# Myocardial Scar and Edema Segmentation with Partial Labels and LGE-Guided Rescue

Xinyi Gao[1]✉, Yuming Zhao[1], Andrew Crean[2], and Fumie Costen[1]

[1] Faculty of Electrical and Electronic Engineering, University of Manchester, Manchester, UK
xinyi.gao@manchester.ac.uk
[2] Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

**Abstract.** Multicentre cardiac MR datasets may independently lack imaging sequences and annotations. We propose APL-Base, which masks losses for unobserved targets and predicts nested scar, pathology parent, and myocardial wall targets. Nested anatomical projection (NAP) restricts pathology predictions to myocardial support, while Component-based Scar Rescue (CSR) converts predictions from an LGE-only expert into residual scar proposals within the fixed parent. On 15 official validation cases, the full pipeline achieved scar and edema Dice scores of 0.6722 and 0.7109. In a five-fold, patient-level out-of-fold (OOF) evaluation of 220 patients, NAP reduced scar HD95 by 11.22 mm but decreased sensitivity, whereas rescue increased scar Dice by approximately 0.02. Among the 80 patients with both pathology labels, this scar gain was offset by lower edema overlap and did not improve equally weighted joint Dice. These results show that anatomical projection and LGE proposal rescue provide complementary boundary control and scar recovery. They also indicate that parent and proposal coverage, rather than selector complexity, limit further improvement.

**Keywords:** Myocardial pathology segmentation · Partial supervision · Nested anatomical projection · Component-based scar rescue · Multisequence CMR

## 1 Introduction

Late gadolinium enhancement (LGE), T2-weighted imaging, and cine bSSFP provide complementary information about myocardial scar, edema, and cardiac anatomy. The MyoPS benchmark formalised joint myocardial pathology segmentation using these sequences, and subsequent alignment work supported multisequence analysis [8, 2]. The CARE 2026 cohort introduces heterogeneous supervision across centres. Only two centres provide complete pathology and anatomy annotations. The remaining centres provide scar annotations with incomplete anatomy annotations or scar annotations alone. A missing sequence and a missing label convey different information and must therefore be represented separately.

RFNet and MyoPS-Net support incomplete modality sets [3, 12], while methods for incomplete class annotations restrict training to observed targets [17, 14, 16, 15, 7]. APL-Base addresses both requirements with a multibranch MyoPS-Net backbone. Modality indicators gate unavailable inputs, whereas separate annotation indicators activate losses only for observed targets. The model learns nested scar, pathology parent, and myocardial wall targets.

At inference, some pathology predictions extend beyond the myocardium. Myocardial delineation and shape priors can suppress anatomically implausible cardiac masks [1, 6, 10, 11]. Topological interactions provide another constraint [5]. Nested anatomical projection (NAP) applies the same predicted myocardial support to the scar and pathology parent predictions. Categorical export then assigns the remaining voxels in the parent to edema. This construction guarantees that scar is contained within the parent and that the scar and edema labels are mutually exclusive.

Projection removes remote false-positive predictions but can delete true scar where myocardial support is incomplete. LGE methods use ordered region constraints, multicentre deep learning, or auxiliary scar prediction [13, 4, 6, 9]. CSR uses an independent LGE expert to propose residual scar components inside the fixed pathology parent. Comparisons among learned selection, all-proposal rescue, and oracle selection distinguish failures of myocardial support and proposal coverage from errors in component classification.

We evaluate the cumulative pipeline using official validation and patient-level OOF predictions. Fixed-backbone ablations examine annotation masking and modality gating. Mechanism, robustness, strictly nested, and seed analyses locate the remaining errors and test the stability of our interpretation.

Our contributions are:

1. APL-Base represents modality and annotation availability separately, preserves the distinction between missing and negative targets, and learns nested pathology targets from heterogeneous centres.
2. NAP constrains scar and parent predictions to common myocardial support. Lesion analysis quantifies both false-positive suppression and sensitivity loss.
3. CSR combines LGE proposals with component selection. Comparisons of all-proposal, learned, and oracle policies identify limitations arising from myocardial support, proposal coverage, and selection.

## 2 Method

### 2.1 Partial supervision and nested targets

For each 2D training sample $i$, $\alpha_i^m \in \{0, 1\}$ records whether modality $m \in \{\text{bSSFP}, \text{LGE}, \text{T2}\}$ is available, while $\beta_i^c \in \{0, 1\}$ records whether target $c \in \{s, p, w, \text{LV}, \text{RV}\}$ is annotated. Both indicators are inherited from the source patient, and $\Omega_i$ denotes the pixel domain of the sample. Missing images are filled with zeros, and modality-specific features are gated by $\alpha_i^m$. Annotation

indicators are not supplied to the network and determine only loss eligibility. An annotated empty mask is therefore distinct from a missing target.

Let $G_i^s$, $G_i^e$, $G_i^w$, $G_i^{\mathrm{LV}}$, and $G_i^{\mathrm{RV}}$ denote the available scar, edema excluding scar, released myocardial wall, LV, and RV regions, respectively. When edema is annotated, the pathology parent is $G_i^p = G_i^s \cup G_i^e$; otherwise, parent supervision is masked rather than replaced by scar. In fully annotated cases, $G_i^s \subseteq G_i^p \subseteq G_i^w$. The pathology parent is an algorithmic target and does not imply biological containment of scar within edema or vice versa.

APL-Base uses a MyoPS-Net backbone with five levels, 2D inputs, and a base width of 64 [12]. It also includes an anatomy branch and separate LGE scar and T2 parent branches. Each available sequence is normalised separately. The internal anatomy U-Net produces a three-class softmax $\pi_i^a$ over background (0), myocardial wall (1), and LV (2). The hard input to both pathology encoders is the detached single-channel map $a_i(v) = \alpha_i^{\mathrm{bSSFP}} \arg\max_{k\in\{0,1,2\}} \pi_{i,k}^a(v)$; RV is coded as 0 in this internal map. At corresponding decoder scales, each pathology branch also fuses features from the opposite encoder. These features remain differentiable during Phase I of training. Separate binary anatomy heads produce $p_i^w$, $p_i^{\mathrm{LV}}$, and $p_i^{\mathrm{RV}}$. The reported outputs are

$$(p_i^s, p_i^p, p_i^w, p_i^{\mathrm{LV}}, p_i^{\mathrm{RV}}) = f_\Theta(x_i, \boldsymbol{\alpha}_i), \qquad p_i^c : \Omega_i \to [0,1].$$

Let $\ell_{i,c}$ be the sum of Dice loss and binary cross-entropy loss for sample $i$ and target $c$. All cases with parent annotations in the released cohort contain T2; therefore, annotation availability alone determines eligibility for supervised loss. Phase I uses unit weights for every target within its group. For a target group $\mathcal{C}_0$,

$$\begin{aligned} \mathcal{L}(\mathcal{C}_0) &= \frac{\displaystyle\sum_{i\in\mathcal{B}} \sum_{c\in\mathcal{C}_0} \beta_i^c \, \ell_{i,c}}{\max\left(1, \displaystyle\sum_{i\in\mathcal{B}} \sum_{c\in\mathcal{C}_0} \beta_i^c\right)}, \\ \mathcal{L}_{\mathrm{avail}} &= \lambda_{\mathrm{path}} \mathcal{L}(\{s,p\}) + \lambda_{\mathrm{anat}} \mathcal{L}(\{\mathrm{LV}, w, \mathrm{RV}\}). \end{aligned} \tag{1}$$

The denominator counts active supervision terms across samples and targets. The group loss is zero when no such term is present. Dice and binary cross-entropy losses are computed over pixels, and all configurations use the same Dice smoothing constant of $10^{-6}$.

For Phase I, let $\mathcal{I}_{\mathrm{miss}} = \{i \in \mathcal{B} : \beta_i^s = 1,\ \beta_i^p = 0\}$ denote samples in the current batch whose source patient has a scar annotation but no complete parent target. We penalise scar probabilities that exceed the corresponding parent probabilities using

$$\mathcal{L}_{\mathrm{inc}} = \frac{1}{|\mathcal{I}_{\mathrm{miss}}|} \sum_{i\in\mathcal{I}_{\mathrm{miss}}} \frac{1}{|\Omega_i|} \sum_{v\in\Omega_i} \left[p_i^s(v) - p_i^p(v)\right]_+, \tag{2}$$

where $[z]_+ = \max(z, 0)$. The term is set to zero when $\mathcal{I}_{\mathrm{miss}}$ is empty. Phase I minimises $\mathcal{L}_{\mathrm{phase1}} = \mathcal{L}_{\mathrm{avail}} + \lambda_{\mathrm{inc}} \mathcal{L}_{\mathrm{inc}}$, with $\lambda_{\mathrm{inc}} = 0.25$.

Phase I updates all parameters using data from seven centres for 200 epochs at $10^{-4}$, with $\lambda_{\text{path}} = 1$ and $\lambda_{\text{anat}} = 0.5$. Phase II uses the two fully annotated centres for 50 epochs at $10^{-5}$. It disables the inclusion loss and updates only the T2 encoder, decoder, and parent head. Because the scar pathway is frozen, only the parent loss produces parameter updates. Both phases use a batch size of 8 and 100 iterations per epoch. Phases I and II denote optimisation periods, whereas Stages 1, 2, and 3 in Fig. 1 denote the framework stages.

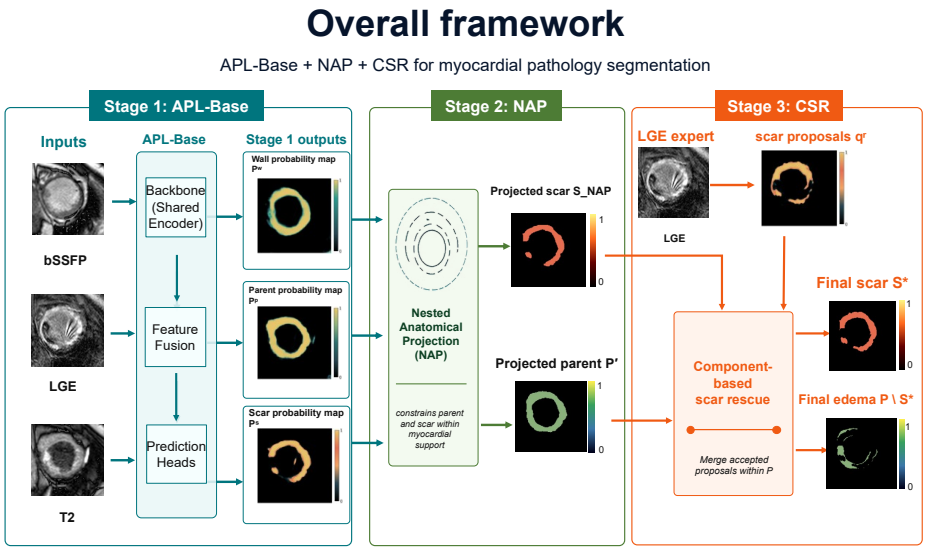


**Fig. 1.** Overview. Stage 1 uses APL-Base to predict myocardial wall, pathology parent, and scar probability maps. Stage 2 applies NAP to constrain pathology to myocardial support. Stage 3 uses LGE-derived component proposals for scar rescue and exports mutually exclusive scar and edema masks.

### 2.2 Anatomical projection and categorical export

Let $p^s$, $p^p$, and $p^w$ denote the predicted probabilities of scar, pathology parent, and myocardial wall. Let $\mathcal{D}_r^{2\text{D}}(A)$ denote binary dilation applied independently to each slice of $A$ with a $(2r+1) \times (2r+1)$ square structuring element. NAP constructs

$$\begin{aligned} M &= \mathcal{D}_1^{2\text{D}}\big(\{v : p^w(v) > 0.50\}\big), \\ S^{\text{NAP}} &= M \cap \{v : p^s(v) > 0.50\}, \\ P &= M \cap \big(\{v : p^p(v) > 0.50\} \cup \{v : p^s(v) > 0.50\}\big). \end{aligned} \tag{3}$$

Thus $S^{\text{NAP}} \subseteq P$.

For any final scar mask $S^\star \subseteq P$ produced by NAP or rescue, categorical export is

$$Y^s = S^\star, \qquad Y^e = P \setminus S^\star. \tag{4}$$

All rescue variants preserve $P$ and modify only its scar partition; a voxel added to scar is therefore removed from edema.

### 2.3 Component-based Scar Rescue (CSR)

CSR uses an independently trained LGE expert to produce the scar probability $q^s$. The expert is a U-Net with five levels, a base width of 64, and one LGE input channel. It produces a softmax over background and scar and uses the same $192 \times 192$ preprocessing and scar loss as APL-Base. For official validation, the expert is fitted using all 220 training cases with scar annotations. In the OOF analysis, each patient is evaluated by an expert trained without that patient's fold.

We form

$$M_{\text{out}} = \mathcal{D}_2^{\text{2D}}(\{v : p^w(v) > 0.35\}) \cup M,$$
$$Q = \{v : q^s(v) \geq 0.35\} \cap M_{\text{out}}.$$

and let $\{C_k\}_{k=1}^K$ be the components of $Q$ under 3D 6-connectivity. Strict and inclusive inequalities follow the locked exporter. No stored probability equals 0.35 or 0.50, so this distinction does not change the reported masks.

A proposal is labelled positive if it recovers at least one reference scar voxel missed by $S^{\text{NAP}}$ and at least 25% of its voxels overlap reference scar. A class-balanced logistic regressor assigns a selection score $r_k \in (0, 1)$ to each standardised 15-feature vector. The features comprise component size and occupied slice count; mean and maximum scar confidence from the LGE expert and APL-Base; and mean and maximum LGE percentile rank per voxel. They also include mean parent probability and overlap with parent probability above 0.50; mean wall probability; and overlap with the core and outer myocardial supports. The remaining features are the mean signed distance and minimum absolute distance to the myocardial boundary. The regressor uses $C = 1$, the LBFGS solver, a maximum of 1000 iterations, and a fixed random state.

All-proposal rescue merges every proposal component with $S^{\text{NAP}}$. Learned rescue retains only components with $r_k \geq 0.50$. Both policies then remove components smaller than three voxels under 3D 6-connectivity and intersect the result with the fixed parent $P$. Cleanup precedes the final intersection, and no additional cleanup is applied, consistent with the locked exporter. The thresholds, radii, 25% labelling rule, and three-voxel cleanup were defined a priori from anatomical considerations. They were fixed before processing the 15 official validation cases, and none was selected from the validation outcomes. Replacing 0.50 with $\tau$ gives the variants used to assess threshold sensitivity.

*Selector fitting protocols.* For official validation, APL-Base and the LGE expert are fitted using all 220 training cases, and the selector is refitted using all 1376

training components. The 15 server cases remain held out. For OOF evaluation, each patient is assessed by base models trained without that patient. The corresponding components are scored by a selector fitted without that patient's component rows. The fold models used to generate rows for selector training are not retrained within a second outer loop. Thus, the evaluation separates patients but is not fully nested.

## 3 Experiments

### 3.1 Data, metrics, and statistical analysis

The CARE training cohort [18, 12, 2] contains 220 patients from seven centres. All 220 patients have scar annotations, 80 have edema annotations, and 185 have myocardial wall annotations. Centres B and C comprise the 80 patients with complete pathology annotations, whereas centres A, E, F, G, and H comprise 140 patients with partial annotations. Because centre, sequence availability, and annotation availability covary, comparisons between centre groups are descriptive rather than causal. The official validation cohort contains 15 cases with all modalities. The OOF analysis uses five fixed patient folds. Each fold trains on 176 patients and evaluates the remaining 44, with the same folds used for all OOF comparisons and training ablations. In Sect. 2.1, $i$ indexes a 2D training sample; throughout this section, evaluation and resampling are performed at the patient level.

The multisequence model uses $192 \times 192$ 2D inputs, with each available sequence normalised separately as described in Sect. 2.1. Training uses Adam, a batch size of 8, and 100 iterations per epoch. Phase I runs for 200 epochs with a learning rate of $10^{-4}$ and weight decay of $5 \times 10^{-4}$. Phase II runs for 50 epochs at $10^{-5}$. Cosine annealing uses $T_{\max} = 50$ and a minimum learning rate of $10^{-6}$. We use OOF checkpoints from the final epoch.

The independent missing-labels-as-background (MBG) baseline is a separately trained 2D U-Net with five levels and the same five output targets. It fills unavailable modalities with zeros but retains all target channels in the loss, thereby coding unobserved labels as background. It uses neither NAP nor proposal rescue. Because its backbone configuration and supervision differ from those of APL-Base, MBG is an independent pipeline comparator rather than a controlled masking ablation.

Both official validation and OOF evaluation compare categorical scar with $G_i^s$ and categorical edema with $G_i^e$. OOF metrics are Dice, precision, sensitivity, and HD95 in millimetres, computed using physical voxel spacing. The patient is the unit of inference and resampling. Tables report patient means, and 95% confidence intervals use 2000 bootstrap replicates with a fixed seed. Paired comparisons use two-sided Wilcoxon signed-rank tests, with Holm correction applied within each reported analysis family. Two empty masks are assigned a Dice score of 1 and an HD95 of 0. If only one mask is empty, the Dice score is 0 and HD95 is undefined. Undefined HD95 values are excluded from distance averages. All

OOF comparisons use the same fixed folds. Morphological settings, probability thresholds, and the primary selector threshold were fixed before official validation. The selector refitted using all data was used only for the 15 server cases withheld from training.

*Joint overlap analysis for scar and edema.* For the 80 patients with both pathology labels, we evaluate both categorical outputs for each patient. For a predicted scar partition $S_i \subseteq P_i$,

$$J_i(S_i) = \frac{1}{2}\left[\mathrm{Dice}(S_i, G_i^s) + \mathrm{Dice}(P_i \setminus S_i, G_i^e)\right]. \tag{5}$$

We report the patient mean of $J_i$ and paired differences between methods.

Lesion recall and deletion use reference scar components under 6-connectivity, and boundary strata use distance to the predicted myocardial wall. The primary selector threshold was fixed at 0.50 before the equally weighted joint Dice analysis, threshold analysis, and selection of qualitative examples.

### 3.2 Official validation

**Table 1.** Official validation results for an independent U-Net baseline that treats missing labels as background (MBG) and four cumulative stages of the proposed pipeline. All methods were evaluated on the same 15 cases by the challenge server, which returned Dice and HD. The MBG baseline changes the backbone, input configuration, and annotation loss, so it is not a controlled ablation.

| Method | Scar Dice ↑ | Scar HD (mm) ↓ | Edema Dice ↑ | Edema HD (mm) ↓ |
|---|---|---|---|---|
| Raw APL-Base | 0.6416 | 29.9506 | 0.6952 | 34.8896 |
| APL-Base + NAP | 0.6452 | 26.0341 | 0.6974 | 29.1286 |
| APL-Base + NAP + all-proposal rescue | 0.6713 | **24.1916** | 0.7105 | **28.2074** |
| APL-Base + NAP + learned rescue | **0.6722** | 24.2551 | **0.7109** | 28.2391 |
| Independent MBG U-Net baseline | 0.6581 | 28.6261 | 0.6826 | 44.1125 |

On official validation, NAP affected boundary accuracy more than overlap by removing anatomically implausible predictions. Rescue produced the main improvement in Dice. The nearly identical results for all-proposal and learned rescue show that component selection added little once suitable proposals were available (Table 1).

### 3.3 OOF performance matched to the official pipeline

NAP primarily improved boundary accuracy. Scar HD95 decreased by 11.22 mm, while sensitivity decreased by 0.036. Both rescue policies increased scar Dice by approximately 0.02. In the matched cohort of 80 patients, this scar gain was offset by lower edema overlap and did not improve joint Dice. Learned rescue differed little from all-proposal rescue and primarily reduced the number of accepted components.

**Table 2.** OOF performance using the official pipeline definitions. Panel A reports patient means; scar metrics use 220 patients and edema metrics use 80 patients. Panel B reports mean scar Dice for the B/C group with complete annotations and the A/E/F/G/H group with partial annotations. Because centre, sequence availability, and annotation availability covary, this comparison is descriptive.

A. Absolute performance

| Method | Scar Dice | Scar HD95 (mm) | Edema Dice | Edema HD95 (mm) |
|---|---|---|---|---|
| Raw APL-Base | 0.4512 | 35.9648 | **0.4169** | 22.5166 |
| APL-Base + NAP | 0.4626 | 24.7438 | 0.4073 | **21.5759** |
| NAP + all-proposal rescue | 0.4813 | 24.8503 | 0.3897 | 21.6284 |
| NAP + learned rescue | **0.4825** | **24.5350** | 0.3903 | 21.9300 |

B. Scar Dice by annotation group

| Group | $n$ | Raw APL-Base | $+NAP$ | + all-proposal rescue | + learned rescue |
|---|---|---|---|---|---|
| Complete B/C | 80 | 0.5921 | 0.5978 | 0.6100 | **0.6112** |
| Partial A/E/F/G/H | 140 | 0.3707 | 0.3853 | 0.4078 | **0.4089** |

**Table 3.** Paired scar and edema analysis of the same 80 OOF patients with complete pathology annotations. Panel A reports patient means. Equally weighted joint Dice is defined for each patient in Eq. 5. Panel B reports paired mean differences; positive values favour the second method. The improvement, decline, and tie counts for joint Dice denote numbers of patients. The three joint Dice tests form a separate Holm family.

A. Absolute performance on the matched cohort

| Method | Scar Dice | Edema Dice | Joint Dice | Pathology parent Dice |
|---|---|---|---|---|
| APL-Base + NAP | 0.5978 | 0.4073 | **0.5025** | 0.7057 |
| NAP + all-proposal rescue | 0.6100 | 0.3897 | 0.4998 | 0.7057 |
| NAP + learned rescue | **0.6112** | 0.3903 | 0.5007 | 0.7057 |

B. Paired scar and edema Dice changes

| Comparison | $\Delta$ scar | $\Delta$ edema | $\Delta$ joint Dice | Joint Dice I/D/T | $p_{\text{Holm}}$ |
|---|---|---|---|---|---|
| NAP → all-proposal rescue | 0.0122 | −0.0176 | −0.0027 | 37/42/1 | 1.000 |
| NAP → learned rescue | 0.0134 | −0.0170 | −0.0018 | 38/40/2 | 1.000 |
| All-proposal rescue → learned rescue | 0.0012 | 0.0006 | 0.0009 | 25/8/47 | 0.031 |

The parent mask is identical across rescue policies.

### 3.4 APL-Base training ablations and supporting controls

BG, AM, MG, Inc, and P2 denote background coding, availability masking, modality gating, Phase I inclusion, and matched Phase II training. Panel A isolates the effects of inclusion and Phase II. Panel B reports native MyoPS-Net SSL as the mean ± SD across three seeds and nnU-Net-style controls using one seed. These controls are not part of the official pipeline. They support transfer of the mechanism rather than superiority across backbones.

**Table 4.** Selected fixed-backbone configurations. Panel A gives OOF patient means (seed 3407). Scar metrics use $n = 220$; edema metrics use $n = 80$. Parent/joint use $n = 80$ and wall uses $n = 185$. Panel B gives supporting controls on one held-out 44-patient fold, which are not directly ranked against Panel A.

*A. Five-fold OOF component ablation*

| Configuration | Scar Dice | Edema Dice | Parent Dice | Joint Dice | Wall Dice |
|---|---|---|---|---|---|
| BG-NoGate | 0.4608 | 0.3461 | 0.6493 | 0.4598 | 0.7753 |
| AM-NoGate | 0.4584 | 0.3750 | 0.6807 | 0.4778 | 0.7708 |
| AM+MG | 0.4580 | 0.3767 | 0.6816 | 0.4833 | 0.7708 |
| AM+MG+Inc | 0.4640 | 0.3811 | 0.6808 | 0.4874 | 0.7708 |
| AM+MG→P2 | 0.4326 | 0.4211 | 0.7135 | 0.5003 | 0.7640 |
| Full APL-Base | 0.4512 | 0.4169 | 0.7145 | 0.5045 | 0.7640 |

*B. Supporting held-out backbone controls*

| Control | Scar Dice | Edema Dice | Mean Dice | Mean HD95 (mm) |
|---|---|---|---|---|
| Native MyoPS-Net SSL | $0.229 \pm 0.015$ | $0.621 \pm 0.021$ | $0.425 \pm 0.018$ | $47.011 \pm 5.746$ |
| nnU-Net-style MBG | 0.562 | 0.700 | 0.631 | 9.600 |
| nnU-Net-style observed-label | 0.564 | 0.700 | 0.632 | 9.806 |
| nnU-Net-style + soft projection | 0.550 | 0.701 | 0.625 | 8.895 |

**Table 5.** Compact mechanism summary. Panel A reports the overall lesion measures and the two deletion strata with highest risk. Deletion rates are pooled over reference lesions that overlap the raw APL-Base prediction before NAP. Panel B reports the parent and proposal coverage terms needed to locate the oracle bottleneck.

*A. NAP lesion behaviour*

| Measure | Raw | NAP |
|---|---|---|
| Lesion recall | 0.8245 | 0.8090 |
| False-positive lesions per patient | 9.445 | 6.773 |
| Deletion of small lesions ($\leq 100$ mm$^3$) | N/A | 24.56% |
| Deletion beyond 3 mm from support | N/A | 77.78% |

*B. Rescue coverage and oracle*

| Quantity | Value |
|---|---|
| Pathology parent recall | 52.3% |
| Feasible proposal recall | 17.1% |
| Learned rescue Dice | 0.4825 |
| Training target oracle Dice | 0.4862 |

### 3.5 Mechanism analysis

We report lesion recall, false-positive lesions per patient, and deletion in the two highest-risk strata. These strata comprise lesions no larger than 100 mm$^3$ and lesions more than 3 mm from predicted myocardial support. NAP reduced false-positive lesions from 9.445 to 6.773 per patient, while lesion recall changed from 0.8245 to 0.8090 (Table 5). It deleted 24.56% of small lesions and 77.78% of lesions remote from predicted support.

None of the tested acceptance thresholds $\tau \in \{0.0, 0.1, \ldots, 0.9\}$ exceeded NAP in joint Dice. At the prespecified $\tau = 0.50$, learned and all-proposal rescue accepted 2.141 and 6.750 candidate components per patient, respectively. These counts precede cleanup and intersection with the parent. A rescue-created false-positive component was defined as final scar disjoint from both NAP scar and reference scar. Learned and all-proposal rescue produced 0.827 and 1.736 such components per patient, respectively. Among reassigned voxels in the 80 patients

with complete pathology annotations, 41.0% were reference scar, 28.7% were reference edema, and 30.3% were outside the reference pathology parent.

For the 212 patients with residual reference scar missed by $S^{\mathrm{NAP}}$, Table 5 reports macro-averaged patient recall within $P$ and within both $Q$ and $P$. The pooled feasible ratio was 12.83% when all residual voxels formed a common denominator. The training target oracle applies the same cleanup and parent constraint but accepts reference-positive proposals.

Learned rescue was only 0.0037 Dice below the training target component oracle. However, feasible proposal recall was 17.1% when macro-averaged across patients and 12.83% when pooled. The main bottleneck therefore precedes component classification.

*Supporting robustness analyses.* Scar Dice was stable across thresholds from 0.40 to 0.60 (0.4619–0.4626) but varied across 13 wall and radius configurations (0.4434 to 0.4652). This difference indicates dependence on morphology. In a strictly nested protocol with three training folds, one calibration fold, and one test fold, learned selection reduced reference-negative candidate components from 5.686 to 1.105 per patient. Dice remained essentially unchanged from accepting all proposals. Absolute Dice is not directly comparable with the primary OOF protocol, which uses four training folds. Across three APL-Base seeds with the LGE expert fixed, learned rescue exceeded all-proposal rescue in scar, edema, and joint Dice. Raw scar Dice was more sensitive to the seed than edema Dice, with sample SDs of 0.0233 and 0.00234, respectively.

Figure 2 illustrates these mechanisms and the tradeoff between scar and edema within the fixed parent.

## 4 Discussion

NAP removes anatomically implausible predictions but can delete true scar outside the fixed myocardial support. LGE proposals recover residual scar only within that parent. Reassignment from edema to scar therefore explains improved scar overlap, reduced edema overlap, and unchanged joint Dice.

Learned selection reduced accepted proposals and false-positive components but added little overlap beyond accepting all proposals. Its small oracle gap, together with low feasible residual scar coverage, indicates that the selector is mainly a conservative diagnostic and that myocardial support and proposal coverage remain the primary targets for improvement.

*Limitations.* Supporting controls were limited to one 44-patient fold. The nnU-Net-style controls used one seed, were not official implementations, and omitted the full five-fold OOF and CSR protocol. External validation was absent, and official validation included only 15 cases. Fixed thresholds and morphology lack prospective validation. Primary OOF used one APL-Base realisation per fold. The strictly nested analysis used fewer training folds, and the three-seed analysis fixed the LGE expert. Centre, sequence, and annotation availability covary; CSR was not assessed as a standalone lesion detector.

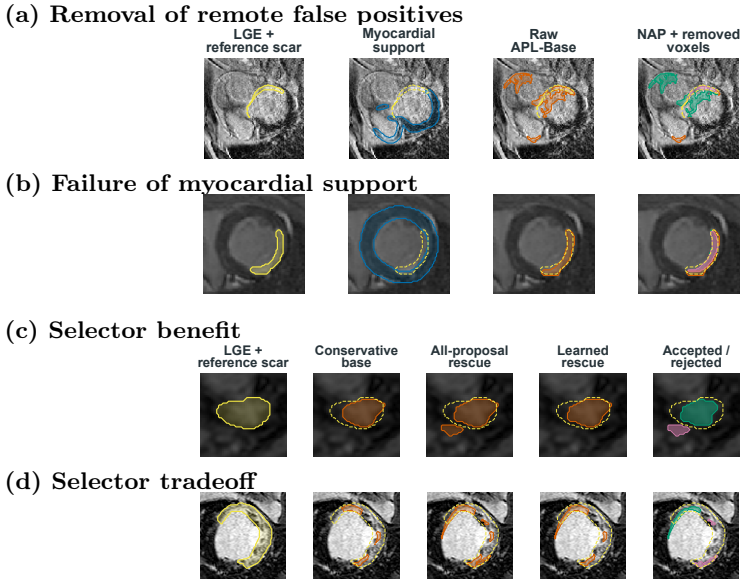


**Fig. 2.** Qualitative OOF examples: (a) removal of remote false positives, (b) loss of true scar because of incomplete support, (c) selective proposal acceptance, and (d) the tradeoff between scar and edema within a fixed parent. The examples illustrate mechanisms, not prevalence.

## 5 Conclusion

APL-Base supports heterogeneous supervision without treating unobserved targets as negatives. NAP suppresses anatomical false positives, while LGE proposals recover residual scar. Learned selection filters conservatively but adds little overlap. Myocardial support and proposal coverage remain the main bottlenecks.

**Acknowledgments and Disclosure.** The authors thank the CARE-MyoPS organisers for providing the data and evaluation framework and declare no competing interests.

## References

1. Bernard, O., Lalande, A., Zotti, C., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
2. Ding, W., Li, L., Qiu, J., et al.: Aligning multi-sequence CMR towards fully automated myocardial pathology segmentation. IEEE Trans. Med. Imaging **42**(12), 3474–3486 (2023)
3. Ding, Y., Yu, X., Yang, Y.: RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 3975–3984 (2021)

4. Fahmy, A.S., Neisius, U., Chan, R.H., et al.: Three-dimensional deep convolutional neural networks for automated myocardial scar quantification in hypertrophic cardiomyopathy: a multicenter multivendor study. Radiology **294**(1), 52–60 (2020)
5. Gupta, S., Hu, X., Kaan, J., et al.: Learning topological interactions for multi-class medical image segmentation. In: European Conference on Computer Vision. pp. 701–718. Springer (2022)
6. Jani, V.P., Ostovaneh, M., Chamera, E., et al.: Deep learning for automatic volumetric segmentation of left ventricular myocardium and ischaemic scar from multislice late gadolinium enhancement cardiovascular magnetic resonance. Eur. Heart J. Cardiovasc. Imaging **25**(6), 829–838 (2024)
7. Li, H., Luo, X., Fu, J., et al.: PL-Seg: Partially labeled abdominal organ segmentation via classwise orthogonal contrastive learning and progressive self-distillation. Medical Image Analysis **108**, 103885 (2026). https://doi.org/10.1016/j.media.2025.103885
8. Li, L., Wu, F., Wang, S., et al.: MyoPS: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. Medical Image Analysis **87**, 102808 (2023)
9. Margolis, I., Buoso, S., Kozerke, S.: Two-stage approach for myocardial scar and edema segmentation using synthetic multi-sequence MRI and auxiliary scar prediction. In: MICCAI CARE Challenge. pp. 112–123. Springer (2026)
10. Oktay, O., Ferrante, E., Kamnitsas, K., et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**(2), 384–395 (2017)
11. Painchaud, N., Skandarani, Y., Judge, T., et al.: Cardiac segmentation with strong anatomical guarantees. IEEE Trans. Med. Imaging **39**(11), 3703–3713 (2020)
12. Qiu, J., Li, L., Wang, S., et al.: MyoPS-Net: Myocardial pathology segmentation with flexible combination of multi-sequence CMR images. Medical Image Analysis **84**, 102694 (2023)
13. Rajchl, M., Yuan, J., White, J.A., et al.: Interactive hierarchical-flow segmentation of scar tissue from late-enhancement cardiac MR images. IEEE Trans. Med. Imaging **33**(1), 159–172 (2013)
14. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Medical Image Analysis **70**, 101979 (2021)
15. Xie, Y., Zhang, J., Xia, Y., Shen, C.: Learning from partially labeled data for multi-organ and tumor segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 14905–14919 (2023)
16. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1195–1204 (2021)
17. Zhou, Y., Li, Z., Bai, S., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 10672–10681 (2019)
18. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 2933–2946 (2018)