Geometric Shape Matching for Explainable and Accurate Medical Image Segmentation: A Post-Processing Refinement Framework

Ruhaib Muhammad[1] **Shivaram Velayutham**[1] **Jey Praveen Sivaraj**[1]

Dhiraj Gambhire[2] **Lakshminarayanan Subramanian**[3,4]

[1]Velai Health Analytics
[2]Dizzaroo Private Limited
[3]Velai Health LLC
[4]Courant Institute of Mathematical Sciences, New York University

Abstract

Deep learning models for medical image segmentation, while achieving remarkable performance, often produce anatomically implausible outputs that compromise clinical trust and adoption. We propose a novel inference-time refinement framework that leverages geometric shape matching against a curated library of high-quality organ segmentations to enhance TotalSegmentator predictions without requiring retraining or ground truth data. Our approach provides interpretable corrections by comparing predicted segmentations with anatomically plausible reference templates through a geometry-based matching framework. The framework operates as a modular post-processing layer, addressing TotalSegmentator's occasional anatomical hallucinations while maintaining compatibility with existing clinical workflows. Proof-of-concept experiments on liver segmentation using the CT-ORG dataset demonstrate an average 15% improvement in Dice scores for poor-performing segmentations. This work presents a promising direction for improving segmentation reliability in clinical deployment while preserving the interpretability required for medical applications.

1 Introduction

Medical image segmentation serves as a cornerstone for diagnosis, treatment planning, and surgical navigation across diverse clinical applications. Multi-atlas segmentation (MAS) is an important class of medical image segmentation, utilizing multiple labeled training images to capture anatomical variation [3]. Recent advances combine multi-atlas approaches with deep learning, where neural networks provide initial segmentations that are refined using atlas-based post-processing [4]. Modern approaches like TotalSegmentator [1] and MedSAM [5] demonstrate the potential of foundation models for universal medical image segmentation across diverse modalities and anatomical structures.

While deep learning has revolutionized automated segmentation with models like TotalSegmentator [1] achieving state-of-the-art performance, these systems remain susceptible to producing anatomically implausible outputs, particularly when encountering imaging artifacts, rare anatomical variants, or pathological conditions outside their training distribution.

TotalSegmentator demonstrates robust performance across 104 anatomical structures but occasionally produces anatomical inconsistencies or hallucinations [1]. Shape-based constraints have been

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

explored to improve segmentation reliability, but most approaches integrate these constraints within the model architecture, compromising interpretability [6]. Recent advances in medical AI have also highlighted the importance of detecting and correcting model hallucinations [10], particularly in clinical deployment scenarios where anatomical accuracy is paramount.

Healthcare professionals require understandable systems that enable validation and correction of automated predictions. Current approaches often operate as black boxes, making it difficult for clinicians to understand and trust segmentation outputs or identify when corrections are needed. We introduce a geometric shape matching framework that addresses these challenges through an interpretable inference-time enhancement approach. Our method leverages a carefully curated library of high-quality organ segmentations to refine and improve predictions from TotalSegmentator without requiring model retraining or access to ground truth data during inference.

Contributions include: (1) an inference-time post-processing framework for enhancing TotalSegmentator segmentations through geometric shape matching, (2) demonstration of segmentation improvements on challenging cases where TotalSegmentator produces poor initial predictions, and (3) a transparent, interpretable refinement process that enables clinical validation.

Comprehensive benchmarking datasets such as AMOS [7], CT-ORG [2], and WORD [8] have established standardized evaluation protocols for abdominal organ segmentation. We demonstrate the effectiveness of our approach on the CT-ORG dataset, comprising 140 CT scans with annotations for six organ classes (liver, lungs, bladder, kidneys, bones, and brain), which provides diverse imaging conditions from multiple medical centers [2]. In this early work, we primarily present results on liver segmentation but the approach is generalizable to other organ classes. In summary, our work specifically addresses TotalSegmentator's hallucinations through interpretable post-processing refinement that maintains clinical validation capabilities while leveraging established benchmarking datasets for comprehensive evaluation.

2 Methodology

Our approach is grounded in the observation that most organs exhibit consistent morphological characteristics across a population. This consistency suggests that the canonical shape of an organ can be captured using high-quality reference geometries. By leveraging these reference shapes, we guide and adjust segmentations to better match typical organ morphology, ensuring anatomical plausibility even in challenging cases.

In practice, we build upon TotalSegmentator, a state-of-the-art multi-organ segmentation model, which produces segmentations with varying Dice scores across different slices and organs (Figure 1). We observe that the highest-quality segmentations for a given organ collectively encode its canonical geometry. For slices where TotalSegmentator performs poorly, we draw on this set of high-quality reference geometries to refine the segmentation, effectively improving the Dice score and creating a modular post-processing layer that enhances the original predictions. Our approach draws inspiration from recent advances in shape-aware segmentation [11] and post-processing refinement techniques [12], while incorporating anatomical consistency principles [13] to ensure clinical validity.

2.1 Mathematical Formulation

Let X be a dataset of CT scans, and $x \in X$ represent a sample scan. For a specific organ i, let $T_i(x)$ denote the segmentation generated by TotalSegmentator, and s_i represent the ground truth segmentation for organ i in scan x.

From the training split of dataset X, we construct a reference set $G = \{g_1, g_2, \ldots, g_k\}$, where each g_j represents one of the top-k highest-quality segmentations:

$$G = \{g_j : \text{Dice}(g_j, s_j^{gt}) \geq 0.9, j = 1, 2, \ldots, k\} \quad (1)$$

Each reference g_j maintains metadata including its source scan and Dice score, enabling clinicians to inspect the quality and appropriateness of references used for refinement decisions. This transparency is crucial for clinical validation and trust.

2.2 Reference Selection Strategy

The success of our approach depends critically on the quality and representativeness of the reference library. We implement a systematic reference selection strategy that ensures anatomical diversity while maintaining high segmentation quality. From the training split of the CT-ORG dataset [2], we rank all TotalSegmentator predictions based on their Dice similarity coefficients with expert-validated ground truth annotations.

The selection process follows multiple criteria to ensure comprehensive anatomical coverage. We establish a minimum quality threshold of 0.9 Dice score, analyze morphological diversity to avoid redundancy, and validate that selected references represent different patients and imaging conditions. For liver segmentation, this process identified 10 high-quality references with Dice scores ranging from 0.944 to 0.947 (Figure 2).

2.3 Spatial Alignment and Geometric Matching

The geometric matching process begins with spatial normalization through bounding box extraction around each reference segmentation and target prediction. Reference segmentations are scaled to match target dimensions through the transformation:

$$\hat{g}_j = \mathcal{A}_j(g_j) = \text{Scale}(g_j, \frac{\text{dim}(B_T)}{\text{dim}(B_j)}) \tag{2}$$

where B_T and B_j represent the bounding boxes of the target and reference j respectively. This approach accommodates variations in slice thickness and spatial resolution across different CT acquisition protocols, making the framework robust to diverse imaging conditions.

The core matching algorithm operates at the slice level, computing similarity weights between target slice $T_i(x)_z$ and each aligned reference slice $\hat{g}_{j,z}$:

$$\begin{aligned}
w_{j,z} =& \alpha \cdot \text{AreaRatio}(T_i(x)_z, \hat{g}_{j,z}) && (3)\\
&+ \beta \cdot \text{NCC}(T_i(x)_z, \hat{g}_{j,z}) && (4)\\
&+ \gamma \cdot \text{CentroidSim}(T_i(x)_z, \hat{g}_{j,z}) && (5)\\
&+ \delta \cdot \text{OverlapSim}(T_i(x)_z, \hat{g}_{j,z}) && (6)
\end{aligned}$$

These metrics include area ratio comparison for size compatibility, normalized cross-correlation for structural similarity, centroid distance for spatial positioning, and overlap similarity for direct anatomical correspondence. Each metric provides interpretable insights into why specific references were selected for refinement.

2.4 Weighted Fusion and Quality Validation

The refinement process combines insights from multiple reference templates through weighted fusion. For each target slice, we identify suitable references based on similarity thresholds and create weighted combinations:

$$T_i'(x)_z = \begin{cases} \frac{\sum_{j:g_j \in G_z} w_{j,z} \cdot \hat{g}_{j,z}}{\sum_{j:g_j \in G_z} w_{j,z}} & \text{if } |G_z| > 0 \text{ and } \max(w_{j,z}) > 0.7 \\ T_i(x)_z & \text{otherwise} \end{cases} \tag{7}$$

where G_z represents the subset of references exceeding minimum similarity thresholds for slice z. Quality validation implements 3D consistency checks that evaluate coherence with adjacent anatomy:

$$C_z = \frac{1}{2} \left[\text{Dice}(T_i'(x)_z, T_i'(x)_{z-1}) + \text{Dice}(T_i'(x)_z, T_i'(x)_{z+1})\right] \tag{8}$$

Slices with consistency scores below 0.6 are reverted to original predictions to maintain segmentation integrity. This conservative approach prioritizes anatomical plausibility over aggressive refinement.

3 Experimental Setup and Results

We demonstrate our framework using the CT-ORG dataset [2], comprising 140 3D whole-body CT scans with expert manual annotations for six anatomical structures. The dataset exhibits wide variety in imaging conditions collected from various medical centers, ensuring generalizability of trained models. From the training split, we identified the top 10 TotalSegmentator liver segmentations with highest Dice scores (ranging from 0.944 to 0.947) to serve as our reference library G. We evaluated our framework on the complete test set of 30 sample scans from the CT-ORG test split.

Our evaluation focuses on poor-performing TotalSegmentator predictions from the test split with initially low Dice scores to demonstrate the enhancement capabilities. The framework achieved an average 15% improvement in Dice scores for poor-performing segmentations with initial scores below 0.6. Improvements ranged from a decrease of 5% to an increase of 25% depending on the degree of initial anatomical inconsistency. The framework's conservative design ensures minimal degradation even in cases where geometric matching is less effective.

Qualitative analysis revealed that improvements primarily occurred in regions where TotalSegmentator produced anatomical hallucinations. The framework successfully corrected boundary irregularities, eliminated spurious connections, and improved overall organ shape consistency by leveraging anatomically plausible reference templates (Figure 3). Detailed quantitative results are provided in Table 1.

4 Discussion and Conclusions

The complete transparency of our approach represents a significant advancement over black-box refinement methods, addressing critical needs in medical AI deployment where hallucination detection [10] and quality assessment are essential for clinical acceptance. Our geometric shape matching framework demonstrates promising capabilities for enhancing TotalSegmentator predictions through interpretable, inference-time refinement.

The interpretability of our framework manifests in several key aspects: (1) clinicians can visualize which reference templates were selected and their similarity scores, (2) the geometric metrics provide intuitive explanations for refinement decisions, (3) metadata preservation enables tracing refinements back to specific high-quality training examples, and (4) the conservative validation mechanism ensures anatomical plausibility is maintained. These features enable healthcare professionals to understand, validate, and trust the refinement process—a critical requirement for clinical adoption.

The approach successfully addresses anatomical inconsistencies while enabling clinical validation of automated corrections. The framework's modular design ensures compatibility with existing clinical workflows. By operating as a post-processing layer, it preserves current TotalSegmentator deployments while adding enhancement capabilities that address the model's occasional anatomical hallucinations.

Our approach makes specific assumptions that may not universally hold. The primary assumption is that anatomical organs exhibit sufficient morphological consistency within populations to enable effective reference-based correction. This assumption may not hold for patients with significant anatomical variants, congenital abnormalities, or extensive pathological conditions that fundamentally alter organ shape. We also assume that high-quality reference segmentations from the training population are representative of the test population's anatomical characteristics. Due to space constraints, we only present early results on liver segmentation on the CT-ORG dataset [2], but the approach is replicable on other organs and datasets.

Future directions include extending the framework to TotalSegmentator's full 104-structure capability, developing enhanced lesion detection through anatomically-aware organ boundary refinement, and systematically addressing model hallucinations across different segmentation architectures including nnU-Net [9], MedSAM [5], and emerging foundation models.

References

[1] Jakob Wasserthal et al. Totalsegmentator: Robust segmentation of 104 anatomical structures in ct images. *Radiology: Artificial Intelligence*, 5(4):e230024, 2023.

[2] Blaine Rister et al. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020.

[3] Juan Eugenio Iglesias et al. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015.

[4] Naga Venkata Annasamudram et al. Deep network and multi-atlas segmentation fusion for automated muscle group segmentation. *Journal of Medical Imaging*, 11(5):054005, 2024.

[5] Jun Ma et al. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[6] Christian Wachinger et al. Atlas-based under-segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pages 315–322, 2014.

[7] Yuanfeng Ji et al. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. *Advances in Neural Information Processing Systems*, 35:26239–26251, 2022.

[8] Xiangde Luo et al. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Medical Image Analysis*, 82:102642, 2022.

[9] Fabian Isensee et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[10] Anil Kumar Parchur et al. Automated hallucination detection for synthetic CT images in MR-only workflows. *Physics in Medicine & Biology*, 70(4):045015, 2025.

[11] Shuxin You et al. Shape-aware organ segmentation by predicting signed distance maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12565–12572, 2020.

[12] Haoyu Tang et al. Boundary-aware segmentation network for medical images. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 201–210, 2021.

[13] Cheng Huang et al. ASA: Learning anatomical consistency, sub-volume spatial relationships and fine-grained appearance for CT images. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 123–133, 2024.

A Appendix

Table 1: Detailed performance results of the geometric shape matching refinement applied to select CT-ORG test split samples with initially low-quality predictions (Dice < 0.6). The table reports the number of slices modified in the final scaled target segmentation and the corresponding absolute Dice score changes. Positive improvements are highlighted, while negative values indicate conservative refinements where the method prioritized anatomical plausibility.

Sample ID	Original Dice	Refined Dice	Net Improvement
28	0.4458	0.5683	+0.1225
29	0.3115	0.5403	+0.2288
30	0.2437	0.4197	+0.1760
31	0.5187	0.5299	+0.0113
32	0.4938	0.5884	+0.0947
33	0.2977	0.3046	+0.0070
34	0.5324	0.5376	+0.0052
35	0.5916	0.5663	-0.0253
36	0.5789	0.6063	+0.0274
37	0.5729	0.6605	+0.0876
38	0.4568	0.5147	+0.0579
40	0.5232	0.5565	+0.0332
41	0.4718	0.5244	+0.0526
42	0.3623	0.5614	+0.1991
43	0.3111	0.5232	+0.2121
44	0.5650	0.5915	+0.0265
45	0.3947	0.5543	+0.1596
47	0.3640	0.4591	+0.0941

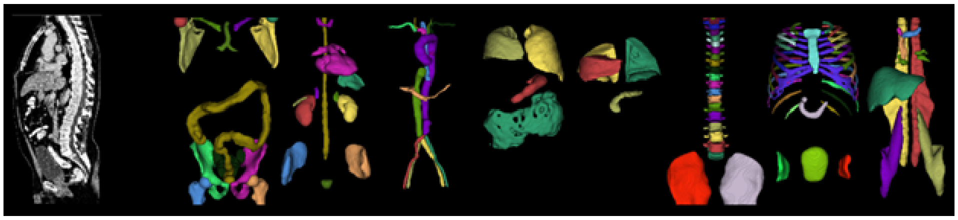
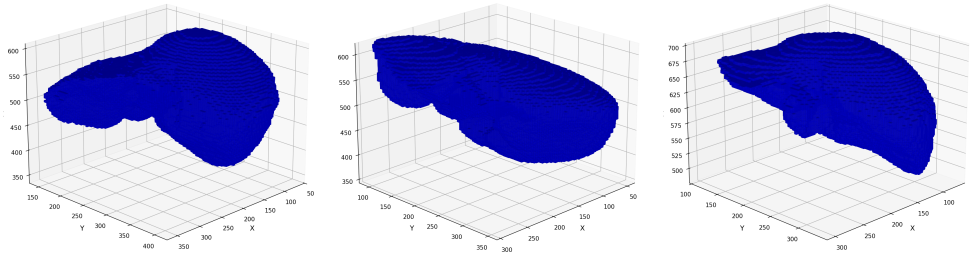


Figure 1: TotalSegmentator outputs for a sample CT scan, showing its multi-organ segmentation capabilities.

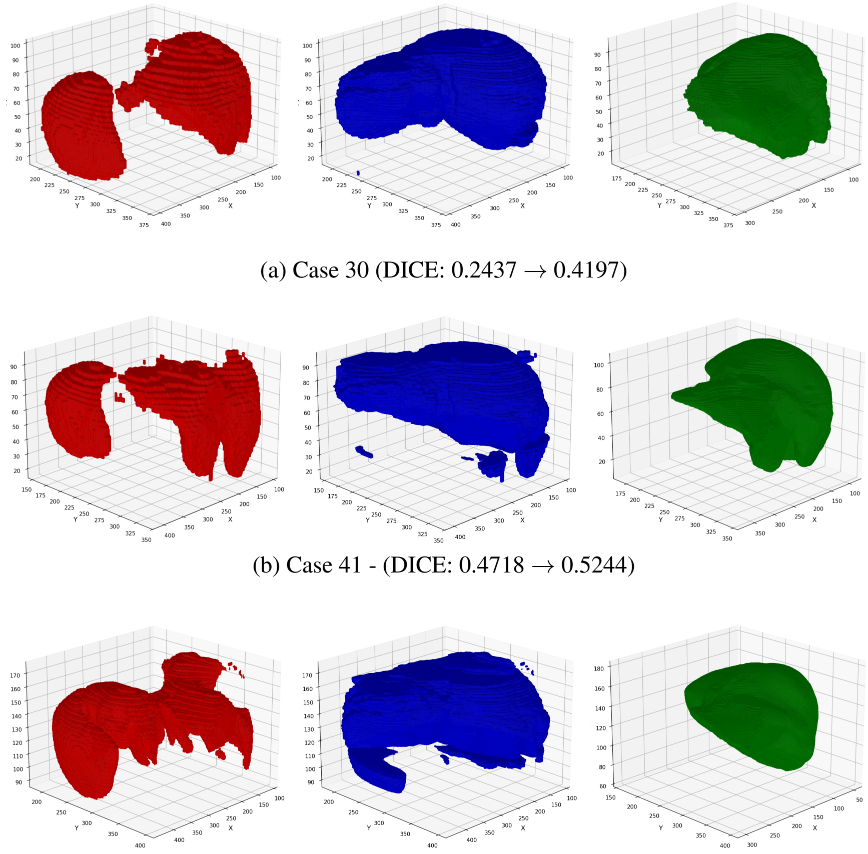


(a) Liver Segment - Scan 85; DSC = 0.947

(b) Liver Segment - Scan 86; DSC = 0.946

(c) Liver Segment - Scan 127; DSC = 0.944

Figure 2: Three out of the 10 reference liver segmentations from TotalSegmentator used for geometric matching, selected based on their high Dice similarity scores with ground truth annotations.



(a) Case 30 (DICE: 0.2437 → 0.4197)

(b) Case 41 - (DICE: 0.4718 → 0.5244)

(c) Case 35 (DICE: 0.5916 → 0.5663)

Figure 3: Representative segmentation refinement results showing three cases. Each row shows (left to right): original TotalSegmentator prediction, refined segmentation after geometric shape matching, and expert-validated ground truth annotation. Subfigures (a) and (b) demonstrate successful refinement with significant DICE improvement, with more anatomically plausible segmentations being formed. Subfigure (c) shows a case where despite the refined segmentation being more anatomically plausible than the original output, it maintained the original segmentation's quality with a slight decrement.