Efficient KernelSHAP Explanations for Patch-based 3D Medical Image Segmentation

Ricardo Coimbra Brioso*, Giulio Sichili*, Damiano Dei‡, Nicola Lambri‡§, Pietro Mancosu†‡, Marta Scorsetti†‡, and Daniele Loiacono*

*Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
†Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy
‡Radiotherapy and Radiosurgery Department, IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy
§Scuola di Specializzazione di Fisica Medica, Università Degli Studi di Milano, Milan, Italy
Email: ricardo.brioso@polimi.it, daniele.loiacono@polimi.it

***Abstract*—Perturbation-based explainability methods such as KernelSHAP provide model-agnostic attributions but are typically impractical for patch-based 3D medical image segmentation due to the large number of coalition evaluations and the high cost of sliding-window inference. We present an efficient KernelSHAP framework for volumetric CT segmentation that restricts computation to a user-defined region of interest and its receptive-field support, and accelerates inference via patch logit caching, reusing baseline predictions for unaffected patches while preserving nnU-Net's fusion scheme. To enable clinically meaningful attributions, we compare three automatically generated feature abstractions within the receptive-field crop: whole-organ units, regular FCC supervoxels, and hybrid organ-aware supervoxels, and we study multiple aggregation/value functions targeting stabilizing evidence (TP/Dice/Soft Dice) or false-positive behavior. Experiments on whole-body CT segmentations show that caching substantially reduces redundant computation (with computational savings ranging from 15% to 30%) and that faithfulness and interpretability exhibit clear trade-offs: regular supervoxels often maximize perturbation-based metrics but lack anatomical alignment, whereas organ-aware units yield more clinically interpretable explanations and are particularly effective for highlighting false-positive drivers under normalized metrics.**

***Index Terms*—Explainability, Semantic Segmentation, Shapley Values, nnU-Net, 3D Medical Imaging**

I. Introduction

Deep learning models have become the de-facto standard for medical image segmentation, enabling robust delineation of anatomical structures and pathological targets across modalities and clinical workflows. In radiotherapy treatment planning, in particular, accurate segmentation of organs-at-risk and target volumes directly impacts dose optimization and safety. Despite their strong empirical performance, deep learning segmentation models remain difficult to audit: errors can be subtle, spatially localized, and strongly influenced by contextual anatomy, acquisition artifacts, or model's biases. This motivates explainable AI (XAI) methods capable of providing faithful, clinically meaningful explanations of *why* a 3D segmentation model produced a given mask rather than another.

Compared to classification, explainability for dense prediction is substantially less consolidated. Recent surveys highlight open challenges specific to segmentation, including (i) how to define interpretable regions of analysis in volumetric data, (ii) how to aggregate voxel-level evidence into robust scores, and (iii) how to preserve anatomical structure in explanations [1]. Gradient-based methods (e.g., segmentation-adapted Grad-CAM variants) are computationally efficient but may be layer-dependent and hard to interpret in multi-structure settings [2]–[4]. Perturbation-based approaches, on the other hand, offer model-agnostic explanations and can be aligned with clinically meaningful units, but they are often prohibitively expensive in 3D—especially for patch-based inference pipelines such as nnU-Net [5]. In particular, Shapley-value methods (e.g., KernelSHAP [6]) require evaluating a value function over many feature coalitions, which naively entails thousands of forward passes over highly overlapping sliding-window patches.

In this work, we propose an efficient perturbation-based explainability framework for *patch-based 3D segmentation* by adapting KernelSHAP to volumetric CT data under two key principles: (1) *localize* computation to a user-defined region of interest (ROI) and its receptive-field support, and (2) *reuse* baseline computations whenever perturbations do not affect a given sliding-window patch. We further investigate how the definition of interpretable units (organs vs. geometric vs. organ-aware supervoxels) and the choice of scalar aggregation function (logit-weighted true/false positive scores, Dice, and Soft Dice) impact the qualitative appearance and the perturbation-based faithfulness of the resulting explanations.

II. Related Works

A recent survey by Gipiškis et al. [1] highlights two broad families that are most relevant to our setting: gradient-based methods and perturbation-/region-based methods. Compared to classification, segmentation explainability remains less explored and still lacks consolidated best practices, especially for *(i)* defining clinically meaningful regions of analysis, *(ii)* aggregating voxel-level evidence into robust scores, and *(iii)* ensuring explanations remain interpretable in a structured anatomical setting.

A. Gradient-based Attribution

Gradient-based explanations extend saliency mechanisms to dense prediction. In segmentation, Grad-CAM-style methods have been adapted (e.g., Seg-Grad-CAM) to focus explanations on selected predicted regions rather than the full

image [2]. Hasany et al. [3] show that the explanatory signal is highly layer-dependent (encoder bottleneck vs. decoder/output), motivating multi-layer inspection to capture how contextual anatomy shapes the final mask. More recently, FM-G-CAM [4] addresses the single-class limitation by integrating gradients from multiple classes, which is particularly relevant in multi-organ settings where inter-structure dependencies are clinically meaningful.

B. Perturbation- and Region-based Methods

Perturbation-based methods explain model behavior through systematic input modifications and their impact on the segmentation output. Early approaches include occlusion/sensitivity analysis to measure segmentation stability under localized masking or transformations [7]. In cardiac MRI, feature ablation studies further suggest that predictions for a target structure may depend on surrounding anatomical context [8], reinforcing the need for explanations that can disentangle target evidence from contextual cues.

However, classic local surrogate approaches such as LIME [9] often struggle in segmentation due to instability and poor semantic alignment of perturbed regions. To mitigate this, Knab et al. [10] propose using anatomically meaningful superpixels derived from SAM to better match the perturbation units to realistic regions. SLICE [11] targets variance reduction via stabilized sampling and superpixel selection, improving consistency across runs. Other works explicitly model context: Grid Saliency [12] separates object evidence from contextual surroundings (at higher computational cost), while U-Noise-style approaches learn input-dependent noise masks and can be regularized for spatial smoothness [13], [14]. Complementarily, MiSuRe [15] optimizes a minimally sufficient region that preserves segmentation quality (e.g., Dice) while shrinking the explanation, providing a fidelity-oriented notion of "necessary" evidence (in contrast to Grad-CAM/RISE [16]).

Overall, perturbation-based explanations are attractive because they are model-agnostic and can be aligned with clinically meaningful units, but their scalability and faithfulness remain open challenges. Chrabaszcz et al. [17] propose Agg^{2}Exp, aggregating voxel-level gradient attributions in 3D, and empirically highlight that gradient aggregation can provide more scalable and faithful insights than computationally heavy perturbation schemes in complex multi-class volumes.

III. Adapting KernelSHAP for Efficient 3D Segmentation Explainability

In this section, we present our explainability framework for interpreting predictions of patch-based 3D segmentation by adapting KernelSHAP [6] to volumetric CT data. Given the high dimensionality of voxel-level features, we estimate Shapley attributions over *interpretable units* (supervoxels or organs) and restrict the analysis to a *region of interest* (ROI). Our pipeline is defined by the following components: (i) ROI definition and corresponding receptive-field (RF) support for spatially localized computation; (ii) a partition of the volume into M interpretable units; (iii) a perturbation operator that removes selected units by masking; (iv) an ROI-restricted scalar *value function* used by KernelSHAP to score coalitions; (v) an efficient coalition-evaluation scheme for patch-based predictors based on sliding-window patch caching.

A. ROI Definition and Receptive-Field Support

Our explanations are conditioned on a baseline segmentation that the user seeks to interpret. In practice, an expert selects a subset of the model prediction for the target class (e.g., a connected component or a clinically relevant subregion), thus defining an ROI mask $R(x) \in \{0, 1\}$. Then, we compute the minimal axis-aligned *cubic* bounding box B enclosing all voxels such that $R(x) = 1$. Since inference is performed patch-wise (e.g., sliding-window evaluation), predictions inside B depend only on a finite neighborhood of input voxels. We therefore define an enlarged box B_{RF} by dilating B to conservatively cover the input support that can influence predictions within B (e.g., using the patch radius along each axis). All perturbations and forward passes are restricted to B_{RF}, while voxels outside are kept fixed. This reduces computation without compromising faithfulness of attributions within the ROI.

B. Interpretable Unit Definition

KernelSHAP operates on M binary features; therefore, the RF-supported crop B_{RF} is partitioned into interpretable units $\{U_j\}_{j=1}^{M}$ with $U_j \subseteq B_{\mathrm{RF}}$. These units are generated automatically inside B_{RF} by either algorithmic tessellation or by intersecting that tessellation with organ masks from TotalSegmentator; they are not manually drawn for each explanation. We favor anatomy-aware variants because alignment with organ boundaries generally makes the resulting attributions easier to interpret clinically, even if no partition can perfectly match the model's latent internal representation. We consider the following three partitions.

Full Organs. Each anatomical structure segmented by TotalSegmentator [18] is treated as one unit (background excluded). This maximizes semantic interpretability and minimizes M, which typically improves KernelSHAP stability under a fixed sampling budget.

Regular. We tessellate B_{RF} using a Face-Centred Cubic (FCC) lattice defined in physical coordinates (mm) [19]. Each voxel is assigned to its nearest lattice center, yielding approximately isotropic supervoxels in real-world units even under anisotropic voxel spacing. A single scale parameter S (mm) controls granularity. This organ-agnostic partition serves as a geometrically regular baseline (see an example in Fig.1).

Hybrid. To combine geometric regularity with anatomical constraints, we start from the FCC tessellation and subdivide each FCC cell according to organ labels obtained from TotalSegmentator [18]. Voxels belonging to different organs within the same FCC cell are assigned distinct unit IDs (background excluded), preventing cross-structure mixing while preserving physical-space isotropy (see an example in Fig.1). This design is conceptually related to organ-splitting strategies

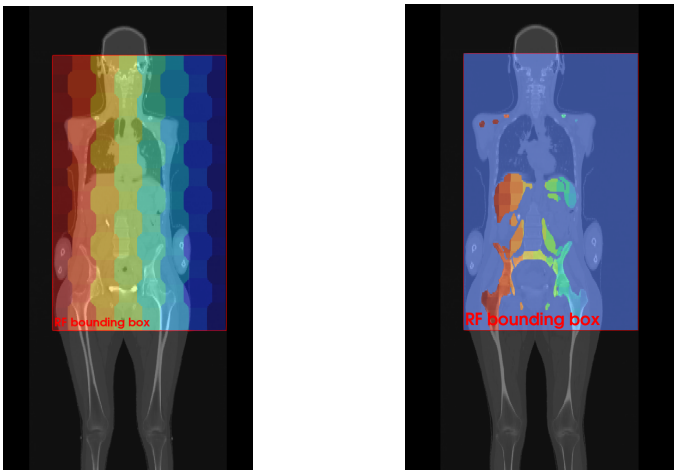

Fig. 1. Coronal examples of (left) Regular FCC supervoxels and (right) Hybrid (organ-aware) FCC supervoxels within the RF-supported crop B_{RF}. Regular FCC provides uniform geometry-driven units, whereas the Hybrid variant further splits FCC cells along organ boundaries to improve anatomical alignment.

such as [17], with the key difference that the underlying grid is FCC rather than voxel-space cubes.

C. Perturbation Operator

Let $\mathbf{m} \in \{0,1\}^M$ denote a coalition mask, where $m_j = 1$ keeps unit U_j and $m_j = 0$ removes it. We implement hard masking in intensity space: for each removed unit, all voxels in U_j are set to a masking value b. In CT we use $b = -1024$ HU, i.e., an air-equivalent value at the lower end of the Hounsfield scale that is commonly used for background/outside-body voxels. This gives a physically meaningful "removal" baseline without introducing spurious high-density tissue and is kept fixed throughout both attribution generation and evaluation. Let $u(x) \in \{1,\dots,M\}$ denote the unit index of voxel $x \in B_{\mathrm{RF}}$. The perturbed input is

$$X^{(\mathbf{m})}(x) = \begin{cases} X(x), & m_{u(x)} = 1, \\ b, & m_{u(x)} = 0, \end{cases} \qquad x \in B_{\mathrm{RF}}. \tag{1}$$

Voxels outside B_{RF} are never perturbed and are copied from the original input. Although hard masking may introduce out-of-distribution artifacts, it provides an explicit "removal" intervention, making the resulting Shapley attributions straightforward to interpret as contributions of the original signal within each unit.

D. Aggregation Strategies for KernelSHAP Scoring

KernelSHAP requires a scalar value function $v(\mathbf{m})$ for each coalition. Let f denote the segmentation network, producing voxel-wise logits $Z^{(\mathbf{m})} = f(X^{(\mathbf{m})})$. For a target class t, define the perturbed hard prediction

$$P_{\mathbf{m}}(x) = \mathbb{1}\left[\arg\max_c Z_c^{(\mathbf{m})}(x) = t\right], \tag{2}$$

and the baseline (unperturbed) prediction P_0 as the special case $\mathbf{m} = \mathbf{1}$. All scores are computed *within the ROI* defined by $R(x)$; let $|R| = \sum_x R(x)$ and let $z_t^{(\mathbf{m})}(x)$ be the target-class logit. We compute the following scores to evaluate each coalition $\mathbf{m}$.

True Positive Score. It rewards logit support for voxels correctly predicted as positive in the baseline segmentation within the ROI:

$$S_{\mathrm{TP}}(\mathbf{m}) = \frac{1}{|R|} \sum_x R(x)\, \mathbb{1}[P_{\mathbf{m}}(x) = 1 \wedge P_0(x) = 1]\; z_t^{(\mathbf{m})}(x). \tag{3}$$

False Positive Score. It penalizes logit support for voxels incorrectly predicted as positive in the baseline segmentation within the ROI:

$$S_{\mathrm{FP}}(\mathbf{m}) = -\frac{1}{|R|} \sum_x R(x)\, \mathbb{1}[P_{\mathbf{m}}(x) = 1 \wedge P_0(x) = 0]\; z_t^{(\mathbf{m})}(x). \tag{4}$$

Dice Score. It exploits Dice similarity to quantify the agreement between perturbed and baseline predictions:

$$S_{\mathrm{Dice}}(\mathbf{m}) = \frac{2\,\big|\,(P_{\mathbf{m}} \odot R) \cap (P_0 \odot R)\,\big|}{\|P_{\mathbf{m}} \odot R\|_1 + \|P_0 \odot R\|_1 + \varepsilon}, \tag{5}$$

where $\odot$ denotes element-wise masking by the ROI and $\varepsilon > 0$ avoids division by zero.

Soft Dice Score. It rewards logit support on voxels consistent with the baseline segmentation and penalizes newly introduced positives within the ROI:

$$S_{\mathrm{Soft}}(\mathbf{m}) = \frac{1}{|R|} \sum_x R(x)\, \mathbb{1}[P_{\mathbf{m}}(x) = 1]\; w(x)\, z_t^{(\mathbf{m})}(x). \tag{6}$$

where $w(x) = 2P_0(x) - 1$, i.e., +1 if voxel x is in the baseline segmentation and -1 otherwise. In other words, $w(x)$ acts as a signed agreement label: logits on voxels already present in the baseline mask are rewarded, whereas logits on newly activated voxels are penalized, making S_{Soft} a soft surrogate of Dice agreement with the baseline segmentation.

E. Efficient Coalition Evaluation via Sliding-Window Patch Caching

KernelSHAP requires evaluating the value function over many coalitions, hence many forward passes of the segmentation model on perturbed inputs. For patch-based predictors using sliding-window inference (e.g., `nnU-Net`), a naive implementation would recompute logits for every overlapping patch at every sampled coalition, which is typically prohibitive even when restricting computation to B_{RF}. To reduce redundant computation, we cache baseline patch logits and reuse them whenever a coalition does not affect a given patch.

Baseline cache construction. We first run a single sliding-window inference on the unperturbed input restricted to B_{RF}. Let $\mathcal{S}$ denote the set of sliding-window patch extractors (slices) used by the inference routine. For each patch location $s \in \mathcal{S}$, we store the predicted patch logits in a dictionary keyed by the patch spatial coordinates (slice key). To limit GPU memory usage, cached logits are stored in CPU memory.

Cached inference under perturbations. For a coalition $\mathbf{m}$, we define a binary perturbation mask $\mathbf{M}^{(\mathbf{m})}$ over B_{RF} such

that $\mathbf{M}^{(\mathbf{m})}(x) = 1$ if voxel x is masked to the masking value $b = -1024$ (see Section III-C). During sliding-window fusion, for each patch slice $s \in \mathcal{S}$ we check whether the coalition affects that patch region. If $\sum_{x \in s} \mathbf{M}^{(\mathbf{m})}(x) = 0$ (no perturbed voxels in the patch), we retrieve the cached baseline logits (cache hit); otherwise, we recompute logits by forwarding the perturbed patch only (cache miss). Cached and recomputed patch logits are fused exactly as in standard `nnU-Net` inference (Gaussian-weighted accumulation followed by normalization), ensuring that caching does not alter the aggregation scheme.

Expected savings and trade-offs. Let h be the cache hit rate, i.e., the fraction of patches retrieved from the cache for a given coalition. Under an idealized constant per-patch cost model and ignoring overheads, the forward-pass time scales with $(1 - h)$, yielding an approximate speedup of $1/(1 - h)$ relative to recomputing all patches. In practice, the realized gain is lower due to cache lookups, intersection tests, and CPU–GPU transfers, but caching remains most effective when perturbations are spatially localized (as typically occurs when interpretable units correspond to anatomically constrained supervoxels within a focused ROI). The main trade-off is memory, since the cache stores logits for all sliding-window patches covering B_{RF}.

IV. Experimental Design

This section details the experimental setup used to compute KernelSHAP attribution maps for the nnU-Net segmentation model and to quantitatively evaluate their faithfulness.

A. Data and Segmentation Task

As testbed for our explainability framework, we use a clinical dataset of whole-body CT images for segmentation of lymph nodes and spleen in the context of Total Marrow and Lymphoid Irradiation (TMLI) treatment planning. The full dataset comprises **40 patients** affected by hematological malignancies and treated with non-myeloablative TMLI. CT scans were acquired in free-breathing and without contrast using a clinical scanner (slice thickness of 5 mm). Volumes have an average shape of $237 \times 512 \times 512$ and anisotropic spacing of approximately $5.0\ \mathrm{mm} \times 1.17\ \mathrm{mm} \times 1.17\ \mathrm{mm}$.

The **segmentation target** is defined as the union of **lymph nodes (CTV_LN)** and **spleen (CTV_Spleen)**, which are challenging due to high inter-patient variability and complex geometries.

The dataset is split into **32 training** and **8 test** volumes. We train a 3D nnU-Net [5] segmentation model on the training set, following the setup described in [20].

On the 8 test volumes, we evaluated the proposed explainability framework using three different interpretable-unit definitions, i.e., Full Organs, Regular, and Hybrid (see Section III-B) and four aggregation/score functions, i.e., True Positive, False Positive, Dice, Soft Dice (see Section III-D). During attribution computation, nnU-Net test-time augmentation was disabled to ensure deterministic coalition evaluations, and inference was accelerated via patch caching (Section III-E).

B. KernelSHAP Convergence Validation and Sampling Budget

To ensure reliable Shapley attributions, we empirically validated the stability of the KernelSHAP approximation as a function of the sampling budget n (number of coalitions). KernelSHAP fits a weighted linear surrogate model to coalition evaluations of the nnU-Net value function. We increased n and monitored surrogate stability using: (i) *coefficient stability* (relative ℓ_1 change of attribution vectors between successive budgets), (ii) *local accuracy* (residual between the coalition value and the sum of attributions), (iii) *numerical stability* (condition number of the weighted design matrix), and (iv) *generalization of the surrogate* on held-out coalitions (MAE and R^2).

Based on this stability analysis, we selected a conservative budget of $\mathbf{n = 2000}$ coalitions for the *Regular* and *Hybrid* settings (typically involving several hundred features), while $\mathbf{n = 1000}$ was sufficient for *Full Organs* due to the much smaller feature space ($M \approx 9$).

C. Attribution Map Evaluation Metrics

We evaluate attribution maps primarily through **perturbation-curve faithfulness**, balancing computational feasibility in 3D with interpretability and direct alignment to the model behavior.

MoRF/LeRF perturbation protocol. For a given volume and configuration, interpretable units are ranked by their SHAP values. Starting from the unperturbed input, we iteratively remove (mask) units according to two complementary orderings: *Most-Relevant-First* (MoRF, descending SHAP) and *Least-Relevant-First* (LeRF, ascending SHAP). After each step, we recompute the model output and the corresponding scalar score using the same ROI-restricted aggregation/value function used for KernelSHAP (Section III-D) and the same masking baseline ($b = -1024$ HU; Section III-C). This yields two curves, $s_{\mathrm{MoRF}}(k)$ and $s_{\mathrm{LeRF}}(k)$, as a function of the amount of removed evidence. Therefore, the reported perturbation-curve metrics validate attribution faithfulness under the exact perturbation operator used to generate the explanations.

Area Over the Perturbation Curve (AOPC). From the MoRF curve, we compute AOPC as the average degradation in score caused by removing highly-ranked units [21]:

$$\mathrm{AOPC} = \frac{1}{K} \sum_{k=1}^{K} \big(s(0) - s_{\mathrm{MoRF}}(k) \big), \tag{7}$$

where $s(0)$ is the unperturbed score and K is the number of perturbation steps. Higher AOPC indicates that the attribution map successfully identifies units whose removal most strongly impacts the model score.

Area Between Perturbation Curves (ABPC). To quantify how well an attribution method separates important from unimportant evidence, we compute ABPC as the average gap between the LeRF and MoRF trajectories:

$$\mathrm{ABPC} = \frac{1}{K} \sum_{k=1}^{K} \big(s_{\mathrm{LeRF}}(k) - s_{\mathrm{MoRF}}(k) \big). \tag{8}$$

Higher ABPC indicates stronger discrimination: removing low-ranked units leaves the score comparatively intact, while removing high-ranked units rapidly degrades it.

Because different aggregation functions (and different volumes) can induce different score ranges, we also report normalized variants of AOPC/ABPC. For each curve, we rescale scores to $[0, 1]$ using the attainable range induced by the perturbation endpoints (analogous in spirit to normalized faithfulness metrics such as NAOPC [22]):

$$\tilde{s}(k) = \frac{s(k) - s_{\min}}{s_{\max} - s_{\min} + \epsilon}, \tag{9}$$

with $s_{\max} = s(0)$ and $s_{\min}$ given by the fully-perturbed score (all units removed), computed separately for each case, configuration, and aggregation function.

Finally, for interpretability across heterogeneous tessellations, perturbation curves are plotted against the fraction (or number) of interpretability units removed.

V. Results

This section reports the results obtained with the KernelSHAP attribution framework described in Section III. We first provide a qualitative inspection of attribution maps across different interpretable-unit definitions and aggregation functions. We then summarize the quantitative evaluation based on MoRF/LeRF perturbation-curve metrics. Finally, we report the computational gains obtained through patch caching. All results are computed on the eight validation volumes defined in the experimental design.

A. Qualitative Analysis of Attribution Maps

Visual inspection of the attribution maps highlights how the choice of interpretable units (supervoxels) and the value-function aggregation affect the resulting explanations. Figures 2–4 show representative examples for **volume 7**. Attribution magnitudes are generally small due to ROI restriction and score normalization. Both strongly positive values (units supporting the baseline prediction) and strongly negative values (units opposing it, e.g., by reducing Dice or increasing false positives) indicate high relevance.

Full Organs (Fig. 2). Attributions are constrained by organ boundaries. The TP-, Dice-, and Soft-Dice-based maps are visually consistent and emphasize regions near the ROI that support the baseline segmentation. In contrast, the FP-based map shows an approximately inverted pattern, highlighting regions whose presence contributes to spurious activations. This suggests that anatomically coherent regions may play a dual role, stabilizing correct predictions while also inducing false positives depending on local context.

Regular (FCC) supervoxels (Fig. 3). Due to the large and spatially uniform units, maps exhibit stronger local effects and sharper transitions. TP, Dice, and Soft Dice again emphasize evidence within or close to the ROI supporting the baseline output. The FP map reveals negative contributions concentrated within the ROI (spurious activations) and occasional positive contributions near boundaries, consistent with a mix

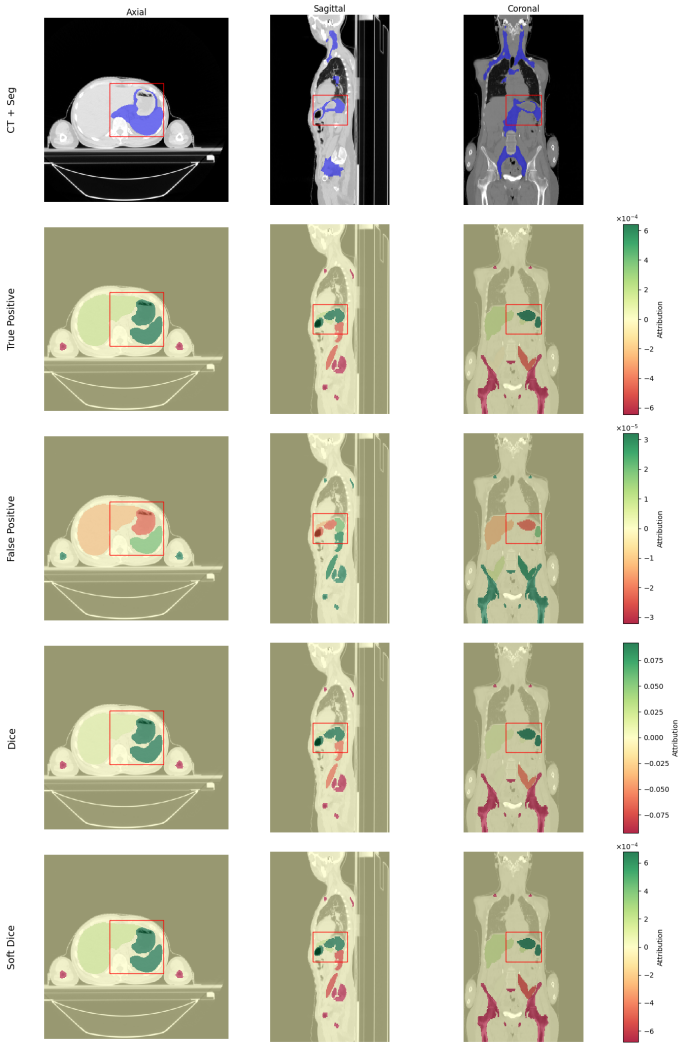

Fig. 2. Qualitative attribution maps for **Full Organs** (volume 7), comparing aggregation functions within the ROI.

of destabilizing and stabilizing effects. Overall, explanations appear noisier, which is expected from an organ-agnostic tessellation and the coarser effective resolution.

Hybrid (Organ-Aware FCC) supervoxels (Fig. 4). Hybrid units preserve organ semantics while enabling finer within-organ granularity. Relative to Full Organs, TP/Dice/Soft-Dice maps capture more detailed intra-organ variations. Importantly, FP-based attributions show marked heterogeneity inside single organs, where different subregions may contribute with opposite signs. This behavior is consistent with localized mechanisms leading to spurious predictions and may provide a more actionable explanation for debugging false positives.

B. Quantitative Evaluation using Perturbation Curves

Figures 5–7 report MoRF and LeRF perturbation curves (median $\pm$ IQR), while Table I summarizes ABPC/AOPC and their normalized variants averaged over the eight validation cases.

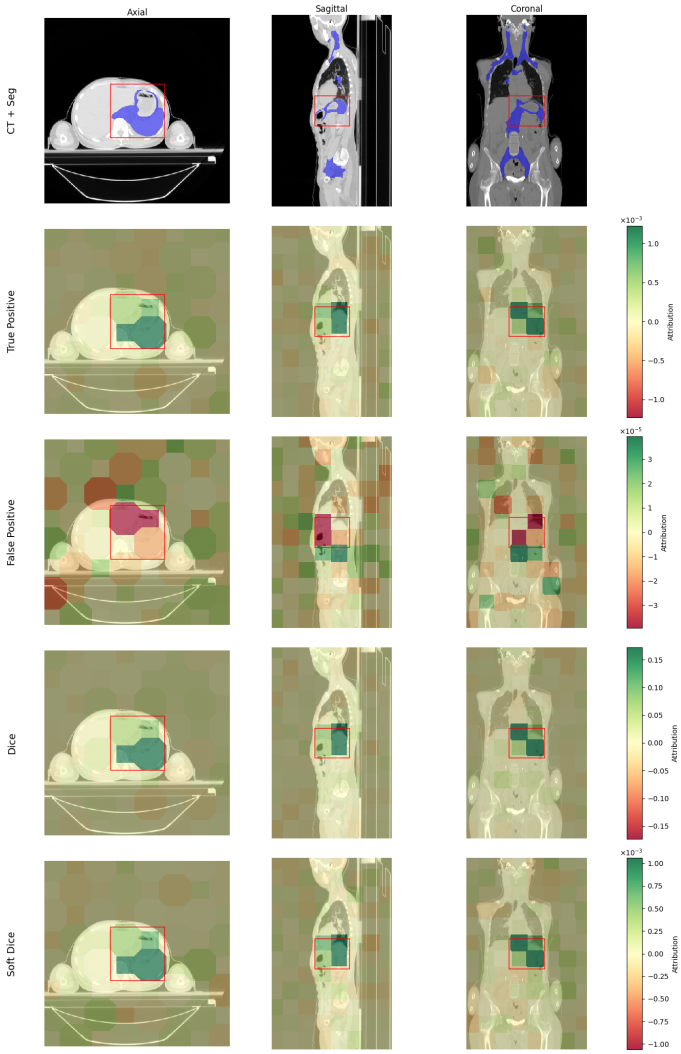

Fig. 3. Qualitative attribution maps for **Regular (FCC)** supervoxels (volume 7).

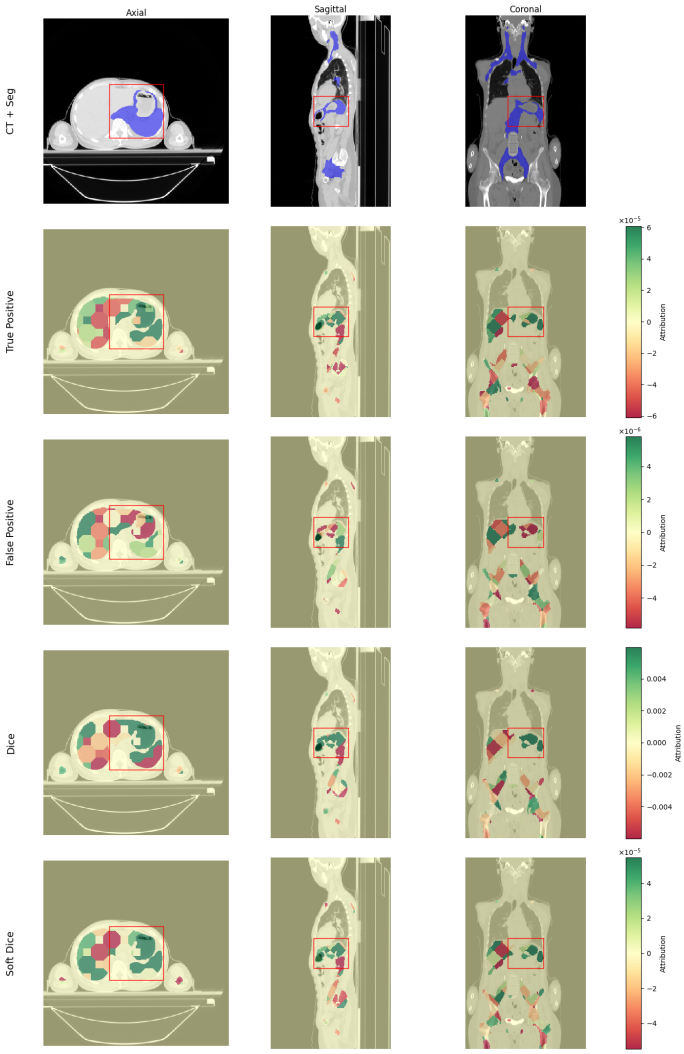

Fig. 4. Qualitative attribution maps for **Hybrid (Organ-Aware FCC)** supervoxels (volume 7).

Across TP, Dice, and Soft Dice aggregations, **Regular** supervoxels achieve the highest raw and normalized ABPC/AOPC values. These results are mainly due to two factors. First, the FCC tessellation is organ-agnostic and typically spans a *larger overall spatial support* than organ-constrained partitions, so that successive perturbations remove (or affect) a broader portion of the input volume. Second—and most critically—FCC units may *overlap the segmentation target* (i.e., voxels that are maximally correlated with the value function). Perturbing units that contain the target can induce a direct and substantial drop in TP/Dice/Soft Dice, which inflates MoRF–LeRF separability and, consequently, ABPC/AOPC (including their normalized variants). Therefore, cross-configuration comparisons based solely on perturbation metrics are not entirely fair, as they conflate attribution quality with feature definition and target overlap.

From an interpretability standpoint, this also highlights a practical limitation of Regular: attributing importance to units that effectively ”contain the answer” (the target) tends to yield explanations that are less clinically actionable, because they emphasize the model’s sensitivity to removing the target itself rather than revealing anatomically meaningful contextual drivers.

For the **False Positive** aggregation, **Hybrid** supervoxels obtain the best performance under normalized metrics (nABPC and nAOPC). This suggests that, at comparable granularity, combining anatomical constraints with within-organ partitioning improves the ability to localize subregions specifically responsible for spurious activations, compared to purely geometric FCC tessellations.

Finally, TP and Soft Dice show consistent trends across supervoxel types, whereas Dice (computed on binarized masks) tends to accentuate differences between configurations, likely due to thresholding effects and discrete changes in overlap.

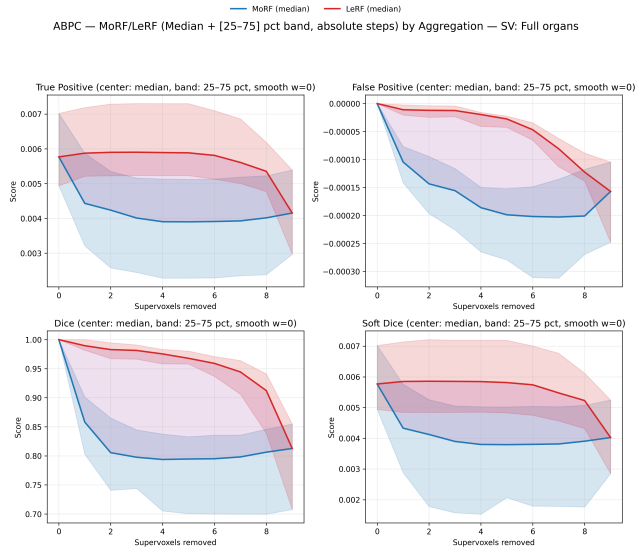

Fig. 5. MoRF and LeRF curves (median ± IQR) for **Full Organs**.

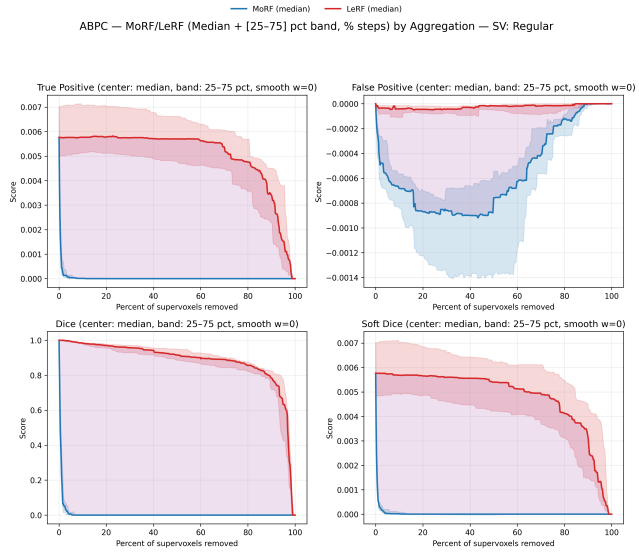

Fig. 6. MoRF and LeRF curves (median ± IQR) for **Regular (FCC)** supervoxels.

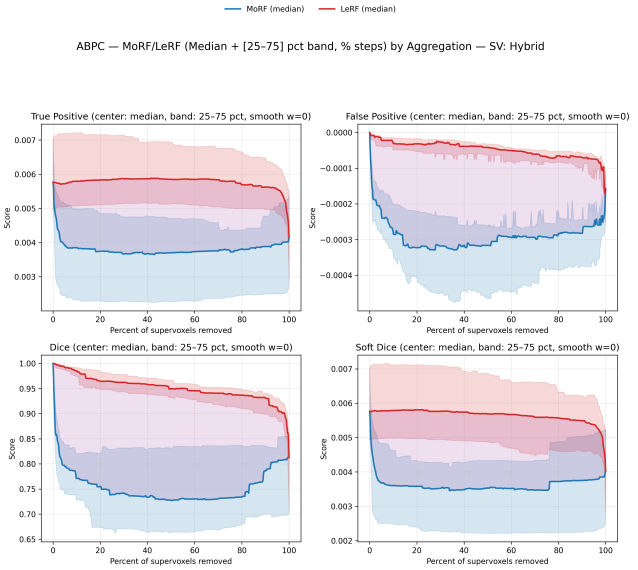

Fig. 7. MoRF and LeRF curves (median ± IQR) for **Hybrid** supervoxels.

TABLE I
SUMMARY OF PERTURBATION-CURVE METRICS (MEAN OVER 8 VALIDATION CASES). BEST RESULTS PER ROW ARE IN BOLD.

Metric	Aggregation	Full organs	Regular	Hybrid
ABPC	True Positive	1.576e-03	**4.980e-03**	2.102e-03
	False Positive	4.065e-04	**8.095e-04**	2.603e-04
	Dice	1.535e-01	**8.387e-01**	2.022e-01
	Soft Dice	1.524e-03	**4.753e-03**	2.005e-03
nABPC	True Positive	0.6776	**0.8341**	0.8112
	False Positive	0.5259	0.4827	**0.6653**
	Dice	0.5869	**0.8388**	0.6960
	Soft Dice	0.6227	**0.7787**	0.7574
AOPC	True Positive	1.565e-03	**5.751e-03**	2.158e-03
	False Positive	4.504e-04	**8.593e-04**	3.118e-04
	Dice	2.017e-01	**9.857e-01**	2.581e-01
	Soft Dice	1.737e-03	**5.811e-03**	2.291e-03
nAOPC	True Positive	0.7625	**0.9898**	0.8948
	False Positive	0.7091	0.5251	**0.8165**
	Dice	0.7745	**0.9858**	0.9002
	Soft Dice	0.7537	**0.9658**	0.9002

C. Computational Performance with Caching

Patch caching (Section III-E) substantially reduces redundant computation during coalition evaluation. Averaged over the eight validation cases, **Full Organs** achieves an average cache hit ratio of $32.4\% \pm 6.3\%$ with an inference time of $3.58\text{s} \pm 0.47\text{s}$ per sample, yielding a total runtime of **1h 01m 45s** for $n = 1000$ coalitions. **Regular (FCC)** shows lower cache reuse ($15.0\% \pm 1.4\%$) and higher inference time ($4.41\text{s} \pm 0.57\text{s}$), resulting in the longest total runtime (**2h 31m 47s**) for $n = 2000$. **Hybrid** preserves high cache reuse ($30.2\% \pm 5.2\%$) with inference time comparable to Full Organs ($3.57\text{s} \pm 0.43\text{s}$), leading to a total runtime of **2h 02m 36s** for $n = 2000$. These results confirm that caching is most effective when perturbations remain spatially localized relative to the sliding-window patch grid, as in organ-constrained or organ-aware unit definitions. Overall, **Hybrid** provides a favorable trade-off between explanation granularity (larger feature space requiring higher sampling budgets) and computational efficiency (high cache reuse).

VI. CONCLUSIONS

We presented an efficient perturbation-based explainability framework for patch-based 3D medical image segmentation by adapting KernelSHAP to volumetric CT. Our pipeline localizes coalition evaluations to a user-defined ROI and its receptive-field support, and accelerates nnU-Net sliding-window inference through patch logit caching that reuses baseline predictions whenever a coalition does not affect a patch. This substantially reduces redundant computation and makes Shapley-style attributions more practical in 3D settings.

Our results show that explanation meaning and quality are tightly coupled to the definition of *interpretable units* and the *aggregation/value function*. Perturbation-curve metrics (AOPC, ABPC, including normalized variants) highlighted clear trade-offs: Regular supervoxels often achieved the highest faithfulness due to their large spatial influence and frequent overlap with the target, but lacked anatomical alignment and may be less actionable clinically. Full Organs provided the most interpretable, stable high-level explanations with a minimal feature set, at the cost of limited granularity. Hybrid organ-aware supervoxels offered a compelling balance, preserving organ semantics while enabling finer within-organ resolution, and were particularly effective at exposing features associated with false positives under normalized metrics. In addition, the aggregation choice modulated the explanatory focus, emphasizing either stabilizing evidence (TP/Dice/Soft Dice) or destabilizing effects linked to spurious activations (FP), underscoring that different clinical questions require different objectives.

Several limitations remain. Validation was performed on only eight held-out volumes due to KernelSHAP's computational cost, so the reported trends should be interpreted as preliminary and confirmed on larger, more diverse cohorts. The approach also inherits KernelSHAP assumptions, including a linear surrogate and sensitivity to the perturbation strategy; hard masking (zero-out to -1024 HU) may introduce out-of-distribution artifacts. Moreover, we did not include clinician-centered evaluation, so quantitative faithfulness/stability metrics do not directly translate to clinical usefulness. Finally, fixed supervoxel strategies may bias explanations when units misalign with clinically meaningful boundaries.

Future work will investigate data-driven, boundary-adherent 3D supervoxels (e.g., 3D SLIC [23] or SEEDS [24], [25]) to improve anatomical adherence while characterizing cost–regularity trade-offs. On the efficiency side, more sophisticated forms of caching (e.g., caching frequently occurring coalitions in low-overlap patches, common with Full Organs) could further reduce runtime. We also plan to explore more realistic in-distribution perturbations, such as inpainting-based removal.

Acknowledgment

The authors disclose the use of OpenAI ChatGPT (version 5.2) to assist with English-language editing and linguistic refinement of the manuscript. The use of AI was limited to language polishing, while all technical and scientific content was produced and verified by the authors.

References

[1] R. Gipiškis, C.-W. Tsai, and O. Kurasova, "Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey," *ICT Express*, vol. 10, pp. 1331–1354, Dec. 2024.

[2] K. Vinogradova, A. Dibrov, and G. Myers, "Towards Interpretable Semantic Segmentation via Gradient-weighted Class Activation Mapping," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13943–13944, Apr. 2020. arXiv:2002.11434 [cs].

[3] S. N. Hasany, F. Mériaudeau, and C. Petitjean, "A Guided Tour of Post-hoc XAI Techniques in Image Segmentation," in *Explainable Artificial Intelligence* (L. Longo, S. Lapuschkin, and C. Seifert, eds.), (Cham), pp. 155–177, Springer Nature Switzerland, 2024.

[4] R. S. R. Silva and J. J. Bird, "FM-G-CAM: A Holistic Approach for Explainable AI in Computer Vision," Apr. 2024. arXiv:2312.05975 [cs].

[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, Feb. 2021. Publisher: Nature Publishing Group.

[6] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017. arXiv:1705.07874 [cs].

[7] M. J. Ankenbrand, L. Shainberg, M. Hock, D. Lohr, and L. M. Schreiber, "Sensitivity analysis for interpretation of machine learning based segmentation models in cardiac MRI," *BMC Medical Imaging*, vol. 21, p. 27, Feb. 2021.

[8] M. Ayoob, O. Nettasinghe, V. Sylvester, H. Bowala, and H. Mohideen, "Peering into the Heart: A Comprehensive Exploration of Semantic Segmentation and Explainable AI on the MnMs-2 Cardiac MRI Dataset," *Applied Computer Systems*, vol. 30, pp. 12–20, Jan. 2025.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Aug. 2016. arXiv:1602.04938 [cs].

[10] P. Knab, S. Marton, and C. Bartelt, "Beyond Pixels: Enhancing LIME with Hierarchical Features and Segmentation Foundation Models," Feb. 2025. arXiv:2403.07733 [cs].

[11] R. P. Bora, P. Terhorst, R. Veldhuis, R. Ramachandra, and K. Raja, "SLICE: Stabilized LIME for Consistent Explanations for Image Classification," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10988–10996, June 2024. ISSN: 2575-7075.

[12] L. Hoyer, M. Munoz, P. Katiyar, A. Khoreva, and V. Fischer, "Grid Saliency for Context Explanations of Semantic Segmentation," Nov. 2019. arXiv:1907.13054 [cs].

[13] T. Koker, F. Mireshghallah, T. Titcombe, and G. Kaissis, "U-Noise: Learnable Noise Masks for Interpretable Image Segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 394–398, Sept. 2021. arXiv:2101.05791 [cs].

[14] T. Okamoto, C. Gu, J. Yu, and C. Zhang, "Generating Smooth Interpretability Map for Explainable Image Segmentation," in *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*, (Nara, Japan), pp. 1023–1025, IEEE, Oct. 2023.

[15] S. N. Hasany, F. Mériaudeau, and C. Petitjean, "MiSuRe is all you need to explain your image segmentation," June 2024. arXiv:2406.12173 [cs].

[16] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," Sept. 2018. arXiv:1806.07421 [cs].

[17] M. Chrabaszcz, H. Baniecki, P. Komorowski, S. Płotka, and P. Biecek, "Aggregated Attributions for Explanatory Analysis of 3D Segmentation Models," Nov. 2024. arXiv:2407.16653 [cs].

[18] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023.

[19] P. Dardouillet, A. Benoit, E. Amri, P. Bolon, D. Dubucq, and A. Crédoz, "Explainability of Image Semantic Segmentation Through SHAP Values," in *ICPR-XAIE -26TH International Conference on Pattern Recognition 2-nd Workshop on Explainable and Ethical AI*, (Montreal, Canada), Aug. 2022.

[20] R. C. Brioso, D. Dei, N. Lambri, P. Mancosu, M. Scorsetti, and D. Loiacono, "Investigating gender bias in lymph-node segmentation with anatomical priors," in *Ethics and Fairness in Medical Imaging* (E. Puyol-Antón, G. Zamzmi, A. Feragen, A. P. King, V. Cheplygina, M. Ganz-Benjaminsen, E. Ferrante, B. Glocker, E. Petersen, J. S. H. Baxter, I. Rekik, and R. Eagleson, eds.), (Cham), pp. 151–160, Springer Nature Switzerland, 2025.

[21] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[22] J. Edin, A. G. Motzfeldt, C. L. Christensen, T. Ruotsalo, L. Maaløe, and M. Maistro, "Normalized AOPC: Fixing Misleading Faithfulness Metrics for Feature Attribution Explainability," May 2025. arXiv:2408.08137 [cs].

[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[24] M. V. d. Bergh, X. Boix, G. Roig, and L. V. Gool, "SEEDS: Superpixels Extracted via Energy-Driven Sampling," Sept. 2013. arXiv:1309.3848 [cs].

[25] C. Zhao, Y. Jiang, and T. C. Hollon, "Extending SEEDS to a Supervoxel Algorithm for Medical Image Analysis," Feb. 2025.