



HAL
open science

Resource-efficient Automatic Refinement of Segmentations via Weak Supervision from Light Feedback

Alix de Langlais, Benjamin Billot, Théo Aguilar Vidal, Marc-Olivier Gauci, Hervé Delingette

► **To cite this version:**

Alix de Langlais, Benjamin Billot, Théo Aguilar Vidal, Marc-Olivier Gauci, Hervé Delingette. Resource-efficient Automatic Refinement of Segmentations via Weak Supervision from Light Feedback. ISBI 2026 - IEEE International Symposium on Biomedical Imaging, Apr 2026, Londres, United Kingdom. ⟨hal-05571851⟩

HAL Id: hal-05571851

<https://inria.hal.science/hal-05571851v1>

Submitted on 30 Mar 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

RESOURCE-EFFICIENT AUTOMATIC REFINEMENT OF SEGMENTATIONS VIA WEAK SUPERVISION FROM LIGHT FEEDBACK

Alix de Langlais^{1,2}, Benjamin Billot¹, Théo Aguilar Vidal^{3,4}, Marc-Olivier Gauci^{2,3}, Hervé Delingette¹

¹ Inria, Université Côte d’Azur, Epione team, Sophia-Antipolis, France

² Institut de Biologie de Valrose, Inserm U1091, ICARE team, Nice, France

³ Institut Universitaire Locomoteur et du Sport, Centre Hospitalier Universitaire, Nice, France

⁴ Université Côte d’Azur, Nice, France

ABSTRACT

Delineating anatomical regions is a key task in medical image analysis. Manual segmentation achieves high accuracy but is labor-intensive and prone to variability, thus prompting the development of automated approaches. Recently, a breadth of foundation models has enabled automated segmentations across diverse anatomies and imaging modalities, but these may not always meet the clinical accuracy standards. While segmentation refinement strategies can improve performance, current methods depend on heavy user interactions or require fully supervised segmentations for training. Here, we present SCORE (Segmentation CORrection from Regional Evaluations), a weakly supervised framework that learns to refine mask predictions only using light feedback during training. Specifically, instead of relying on dense training image annotations, SCORE introduces a novel loss that leverages region-wise quality scores and over/under-segmentation error labels. We demonstrate SCORE on humerus CT scans, where it considerably improves initial predictions from TotalSegmentator, and achieves performance on par with existing refinement methods, while greatly reducing their supervision requirements and annotation time. Our code is available at: <https://gitlab.inria.fr/adelangl/SCORE>.

Index Terms— Weak Supervision, Efficient Segmentation Correction, Reaching Clinical Accuracy

1. INTRODUCTION

Modern automated segmentation approaches, best represented by deep-learning networks [1, 2], have been developed to alleviate the labor and variability associated with manual delineation of anatomical regions in medical images. The performance of these networks is intrinsically linked to the scope and variety of their training data, thus prompting the use of large, heterogeneous datasets. This data-centric approach is exemplified by foundation models, such as TotalSegmentator [3], which are pre-trained on thousands of multi-source scans to enable robust segmentations across a wide range

of input scans. While this strategy yields powerful general-purpose models, their resulting predictions may exhibit local inaccuracies that limit their adoption in clinical applications.

Several methods have been developed to refine automated segmentations. First, semi-automated interactive frameworks propose to place human experts in the correction loop [4], by allowing them to iteratively correct model predictions. Nevertheless, these systems demand substantial manual effort, thus limiting practical applicability. In contrast, fully automated post-processing techniques aim to remove the need for human intervention. Among these approaches, surface deformation-based methods [5] have been proposed to adjust boundaries toward object edges by predicting local displacements. Other fully automated approaches employ secondary refinement networks for correction, either to reclassify voxels within small image patches along the boundary [6], or to progressively improve results through a cascaded architecture [7]. Finally, a last class of automatic methods leverages anatomical priors to frame segmentation correction as a denoising problem, where Denoising Autoencoders (DAEs) [8] or Generative Adversarial Networks (GANs) [9] are trained to correct segmentation inconsistencies. However, these automated approaches rely on fully supervised ground-truth (GT) segmentations for training, which are laborious to obtain.

In this paper, we present SCORE (Segmentation CORrection from Regional Evaluations), a weakly supervised framework that learns to refine initial segmentation masks based on light feedback (only used during training). Specifically, we propose to train a refinement network using novel morphology-inspired losses that leverage region-wise quality scores and associated error labels (i.e., under- or over-segmentation). Here, we demonstrate our approach on computed tomography (CT) scans of the humerus, where it substantially improves the accuracy of the TotalSegmentator foundation model [3]. Overall, SCORE achieves comparable performance to existing semi- and fully automated refinement methods, while requiring neither intensive user interactions nor GT segmentations. These results open perspectives for the adoption of refinement methods in clinical applications.

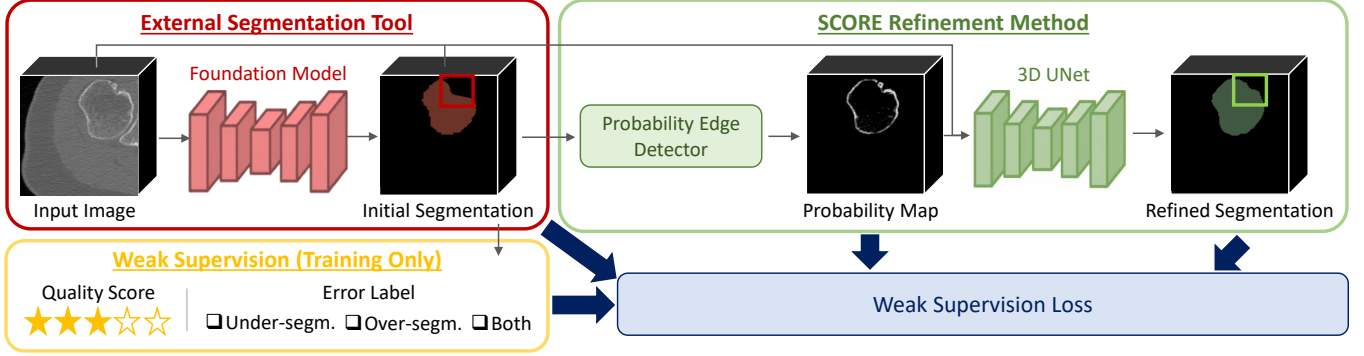


Fig. 1. Overview of SCORE, our weakly supervised framework to refine segmentation from an external tool using only light feedback. The network takes as input a 3D image, its initial segmentation, and a probability map for additional edge priors.

2. METHODS

Problem formulation. Let I be a 3D image associated with GT segmentation S taking values in $\{1, \dots, K\}$, where K is the number of regions of interest (ROI). As S is generally not accessible, one can use an automated segmentation model F (e.g., a foundation model) to approximate S with $\hat{S} = F(I)$. Yet, such models might not always reach accuracy levels required for clinical use. Here, we aim to train a refinement network φ to yield segmentations \tilde{S} of improved quality. Importantly, instead of training φ with previous fully-supervised refinement methods, we introduce SCORE, a new weakly supervised strategy (illustrated in Fig. 1) that we describe below.

Network inputs and boundary prior. In order to correct the initial segmentations given by F , our network φ takes a multi-channel input tensor with three components: (1) the 3D image I , to provide anatomical context; (2) the initial segmentation \hat{S} that we wish to correct; (3) a probabilistic map P of structure contours to serve as a prior to steer the corrections of φ towards visible anatomical boundaries. Here, we obtain P by clipping intensities between an Otsu-derived lower bound [10] and a percentile-based upper bound, and normalizing them in $[0, 1]$. Overall, \hat{S} is given by $\hat{S} = \varphi(I, \tilde{S}, P)$.

Weak supervision from quality scores. SCORE alleviates the needs of previous refinement methods for full supervision by only using light feedback. Specifically, each training pair (I, \tilde{S}) is associated with a set of quality scores $\{q_k\}_{k=1}^K$ ranking the accuracy of \tilde{S} for each region k from 0 (unusable) to 5 (excellent). Each quality score q_k is then associated with an error label $l_k \in \{-1, 0, 1, 2\}$ describing the error type for each region: -1 for under-segmentation, 1 for over-segmentation, 2 when a region presents both error types, and 0 for no error (when $q_k=5$). Overall, we hypothesize that these scores (used for training only) are sufficient to train a refinement network in a data-driven way. In addition to greatly decreasing annotation time, they are easily collected

using online forms via a simple visualization interface, thus bypassing the need for heavier medical image softwares.

Learning strategy. We train φ by designing a morphology-inspired loss function \mathcal{L} that leverages the aforementioned light training feedback. For this purpose, we define a three-fold function that converts the region-wise scores and labels into a voxel-based loss. Specifically, the three components of \mathcal{L} combine two targeted correction losses for over- and under-segmentations (\mathcal{L}^- , \mathcal{L}^+) together with a regularizing stability loss $\mathcal{L}^{\text{stab}}$ that penalizes deviations from correctly segmented areas in \tilde{S} . We now detail each of these losses.

In this paper, we assume that the foundation model F is relatively accurate, such that its errors lie on the outer parts of \hat{S} . Hence, we first propose a stability loss $\mathcal{L}^{\text{stab}}$ to encourage the interior parts of \hat{S} and \tilde{S} to be the same. More precisely, if k indexes the ROIs in \tilde{S} and \hat{S} , we define these ROI-specific interior areas Ω_k^{stab} as follows: if $l_k \neq 0$, Ω_k^{stab} corresponds to the area covered by \tilde{S}_k eroded by a factor η ; otherwise ($l_k = 0$), since \tilde{S}_k is excellent, we extend the stability area Ω_k^{stab} to the whole area occupied by \tilde{S}_k . As such, we define $\mathcal{L}^{\text{stab}}$ as a binary cross-entropy between \hat{S}_k and \tilde{S}_k inside Ω_k^{stab} :

$$\mathcal{L}^{\text{stab}} = \frac{1}{K} \sum_k \frac{1}{\Omega_k^{\text{stab}}} \sum_{v \in \Omega_k^{\text{stab}}} -\tilde{S}_{k,v} \log(\hat{S}_{k,v}), \quad (1)$$

where v indexes the coordinates in 3D tensors like \hat{S}_k and \tilde{S}_k .

Next, we correct regions with under-segmentations (i.e., $l_k \in \{-1, 2\}$) by using an expansive loss \mathcal{L}^+ that encourages φ to add voxels to \tilde{S}_k . Unlike $\mathcal{L}^{\text{stab}}$, \mathcal{L}^+ is applied to the outer part of \tilde{S}_k (named Ω_k^{orr}), defined as the band obtained by (1) dilating and eroding \tilde{S}_k by a factor η , (2) subtracting the eroded mask from the dilated one. Importantly, abusive expansions of \tilde{S}_k are prevented with a dual-mechanism. First, we ensure that new voxels are only added in anatomically plausible locations by weighting \mathcal{L}^+ with the boundary probability map P . Second, we use correction weights $w_k = (5 - q_k)/5 \in [0, 1]$ to adapt the magnitude of \mathcal{L}^+ to the

Table 1. Performance comparison of all refinement methods on each test set in terms of: Dice and HD95 statistics (mean \pm std), training supervision requirements, annotation types, and scan annotation times (both during training and inference). Statistical superiority of SCORE over tested baselines is denoted with $+$ and inferiority with $-$ (Wilcoxon signed-rank test, 5% level).

Methods	Dice			HD95 (mm)			Training supervision	Annot. type	Annot. time	
	CHU-Full	CHU-Prx	Wang-Dst	CHU-Full	CHU-Prx	Wang-Dst			Train	Test
TotalSegmentator [3]	92.4 (0.7) ⁺	93.7 (0.8) ⁺	91.9 (2.2) ⁺	2.1 (0.4) ⁺	2.0 (0.2) ⁺	5.5 (11.0) ⁺	-	-	-	-
Manual Correction	97.9 (0.7) ⁻	-	-	0.6 (0.1) ⁻	-	-	-	Manual	-	1h30
nnInteractive [4]	98.6 (0.2) ⁻	-	-	0.5 (0.1) ⁻	-	-	-	Semi-autom.	-	25min
Post-DAE [8]	96.1 (0.5)	96.6 (0.5) ⁺	94.0 (2.1) ⁻	1.5 (0.5)	1.4 (0.2) ⁺	5.4 (10.6)	GT segm.	Automated	1h30	4s
Post-DAE+Scan	97.3 (0.2) ⁻	97.6 (0.5) ⁻	91.9 (2.0) ⁺	0.8 (0.2) ⁻	1.1 (0.3) ⁻	4.8 (7.6) ⁻	GT segm.	Automated	1h30	4s
SCORE (ours)	96.2 (0.2)	97.0 (0.6)	93.0 (2.0)	1.4 (0.5)	1.2 (0.3)	5.4 (10.9)	Weak	Automated	4min	5s

quality scores, where $w_k = 1$ inflicts full penalty, and $w_k = 0$ means no correction. If $|\cdot|$ is the cardinality operator, we have:

$$\mathcal{L}^+ = \frac{1}{K} \sum_k \frac{1}{|\Omega_k^{\text{corr}}|} \sum_{v \in \Omega_k^{\text{corr}}} -w_k P_v \log(\hat{S}_{k,v}). \quad (2)$$

Finally, we tackle over-segmentation errors (i.e., $l_k \in \{1, 2\}$) using a subtractive loss \mathcal{L}^- that encourages the removal of boundary voxels. As before, \mathcal{L}^- operates in Ω_k^{corr} , and is also weighted by w_k . However, we modify the boundary probability weighting by now considering $(1 - P_v)$ to preferentially remove voxels away from the contours:

$$\mathcal{L}^- = \frac{1}{K} \sum_k \frac{1}{|\Omega_k^{\text{corr}}|} \sum_{v \in \Omega_k^{\text{corr}}} -w_k (1 - P_v) \log(1 - \hat{S}_{k,v}). \quad (3)$$

Overall, if λ^{stab} , λ^+ , and λ^- denote tunable coefficients, φ is trained using a total loss $\mathcal{L}(\tilde{S}, \hat{S}, \{q_k\}, \{l_k\}, P)$ given by:

$$\mathcal{L} = \lambda^{\text{stab}} \mathcal{L}^{\text{stab}} + \lambda^+ \mathcal{L}^+ + \lambda^- \mathcal{L}^-. \quad (4)$$

Data augmentation. To enhance the robustness of SCORE, we develop a data augmentation pipeline that simulates variability in input images I and initial segmentations \tilde{S} . First, we apply standard augmentations to the input 3D image I in the form of intensity perturbations (Gaussian blurring, Gaussian noise, contrast corruption). Then, we spatially deform the training pairs (I, \tilde{S}) using random affine transforms and flips in the left-right direction. Finally, we augment (\tilde{S}_k, q_k) pairs using score-guided morphological augmentation to increase robustness against different “strengths” of segmentation errors. For this purpose, we degrade under- and over-segmentations by respectively applying random erosions and dilations. Importantly, these operations are applied non-uniformly using spatially-varying 3D erosion/dilation kernels, whose radii (in $[-2, 2]$) at each location v are determined by a smooth random field. Quality scores q_k are updated based on the volumetric changes in \tilde{S}_k .

Implementation details. φ is implemented as a 3D UNet [1], and trained with the Adam optimizer [11] (1e-4 learning rate, batch size of 1). The final model is selected based on a validation set, which includes scans of diverse fields-of-view

(FoVs) to represent varying clinical acquisition setups. Finally, all hyperparameters ($\eta = 2$, $\lambda^{\text{stab}} = 5$, $\lambda^+ = 1$, and $\lambda^- = 1$) are fine-tuned on the validation set.

3. EXPERIMENTS AND RESULTS

Datasets. In this paper, we apply SCORE to refine CT segmentations of the humerus given by TotalSegmentator [3]. SCORE is trained on a private dataset of 31 humerus scans from CHU Nice (FoVs of $512 \times 512 \times 343$ to $768 \times 768 \times 1417$ voxels; resolutions of $0.23 \times 0.23 \times 0.3$ to $0.61 \times 0.61 \times 1.5$ mm³). We then use three test sets. The first two are private datasets from CHU Nice: **CHU-Full** with 5 complete humeral scans of similar FoVs and resolutions as before; **CHU-Prx** with 35 proximal (i.e., shoulder end) scans with the same resolutions but smaller FoVs ($512 \times 512 \times [158-662]$ voxels). The third test set, **Wang-Dst**, is a public dataset of 85 distal (i.e., elbow end) scans with FoVs of $512 \times 512 \times [188-497]$ voxels, and resolutions of $0.24 \times 0.24 \times 0.5$ to $0.78 \times 0.78 \times 1$ mm³ [12].

Baselines. We first compare SCORE against **manual correction** and **nnInteractive** [4], the state-of-the-art in semi-automated correction methods. Then, we assess **Post-DAE** [8], the state-of-the-art method in fully automated refinement. In addition to its original configuration (which only takes \tilde{S} as input), we extend Post-DAE into **Post-DAE+Scan**, a variant that also uses I as input for improved context.

Results. As shown in Table 1, SCORE delivers statistically significant improvements over the initial TotalSegmentator predictions across all test sets. On CHU-Full, our method increases the mean Dice from 92.4% to 96.2% and reduces the HD95 from 2.1 to 1.4 mm. This performance generalizes to the other datasets, with significant Dice improvements of 3.3% and 1.1% on CHU-Prx and Wang-Dst, respectively.

Interestingly, SCORE performs on par with the fully supervised methods Post-DAE and Post-DAE+Scan. Compared to Post-DAE, SCORE shows no statistical difference on CHU-Full, is superior on CHU-Prx, and yields the same HD95 but inferior Dice on Wang-Dst. While Post-DAE+Scan performs better on both CHU test sets, SCORE significantly

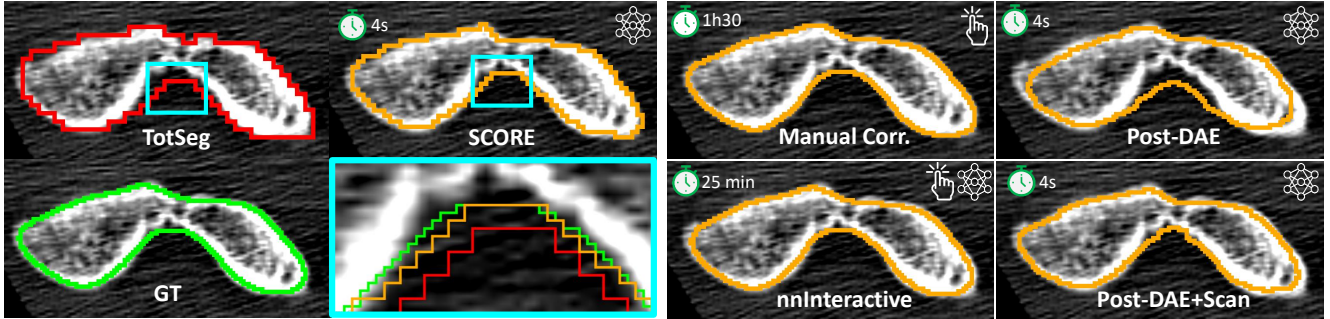


Fig. 2. Example segmentation for all methods on a CHU-Full subject, with associated correction time and interaction type.

Table 2. Dice statistics (mean, std) for the ablation study.

Methods	CHU-Full	CHU-Prx	Wang-Dst
SCORE (ours)	96.2 (0.2)	97.0 (0.6)	93.0 (2.0)
$l_k \in \{-1, 0, 1\}$	95.1 (0.4)	96.0 (0.9)	92.4 (2.0)
$\ominus \mathcal{L}^{\text{stab}}$	54.6 (3.9)	36.1 (10.4)	65.6 (7.9)
$\ominus P$	93.9 (0.6)	95.4 (0.9)	91.8 (2.1)
\ominus morph. augm.	95.9 (0.2)	96.4 (0.8)	92.3 (2.1)

outperforms it on Wang-Dst in Dice. While these methods perform very similarly in terms of accuracy, we note that SCORE only needs weak labels for training, here enabling us to reduce the annotation time by approximately 95%.

Finally, manual correction and the semi-automated nnInteractive method yield the highest accuracies, with respective Dice improvements of 1.7% and 2.4% over our method on CHU-Full. However, these slight performance gains come at prohibitive annotation costs (25 and 90 minutes per scan, respectively), thus preventing deployment at scale, such as on CHU-Prx (n=35) and Wang-Dst (n=85). In contrast, SCORE only requires about 4 minutes to gather the necessary feedback for training, positioning it as a practical solution to balance accuracy with annotation efficiency.

Ablation study. We now assess design choices in SCORE (Table 2). First, ablating multi-class labels (i.e., l_k cannot be 2) shows the benefits of accounting for several errors in each \tilde{S}_k . By ensuring consistency in interior parts, $\mathcal{L}^{\text{stab}}$ greatly contributes to accurate results. Then, removing boundary priors (i.e., P) leads to a small decrease in accuracy. Finally, our proposed non-homogeneous morphological augmentation of (\tilde{S}_k, q_k) pairs slightly increases the robustness of SCORE.

4. CONCLUSION

In this paper, we present SCORE, a weakly supervised framework for refining errors in initial segmentations from foundation models. Our core contribution is a morphology-inspired three-fold loss that translates region-specific quality scores and error-type labels into voxel-wise corrections. Our results show that SCORE improves initial masks from TotalSeg-

mentator to a performance level comparable to fully supervised methods, while reducing training annotation requirements by 95%. Current validation focuses on high contrast humerus CT, where Otsu-based thresholding reliably detects bone edges. Future work will extend SCORE across other anatomies, datasets, and foundation models. Overall, SCORE offers a path toward making automated segmentation refinement practical and scalable for routine clinical use.

Acknowledgments. This work is funded by an Inserm–Inria PhD fellowship, and 3IA Côte d’Azur (ANR-23-IACL-0001). **Compliance with ethical standards.** This work is a numerical simulation study that did not require ethical approval.

5. REFERENCES

- [1] Ronneberger et al., “U-Net: convolutional networks for biomedical image segmentation,” *MICCAI*, 2015.
- [2] Vaswani et al., “Attention is all you need,” *NeurIPS*, 2023.
- [3] Wasserthal et al., “TotalSegmentator: robust segmentation of 104 anatomical structures in CT images,” *Radiology: AI*, 2023.
- [4] Isensee et al., “nnInteractive: redefining 3D promptable segmentation,” arXiv:2503.08373, 2025.
- [5] Hu et al., “Deep level sets for object detection,” *CVPR*, 2017.
- [6] Tang et al., “Look closer to segment better: boundary patch refinement for instance segmentation,” *CVPR*, 2021.
- [7] Schnider et al., “Improved distinct bone segmentation in upper-body CT through multi-resolution networks,” *International Journal of Computer Assisted Radiology and Surgery*, 2023.
- [8] Larrazabal et al., “Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders,” *IEEE Transactions on Medical Imaging*, 2020.
- [9] Iqbal et al., “Generative adversarial networks and its applications in the biomedical image segmentation: a comprehensive survey,” *International J of Multimedia Info Retrieval*, 2022.
- [10] Otsu, “A threshold selection method from gray-level histograms,” *IEEE transactions on sys, man, and cyber*, 1979.
- [11] Kingma et al., “Adam: a method for stochastic optimization,” arXiv:1412.6980, 2017.
- [12] Wang et al., “A comprehensive survey on deep active learning in medical image analysis,” *Medical Image Analysis*, 2024.