencies and dynamically weights modality contributions across the cardiac cycle, achieving state-of-the-art classification performance on a cine CMR dataset with improved interpretability.

In future work, we plan to extend ShapeFuse to multi-label pathology classification, where multiple co-occurring conditions require simultaneous localization, and to explore self-supervised pretraining of the cross-modal attention module to reduce reliance on labeled data in low-resource clinical settings.

Acknowledgments. This work was supported by NSF CAREER Grant 2239977 and NIH 1R21EB032597.

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

References

1. Alis, D., Yergin, M., Asmakutlu, O., Topel, C., Karaarslan, E.: The influence of cardiac motion on radiomics features: radiomics features of non-enhanced cmr cine images greatly vary through the cardiac cycle. European Radiology **31**(5), 2706–2715 (2021)
2. Amyar, A., Nakamori, S., Morales, M., Yoon, S., Rodriguez, J., Kim, J., Judd, R.M., Weinsaft, J.W., Nezafat, R.: Gadolinium-free cardiac mri myocardial scar detection by 4d convolution factorization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 639–648. Springer (2023)
3. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 924–931. Springer (2006)
4. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis **12**(1), 26–41 (2008)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
6. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
7. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**, 139–157 (2005)
8. Cau, R., Pisu, F., Pintus, A., Palmisano, V., Montisci, R., Suri, J.S., Salgado, R., Saba, L.: Cine-cardiac magnetic resonance to distinguish between ischemic and

non-ischemic cardiomyopathies: a machine learning approach. European Radiology **34**(9), 5691–5704 (2024)
9. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 865–872 (2019)
10. Clough, J.R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A.P., Schnabel, J.A.: Global and local interpretability for cardiac mri classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656–664. Springer (2019)
11. Foley, J.R., Swoboda, P.P., Fent, G.J., Garg, P., McDiarmid, A.K., Ripley, D.P., Erhayiem, B., Musa, T.A., Dobson, L.E., Plein, S., et al.: Quantitative deformation analysis differentiates ischaemic and non-ischaemic cardiomyopathy: sub-group analysis of the vindicate trial. European Heart Journal-Cardiovascular Imaging **19**(7), 816–823 (2018)
12. Gao, Q., Yi, W., Gao, C., Qi, T., Li, L., Xie, K., Zhao, W., Chen, W.: Cardiac magnetic resonance feature tracking myocardial strain analysis in suspected acute myocarditis: diagnostic value and association with severity of myocardial injury. BMC Cardiovascular Disorders **23**(1), 162 (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hossain, T., Ma, J., Li, J., Zhang, M.: Invariant shape representation learning for image classification. arXiv preprint arXiv:2411.12201 (2024)
15. Hossain, T., Zhang, M.: Corld: Contrastive representation learning of deformable shapes in images. In: International Conference on Information Processing in Medical Imaging. pp. 342–357. Springer (2025)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
17. Jacob, A.J., Chitiboi, T., Schoepf, U.J., Sharma, P., Aldinger, J., Baker, C., Lautenschlager, C., Emrich, T., Varga-Szemes, A.: Deep-learning-based disease classification in patients undergoing cine cardiac mri. Journal of Magnetic Resonance Imaging **61**(4), 1635–1647 (2025)
18. Jayakumar, N., Hossain, T., Zhang, M.: Sadir: Shape-aware diffusion models for 3d image reconstruction. In: International Workshop on Shape in Medical Imaging. pp. 287–300. Springer (2023)
19. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Ni, Z.L., Bian, G.B., Li, Z., Zhou, X.H., Li, R.Q., Hou, Z.G.: Space squeeze reasoning and low-rank bilinear feature fusion for surgical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(7), 3209–3217 (2022)
21. Qin, C., Wang, S., Chen, C., Bai, W., Rueckert, D.: Generative myocardial motion tracking via latent space exploration with biomechanics-informed prior. Medical Image Analysis **83**, 102682 (2023)
22. Tan, M.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

24. Wang, J., Zhang, M.: Geo-sic: Learning deformable geometric shapes in deep image classifiers. The Conference on Neural Information Processing Systems (2022)
25. Wang, S., Patel, H., Miller, T., Ameyaw, K., Narang, A., Chauhan, D., Anand, S., Anyanwu, E., Besser, S.A., Kawaji, K., et al.: Ai based cmr assessment of biventricular function: clinical significance of intervendor variability and measurement errors. Cardiovascular Imaging **15**(3), 413–427 (2022)
26. Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R.: Multi-modal volume registration by maximization of mutual information. Medical image analysis **1**(1), 35–51 (1996)
27. Wu, N., Xing, J., Zhang, M.: Tlrn: Temporal latent residual networks for large deformation image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 728–738. Springer (2024)
28. Xie, Y., Zhong, H., Wu, J., Zhao, W., Hou, R., Zhao, L., Xu, X., Zhang, M., Zhao, J.: Automatic classification of heart failure based on cine-cmr images. International Journal of Computer Assisted Radiology and Surgery **19**(2), 355–365 (2024)
29. Xing, J., Jayakumar, N., Wu, N., Wang, Y., Epstein, F.H., Zhang, M.: Lamod: Latent motion diffusion model for myocardial strain generation. In: International Workshop on Shape in Medical Imaging. pp. 164–177. Springer (2024)
30. Xing, J., Wu, N., Bilchick, K.C., Epstein, F.H., Zhang, M.: Multimodal learning to improve cardiac late mechanical activation detection from cine mr images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2024)
31. Zheng, Q., Delingette, H., Ayache, N.: Explainable cardiac pathology classification on cine mri with motion characterization by semi-supervised learning of apparent flow. Medical image analysis **56**, 80–95 (2019)