Self-Supervised Temporal Regularization for Landmark-Based Cardiac Segmentation with Automatic AHA Regional Mapping

David Montalvo-García[1,2(✉)], Nicolás Gaggion[3,4,5], María J. Ledesma-Carbayo[1,2], and Enzo Ferrante[5(✉)]

[1] Biomedical Image Technologies, ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
david.montalvo@upm.es
[2] Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Instituto de Salud Carlos III, Madrid, Spain
[3] APOLO Biotech, Ciudad Autónoma de Buenos Aires, Argentina
[4] Departamento de Computación, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
[5] Instituto de Ciencias de la Computación (ICC), CONICET-Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
eferrante@dc.uba.ar

Abstract. Graph-based cardiac segmentation with implicit anatomical correspondences provides topological guarantees and population-level analysis capabilities, but models trained on independent frames of image sequences exhibit temporal discontinuities that affect reliable clinical measurements, particularly in cardiac ultrasound. In this work, we introduce self-supervised temporal regularization as a post-training refinement stage that exploits the temporal coherence in image sequences to enforce consistent cardiac segmentation and motion estimation over time, without requiring per-frame annotations. By penalizing velocity and acceleration discontinuities across consecutive frames, our method achieves temporally consistent segmentations while maintaining the learned anatomical correspondences. We further leverage these correspondences to automatically map landmarks to the AHA 17-segment clinical standard, enabling standardized regional assessment and detection of pathological myocardial motion patterns. Validation on CAMUS dataset demonstrates the clinical utility of combining temporal consistency with automatic regional mapping. The code is publicly available at https://github.com/david-montalvoo/MaskHybridGNet-TempReg.

Keywords: Cardiac segmentation · Temporal consistency · Self-supervised learning · AHA segments · Implicit correspondences

1 Introduction

Deep learning has transformed medical image segmentation, with U-Net architectures [16] and Vision Transformers [19] achieving remarkable performance across

modalities [13]. However, pixel-based methods trained with cross-entropy or Dice losses [14] lack explicit anatomical knowledge, producing segmentations with topological inconsistencies, holes, and irregular boundaries under challenging conditions [3, 4]. Graph-based segmentation addresses these limitations by representing anatomical boundaries as connected graphs with guaranteed topology [2]. Recent advances in implicit anatomical correspondence learning [6] enable training on standard pixel-wise masks using Chamfer distance, where the i-th predicted landmark emerges to consistently represent the same anatomical location across patients, providing competitive accuracy with superior robustness and enabling automatic atlas generation.

Despite anatomical consistency across patients, these models process frames independently, ignoring temporal coherence. For cardiac imaging, this produces frame-to-frame discontinuities that degrade ejection fraction estimation, prevent accurate wall motion tracking required for clinical assessment, introduce artificial jitter incompatible with physiological cardiac dynamics, and limit translation to clinical workflows. Existing temporal consistency approaches rely on post-processing with optical flow [12], Kalman filtering, or autoencoder refinement [15]. While effective, these operate independently of the segmentation model and cannot leverage learned anatomical structure.

In this work, we introduce self-supervised temporal regularization as a post-training refinement stage for landmark-based cardiac segmentation. Starting from a model with established implicit correspondences, we fine-tune on ultrasound sequences by penalizing velocity and acceleration discontinuities between consecutive frames. This exploits the fundamental principle that anatomically-corresponding landmarks should move smoothly due to physiological constraints. Critically, this requires no per-frame annotations: the temporal behavior itself provides the supervisory signal. We further leverage learned correspondences to automatically map landmarks to the AHA 17-segment standard [1], enabling clinically-interpretable regional analysis. This combination of temporal consistency and automatic regional mapping allows detection of pathological wall motion patterns such as regional dyssynchrony and hypokinesis, making the approach directly applicable to clinical workflows.

1.1 Related Work

Landmark-Based Cardiac Segmentation. Point Distribution Models [5] and Active Shape Models pioneered landmark-based anatomical modeling through statistical shape variation. HybridGNet [7, 8] combined CNNs with spectral graph convolutions for variational cardiac segmentation, demonstrating superior domain shift robustness but requiring manually annotated landmarks with anatomical correspondences. Recent work [6] eliminates this requirement by training on standard segmentation masks, where the i-th landmark emerges to consistently represent the same anatomical location through optimization with Chamfer distance and edge regularization. This enables topologically-guaranteed segmentation with automatic correspondence discovery. However, temporal consistency

remains unaddressed as models process frames independently.

Temporal Consistency in biomedical image sequences. Classical approaches employ registration or optical flow based techniques [12] as well as Kalman filtering for exploiting temporal coherence. Ledesma-Carbayo et al. [12] introduced a B-spline based spatio-temporal registration framework to exploit temporal coherence for the estimation of cardiac motion on ultrasound sequences. Sundar et al. [17] proposed simultaneous 4D registration where all cardiac phases are jointly optimized for temporal smoothness. Recent deep learning methods include recurrent architectures, test-time refinement, and semi-supervised approaches. Painchaud et al. [15] perform post-hoc temporal refinement using a constrained autoencoder in the attributes driven latent representation. Tziritas [18] improves temporal coherence through post-processing of compact boundary representations. Wu et al. [20] exploit inter-frame context within a semi-supervised video segmentation framework. Our work differs by integrating temporal regularization as a fine-tuning stage for landmark-based segmentation that leverages learned anatomical correspondences, enabling the model to refine its temporal representations while preserving spatial accuracy and structure correspondence.

2 Methods

2.1 Background: Implicit Correspondence Learning

Our framework builds on the implicit correspondence learning approach [6] where a graph-based VAE model learns to map input images $\mathbf{I}$ to fixed-size landmark graphs $\mathbf{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$. Here V represents a set of M nodes corresponding to anatomical landmarks, $\mathbf{A}$ is an adjacency matrix encoding connectivity between organ contours, and $\mathbf{X} \in \mathbb{R}^{M \times 2}$ contains the landmark coordinates. The encoder maps images to a latent distribution $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, which the graph-convolutional decoder transforms into landmark coordinates. Completing the Mask-HybridGNet Dual architecture, an auxiliary convolutional decoder generates dense pixel-level masks and routes its multi-scale features directly to the graph decoder via image-to-graph skip connections.

The graph adjacency matrix is constructed automatically from an atlas derived from CAMUS ground-truth segmentation masks to represent the left ventricular endocardium, epicardium, and left atrium as a unified graph. This structure enforces anatomical constraints through shared boundaries: the left ventricular endocardium shares nodes with the inner myocardial wall, and the basal endocardium shares nodes with the left atrium at the mitral valve plane.

We train the base model with a compound loss functions which uses Chamfer distance ($\mathcal{L}_{CD}$) to match variable-length ground truth contours $\mathbf{P}$ with the fixed-size predictions $\mathbf{G}$, combined with edge regularization for uniform spacing between landmarks ($\mathcal{L}_{uniform}$), elasticity regularization to encourage compact contours ($\mathcal{L}_{elastic}$), and curvature regularization ($\mathcal{L}_{smooth}$) to promote local smoothness by penalizing abrupt direction changes between consecutive edges, along with a

standard VAE Kullback–Leibler divergence on the variational latents ($\mathcal{L}_{KL}$). We also include a Dice pixel-loss ($\mathcal{L}_{pixel}$) computed by comparing the ground-truth masks with both, the rasterized graph-contour predictions and the dense output of the auxiliary CNN decoder. This process discovers anatomical correspondences where landmark i represents consistent anatomy across all subjects. The standard frame-independent training objective is:

$$\mathcal{L}_{base} = \mathcal{L}_{CD} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_e\mathcal{L}_{elastic} + \lambda_u\mathcal{L}_{uniform} + \lambda_s\mathcal{L}_{smooth} + \lambda_p\mathcal{L}_{pixel} \quad (1)$$

For a more detailed description of these loss terms and the Mask-HybridGNet Dual architecture we refer the reader to [6]. After training with this objective, the model produces anatomically consistent segmentations with implicit correspondences. However, the combination of fixed-size predictions and Chamfer distance matching can cause temporal instability: landmarks may shift by one or two positions along the circular contour between consecutive frames, resulting in correspondence oscillations and abrupt temporal transitions.

2.2 Self-Supervised Temporal Regularization

To address the temporal instability of the predicted landmarks while preserving learned correspondences, we introduce temporal regularization as a post-training refinement stage. Starting from a converged model with established anatomical correspondences, we fine-tune on video sequences by adding temporal constraints that exploit the principle that anatomically-corresponding landmarks should exhibit smooth motion profiles characteristic of physiological cardiac dynamics.

Let $\{\mathbf{I}^{(t)}\}_{t=1}^{T}$ denote a temporal sequence of T consecutive frames from the same cardiac cycle. For each frame, the encoder produces latent code $\mathbf{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2)$ and the decoder generates a set of landmarks $\mathbf{X}(t) = \left\{x_i(t)\right\}_{i=1}^{M} \in \mathbb{R}^{M\times 2}$. We introduce two complementary temporal regularization terms that penalize motion discontinuities while maintaining the spatial accuracy learned during initial training.

Velocity regularization encourages smooth trajectories for anatomically corresponding landmarks by penalizing large frame-to-frame displacements:

$$\mathcal{L}_{vel} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\,\mathbf{x}_i(t+1) - \mathbf{x}_i(t)\right\|^2\right) \quad (2)$$

Acceleration regularization penalizes abrupt changes in velocity by minimizing the difference between consecutive velocity vectors, enforcing physiologically plausible acceleration and deceleration patterns:

$$\mathcal{L}_{accel} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{T-2}\sum_{t=1}^{T-2}\left\|\mathbf{x}_i(t+2) - 2\,\mathbf{x}_i(t+1) + \mathbf{x}_i(t)\right\|^2\right) \quad (3)$$

The refinement objective combines the original spatial terms with temporal regularization:

$$\mathcal{L}_{refine} = \mathcal{L}_{base} + \lambda_v \mathcal{L}_{vel} + \lambda_a \mathcal{L}_{accel} \tag{4}$$

where λ_v, λ_a control the strength of temporal regularization.

Training Procedure. We implement an alternating optimization strategy to balance spatial accuracy with temporal consistency, based on a composite $\mathcal{L}_{refine}$ function. Each iteration first processes N independent annotated ED/ES frames to compute the supervised spatial loss ($\mathcal{L}_{base}$), which anchors the model and prevents collapse to a trivial zero-velocity solution. The model then processes a temporal sequence of T consecutive unannotated frames to compute the temporal regularization loss terms ($\mathcal{L}_{vel}$ and $\mathcal{L}_{accel}$). The total $\mathcal{L}_{refine}$ combines the gradients from these two forward passes. Since the T sequence frames use no GT masks, the temporal regularization component is self-supervised. The sequence length T is constrained by GPU memory requirements, as processing full sequences requires maintaining activations for all T frames during backpropagation.

2.3 Automatic AHA Mapping

We leverage the learned implicit correspondences to automatically map landmarks to the standardized AHA 17-segment model [1], translating established clinical echocardiography workflows [10] into our automated pipeline. Because the landmark i consistently represents the same anatomical location across subjects and frames, this mapping is computed just once on a population-averaged reference atlas, constructed by averaging the spatial coordinates of all subjects at the end-diastolic frame, and applied universally. As illustrated for the 4-chamber view in Figure 1, the procedure automatically identifies the mitral valve center from shared left ventricular and atrial nodes, establishing the primary long-axis vector toward the endocardial apex. The left ventricle is then divided along this long axis into basal (up to 33%), mid-cavity (33% to 66%), and apical (66% to 98%) levels, reserving the final 2% for the apex cap. These levels are perpendicularly bisected to separate the opposing walls (anterolateral and inferoseptal in the 4-chamber view; anterior and inferior in the 2-chamber view). Finally, epicardial nodes are assigned to the segment of their nearest endocardial neighbor.

3 Experiments, results & discussion

Datasets and Evaluation. The CAMUS dataset [11] provides 500 patients with 2-chamber and 4-chamber apical echocardiography sequences of approximately 20 frames each. Expert annotations mark LV endocardium, epicardium, and left atrium at end-diastole and end-systole, with metadata including ejection fraction and image quality scores. We use the 400-50-50 train/val/test split provided.

For spatial accuracy, we compute the Dice coefficient by rasterizing the predicted landmarks into binary masks, and the 95th percentile Hausdorff distance (HD95) to quantify boundary error while reducing sensitivity to outlier points.

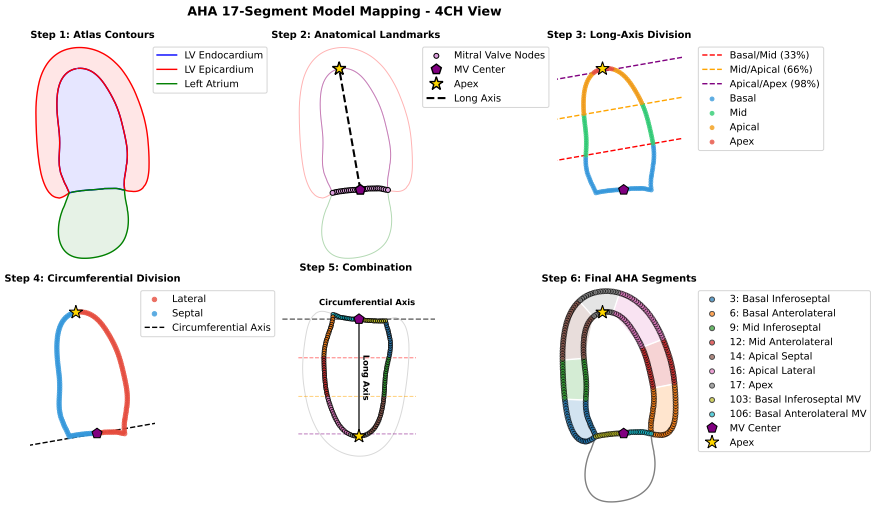

Fig. 1. Automatic AHA mapping via implicit correspondence. Step-by-step procedure for the apical 4-chamber view: (1) population atlas generation at ED; (2) anatomical landmark identification and long axis definition; (3) long-axis division into basal (0–33%), mid (33–66%), and apical (66–98%) levels; (4) circumferential bisection; (5) combined classification; and (6) final AHA segment assignments. The 2-chamber view follows an identical procedure to map its respective anterior and inferior segments.

Clinical metrics were selected to reflect standard measures used in routine cardiac practice, and performance was evaluated by computing the mean absolute error (MAE) between the reference and predicted values for ejection fraction (EF), end-systolic volume (ESV), and end-diastolic volume (EDV), with ventricular volumes computed from masks using the biplane Simpson's method.

To quantify the temporal regularity of the trajectories while excluding the expected displacements induced by cardiac motion, we use a jitter metric that captures high-frequency perturbations. Specifically, jitter is defined as the root-mean-square (RMS) deviation between each aligned trajectory and a low-pass filtered version of the same trajectory over time:

$$\text{Jitter} = \frac{1}{M}\sum_{i=1}^{M}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\| \mathbf{x}_i(t) - \bar{\mathbf{x}}_i(t) \right\|^2}\right) \tag{5}$$

where $\bar{\mathbf{x}}_i(t)$ denotes the smoothed trajectory obtained by applying a Savitzky–Golay filter independently to each aligned trajectory. As complementary measures of trajectory quality and smoothness, we also report frame-to-frame displacement (FTD), which quantifies the RMS excursion distance for each node

$$\text{FTD} = \frac{1}{M}\sum_{i=1}^{M}\left(\sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}\left\| \mathbf{x}_i(t+1) - \mathbf{x}_i(t) \right\|^2}\right) \tag{6}$$

and RMS jerk, which measures the temporal variation of acceleration and penalizes abrupt motion changes using the 3rd-order finite difference

$$\text{Jerk} = \frac{1}{M}\sum_{i=1}^{M}\left(\sqrt{\frac{1}{T-3}\sum_{t=1}^{T-3}\left\|\,\Delta^3\mathbf{x}_i(t)\right\|^2}\right) \tag{7}$$

Results. We first trained Mask-HybridGNet Dual on CAMUS without temporal regularization, using the loss in Eq. 1 and ED and ES frames with ground-truth masks from both 2-chamber and 4-chamber views. We selected the dual branch architecture because it showed the best performance in the experiments reported in [6]. Figure 2 summarizes all spatial, clinical, and trajectory metrics. This baseline achieves strong segmentation performance (Dice and HD95), but relatively poor trajectory quality and smoothness metrics, consistent with the absence of temporal regularization during training.

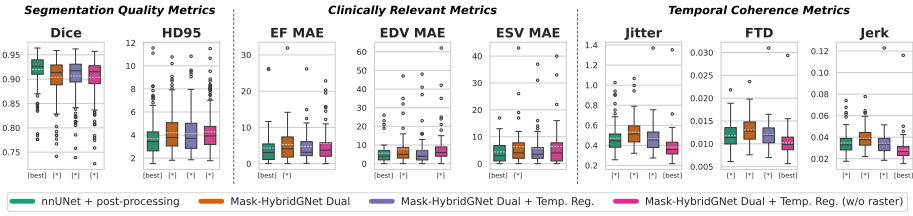

Fig. 2. Quantitative comparison across segmentation, clinical, and temporal coherence metrics. Boxplots report the distribution of Dice and HD95 (left), MAE for EF, EDV, and ESV (center), and Jitter, FTD and Jerk (right) computed from predicted landmark trajectories. The label [best] indicates the best mean performance for each metric (higher is better for Dice; lower is better for all others), and [*] denotes statistically significant differences with respect to the best-performing model for that metric according to the Wilcoxon signed-rank test

For comparison, we also trained nnUNet [9] on the same CAMUS images and masks. Since nnUNet is a pixel-wise segmentation model and does not provide landmarks or temporal correspondences, trajectory-based temporal consistency metrics cannot be computed directly from its output. To enable this analysis, we trained a Mask-HybridGNet model that takes multiclass segmentation masks as input (instead of images) and predicts spatially consistent nodes using a unified multiclass contour. This enables the post-processing of nnUNet predictions over the cardiac cycle, converting them into trajectories and quantifying their temporal regularity. Although this pipeline enables trajectory analysis, its temporal coherence metrics show statistically significant differences compared to those of the proposed temporally regularized approach. (see Figure 2).

Starting from the supervised Mask-HybridGNet Dual baseline, we then performed post-training temporal regularization following Section 2.2, using the loss in Eq. 4, which includes $\mathcal{L}_{vel}$ (Eq. 2) and $\mathcal{L}_{accel}$ (Eq. 3) terms with λ_v and λ_a

linearly scheduled from 0.01 to 0.001 over 500k iterations. We evaluated two variants, with and without rasterization in the $\mathcal{L}_{pixel}$ loss-term (Mask-HybridGNet Dual + Temp. Reg. and Mask-HybridGNet Dual + Temp. Reg. w/o Raster).

Among the evaluated models, Mask-HybridGNet Dual + Temp. Reg. w/o Raster achieves the best trajectory quality metrics, with smoother trajectories than both the non-regularized Mask-HybridGNet baseline and the nnUNet + post-processing pipeline. To perform a qualitative regional analysis, predicted landmarks were assigned to AHA segments using the self-supervised procedure described in Section 2.3. Figure 3 shows AHA segment identification and landmark trajectories from ED to ES in two test cases, together with longitudinal displacements of the segment centroids. The temporally regularized model exhibits smoother segment-centroid trajectories than the post-processed nnUNet, consistent with the improvements observed in the temporal coherence metrics. As illustrated in Figure 3, graph-based models with implicit correspondence learning show strong potential for downstream clinical tasks, including segment-wise AHA longitudinal strain estimation. The Supplementary Video shows the dynamic representation of the comparative performance for 6 test cases.

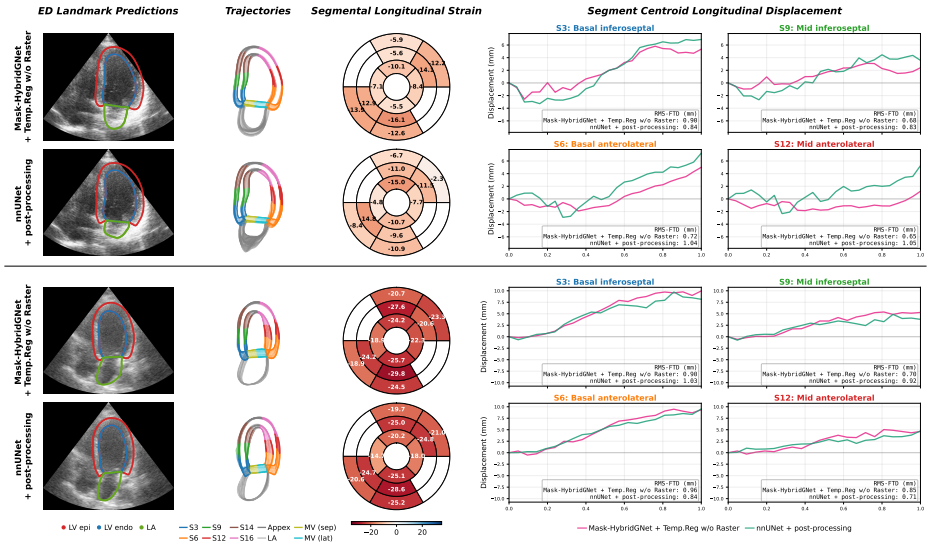

Fig. 3. Qualitative comparison of regional motion trajectories and strain in two representative test cases. For each patient, the figure compares *Mask-HybridGNet + Temp.Reg. w/o Raster* and *nnUNet + post-processing* using (from left to right) ED landmark predictions, full landmark trajectories (4CH) and segmental longitudinal strain based on AHA segments (4CH/2CH), and longitudinal displacement curves of selected AHA segment centroids (S3, S6, S9, and S12), with RMS-FTD values reported in each subplot. Top: Patient0051 (EF=29%, reduced EF). Bottom: Patient0243 (EF=55%, normal EF). The temporally regularized model yields smoother segment-centroid excursions and more regular trajectories, especially at the left atrium.

Discussion & Conclusions. In this work, we introduced a self-supervised post-training temporal regularization strategy for graph-based cardiac ultrasound segmentation that improves temporal coherence while preserving anatomical correspondences, allowing automatic AHA-based regional motion analysis. Our results show that high pixel-level segmentation accuracy does not necessarily ensure temporally consistent trajectories. Although nnUNet performs strongly on conventional spatial and clinical metrics, it lacks explicit landmark correspondences and temporal regularization, limiting direct trajectory-based analysis. Converting nnUNet masks into trajectories partially addresses this limitation, but yields inferior temporal quality compared to models enforcing consistency. The proposed regularized Mask-HybridGNet improves trajectory smoothness while maintaining spatial and clinical performance comparable to nnUNet, and naturally supports standardized regional motion analysis. Future work will evaluate the method on datasets with temporal annotations, such as TED [15], and extend regional strain analysis to hypokinesis and dyssynchrony detection.

Acknowledgments. EF was supported by the Google Award for Inclusion Research, and a Googler Initiated Grant. MJL acknowledges the support of the Spanish Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, under grant PID2022-141493OB-I00 (10.13039/501100011033/MCIN/AEI/ERDF, UE), cofinanced by European Regional Development Fund (ERDF), 'A way of making Europe, and partial funding from the MAGERIT-CM project (TEC-2024/COM-44), supported by the Research and Development Activities Program by the Comunidad de Madrid. DMG was supported by an FPU PhD fellowship funded by the Spanish Ministerio de Ciencia, Innovación y Universidades.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging, Cerqueira, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., et al.: Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. Circulation **105**(4), 539–542 (2002)
2. Boussaid, H., Kokkinos, I., Paragios, N.: Discriminative learning of deformable contour models. In: IEEE International Symposium on Biomedical Imaging (ISBI). pp. 624–628. IEEE (2014)
3. Choudhary, A., Tong, L., Zhu, Y., Wang, M.D.: Advancing medical imaging informatics by deep learning-based domain adaptation. Yearbook of medical informatics **29**(01), 129–138 (2020)
4. Cohen, J.P., Hashir, M., Brooks, R., Bertrand, H.: On the limits of cross-domain generalization in automated X-ray prediction. In: Medical Imaging with Deep Learning. vol. 121, pp. 136–155. PMLR (2020), https://arxiv.org/abs/2002.02497

5. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding **61**(1), 38–59 (1995)
6. Gaggion, N., Ledesma-Carbayo, M.J., Christodoulidis, S., Vakalopoulou, M., Ferrante, E.: Mask-HybridGNet: Graph-based segmentation with emergent anatomical correspondence from pixel-level supervision (2026), https://arxiv.org/abs/2602.21179
7. Gaggion, N., Mansilla, L., Milone, D.H., Ferrante, E.: Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 600–610. Springer (2021)
8. Gaggion, N., Mansilla, L., Mosquera, C., Milone, D.H., Ferrante, E.: Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: Applications to chest X-ray analysis. IEEE Transactions on Medical Imaging **42**(2), 546–556 (2023)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
10. Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., et al.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. European Heart Journal-Cardiovascular Imaging **16**(3), 233–271 (2015)
11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
12. Ledesma-Carbayo, M.J., Kybic, J., Desco, M., Santos, A., Suhling, M., Hunziker, P., Unser, M.: Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation. IEEE Transactions on Medical Imaging **24**(9), 1113–1126 (2005)
13. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
15. Painchaud, N., Duchateau, N., Bernard, O., Jodoin, P.M.: Echocardiography segmentation with enforced temporal consistency. IEEE Transactions on Medical Imaging **41**(10), 2867–2878 (2022)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
17. Sundar, H., Litt, H., Shen, D.: Estimating myocardial motion by 4D image warping. Pattern Recognition **42**(11), 2514–2526 (2009)
18. Tziritas, G.: Eigenboundaries for temporally regularized segmentation of echocardiographic images. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 398–410. Springer (2024)
19. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2021)
20. Wu, H., Lin, J., Xie, W., Qin, J.: Super-efficient echocardiography video segmentation via proxy- and kernel-based semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2803–2811 (2023)