OSA: Echocardiography Video Segmentation via Orthogonalized State Update and Anatomical Prior-aware Feature Enhancement

Supplementary Material

Contents

A. Visualizing Optimization Dynamics

We analyze the geometric properties of the proposed Orthogonalized State Update (OSU) by comparing its optimization trajectories with standard Euclidean gradients.

Manifold-Constrained Optimization. As illustrated in Fig. A.1, Euclidean updates operate in unconstrained space, treating the state matrix as a flattened vector. This approach risks violating the orthogonality constraints of the Stiefel manifold. In contrast, OSU employs an orthogonalized update via Newton-Schulz iteration, first performing a standard gradient step in ambient space then projecting back to the manifold, ensuring strict orthogonality of the state. By enforcing strict orthogonality of the state $\mathbf{S}$, OSU ensures numerical stability and prevents rank collapse over long sequences.

Convergence Stability. The vector fields in Fig. 5 demonstrate that Euclidean gradients exhibit high-frequency oscillations due to a lack of manifold curvature awareness. In video segmentation tasks, such instabilities result in temporal inconsistency across frames. Conversely, OSU functions as an implicit preconditioner, aligning the gradient flow with the optimization landscape. This results in smoother convergence trajectories, which are essential for modeling continuous physiological dynamics, such as the cardiac cycle, without introducing high-frequency noise into the latent representation.

Feature Decorrelation and Consistency. The orthogonality constraints imposed by OSU promote semantic decorrelation within the latent state. By enforcing $\mathbf{S}^\top\mathbf{S} = \mathbf{I}$, OSU reduces feature redundancy and prevents mode collapse. This structural prior enables distinct state dimensions to represent independent dynamics, such as separating low-frequency ventricular contractions from high-frequency valve mechanics. Unlike the Euclidean regime where feature interference leads to temporal instability, OSU ensures that diverse motion patterns evolve along orthogonal trajectories, thereby improving the temporal consistency of the resulting segmentation masks.

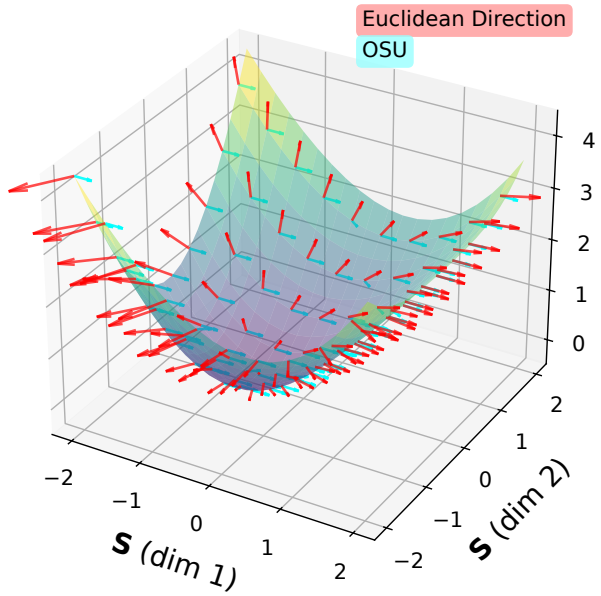


Figure A.1. **3D geometric intuition** of update directions on the loss landscape. The Euclidean Direction updates the state element-wise along the steepest descent, often ignoring local curvature. OSU evolves the state by projecting gradients via the matrix sign function, thereby maintaining the consistency of the underlying subspace structure. The x and y axes represent the two dimensions of the state $\mathbf{S}$. The movement toward the lower regions of the surface represents the direction of loss optimization.

B. Problem Setting

Exhaustive frame-by-frame annotation of echocardiography videos is labor-intensive in clinical settings. Consequently, this work addresses semantic segmentation under a sparse temporal supervision setting [47].

Let the input echocardiography video sequence of length T be denoted as

$$V = \{x_1, x_2, \ldots, x_T\}.$$

where x_t represents the image frame at time step t.

Under the target sparse supervision setting, ground truth segmentation masks are provided exclusively for the first ($t = 1$) and the last ($t = T$) frames, which typically correspond to the end-diastole and end-systole phases. Thus, the index set of annotated frames is strictly defined as

$$S = \{1, T\}.$$

The available ground truth segmentation masks during the training phase are defined as

$$Y_S = \{y_1, y_T\}.$$

This formulation indicates that labels for frames $1 < t < T$ are unavailable during training.

The objective is to train a segmentation model f_θ parameterized by θ. The model processes the entire video sequence V to generate dense segmentation predictions for all frames

$$\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\} = f_\theta(V).$$

Given the sparse annotations, the objective function is computed exclusively on the first and last frames. The loss $\mathcal{L}$ is formulated as

$$\mathcal{L}(\theta) = \ell(\hat{y}_1, y_1) + \ell(\hat{y}_T, y_T),$$

where ℓ denotes a standard segmentation loss function. Optimizing this objective requires the model to utilize temporal context to generate segmentations for the unannotated frames. During the inference phase, the model operates in a fully automated manner, generating predictions for the entire sequence without requiring any additional prompts or manual interventions, simulating a real-world clinical workflow.

C. Experiments Continued

Segmentation evaluation. For a fair comparison, the segmentation metrics (mDice, mHD95) reported in Tab. 1 are evaluated exclusively on the annotated ED and ES frames, consistent with the sparse supervision protocol.

Data and calibration. All results for the medical metrics in Tab. 1 are derived from the official CAMUS test subset. Notably, we report the raw correlation, bias, and standard deviation without applying any bias-correction or post-calibration techniques to reflect the model's direct predictive performance.

Training dynamics and convergence. As illustrated in Fig. A.2, OSA exhibits a clear transition in both segmentation accuracy and feature localization. In the initial phase (iterations 10–30), the segmentation masks are characterized by coarse boundaries, while the corresponding feature activation maps (Figs. A.2a to A.2c, bottom rows) show diffused activations across the spatial domain. By iteration 1250 (epoch 19), the model produces spatially precise segmentations (Fig. A.2d), with feature activations concentrated strictly on the target anatomical structures. This progression indicates that the training protocol facilitates the transition from global spatial search to fine-grained structural delineation without performance degradation.

Temporal consistency analysis. The visualization across the frame sequences in Fig. A.2 demonstrates the capacity of the model to maintain temporal stability. In the early stages, predictions are temporally inconsistent, with visible fluctuations in mask geometry across frames. However, the converged model (Fig. A.2d) preserves structural identity and boundary continuity throughout the cardiac cycle. This stability is attributed to the following mechanisms.

Spatiotemporal feature aggregation. The integration of contextual information from the temporal sequence facilitates the reduction of frame-wise prediction noise by utilizing the temporal evolution of anatomical structures.

Gradient propagation. Backpropagating the loss through the temporal network enforces a consistency constraint on intermediate frames, ensuring that non-annotated frames align with the patterns learned from supervised frames.

Sequential context integration. The refinement of internal representations as the sequence progresses allows the model to utilize cumulative temporal evidence for more accurate boundary localization, effectively correcting initial segmentation noise.

As shown in Fig. A.2, the segmentation quality of intermediate frames without direct supervision remains consistent with the supervised frames. This indicates that the architecture effectively enforces temporal consistency and preserves structural identity throughout the sequence.

D. Clinical Applications

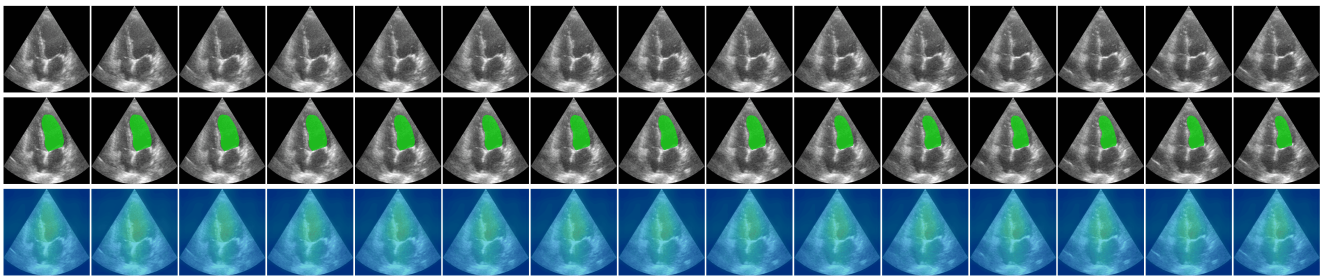
The Left Ventricular Ejection Fraction is a critical metric for cardiac function assessment. Traditionally, it requires accurate measurement of End-Diastolic Volume (V_{ED}) at peak filling and End-Systolic Volume (V_{ES}) at peak contraction. The standard volumetric LV_{EF} formula is defined as:

$$\text{LV}_{\text{EF}} = \frac{V_{\text{ED}} - V_{\text{ES}}}{V_{\text{ED}}} \times 100\%.$$

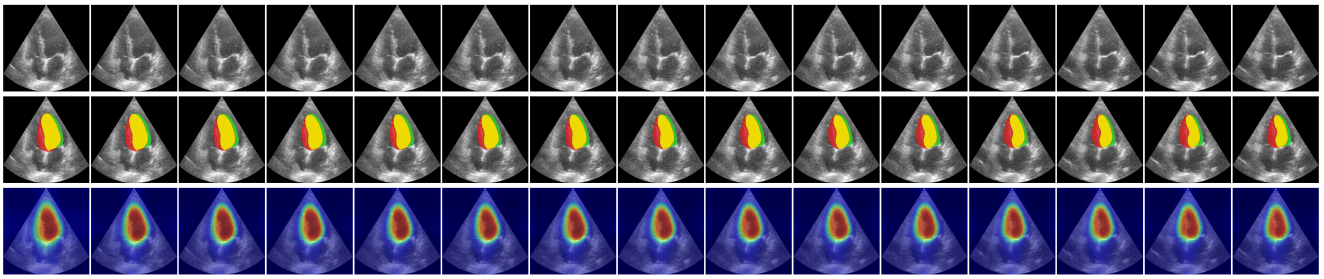
While the Biplane Method of Disks (modified Simpson's rule) is the clinical gold standard for computing these volumes, it requires consistent dual-view data. To ensure compatibility with single-view datasets (*e.g.*, EchoNet-Dynamic) and to directly leverage our automated segmentation framework, we employ a surrogate EF based on 2D area changes:

$$EF_{area} = \frac{A_{ED} - A_{ES}}{A_{ED}}.$$

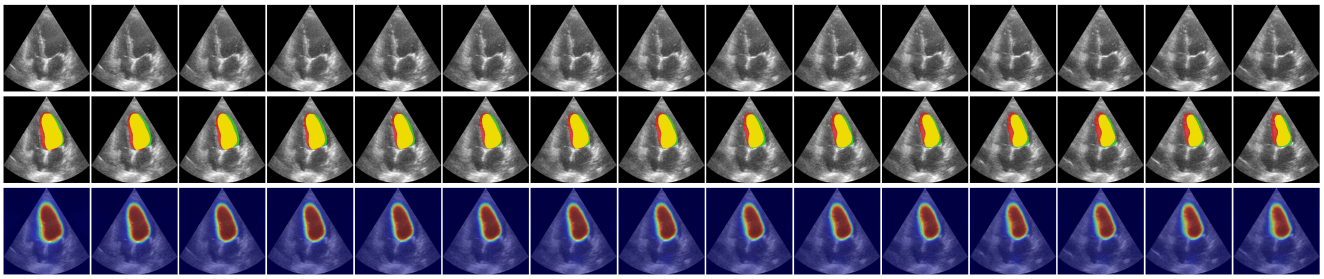
where A_{ED} and A_{ES} denote the segmented Left Ventricular areas at end-diastole and end-systole, respectively. This



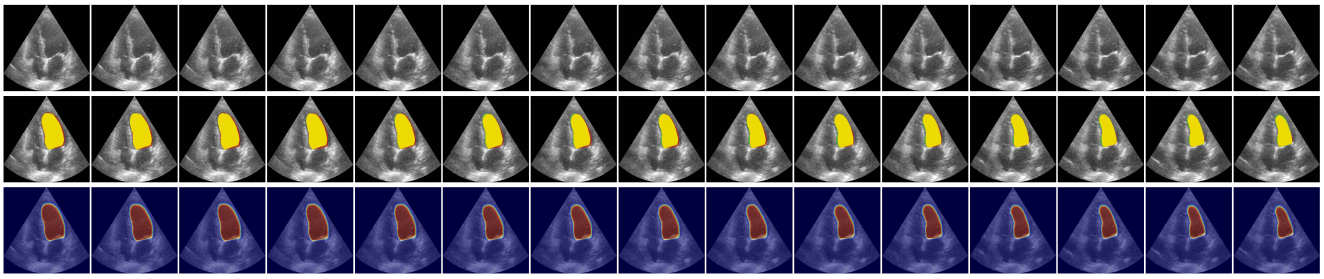
(a) Iteration 10, Epoch 1



(b) Iteration 20, Epoch 1



(c) Iteration 30, Epoch 1



(d) Iteration 1250, Epoch 19

Figure A.2. **Spatiotemporal training dynamics visualization.** Qualitative results on CAMUS patient 0225 (4-chamber view) at four training stages. For each stage (a-d), the top row displays input frames with segmentation masks, and the bottom row shows the corresponding feature activation maps. The progression from (a) to (d) illustrates the transition from diffused spatial activations and coarse masks to localized feature representations and temporally consistent segmentations throughout the cardiac cycle.

surrogate metric is calculated independently for apical 4-chamber and 2-chamber views without biplane fusion.

By utilizing this area-based approach, our method provides a direct and robust pathway from automated single-view echocardiographic segmentation to actionable diagnostic metrics, enabling efficient and clinically relevant cardiac assessment.

E. Temporal Consistency in Video Segmentation

Temporal consistency in video segmentation is defined as the stability of predicted masks across consecutive frames. While optical flow-based warping is a standard approach for quantifying such stability, its efficacy is constrained in echocardiography due to significant speckle noise, which degrades motion estimation accuracy. To address this, the

temporal matching error $\mathcal{E}_{tme}$ is employed, defined as the mean absolute difference between the temporal Dice evolution of predicted masks M and ground truth masks G:

$$\mathcal{E}_{tme} = \frac{1}{T-1} \sum_{t=1}^{T-1} |\text{Dice}(M_t, M_{t-1}) - \text{Dice}(G_t, G_{t-1})| \tag{A.1}$$

where T denotes the total number of frames. This metric quantifies the deviation of predicted temporal dynamics from the reference anatomical motion.