Decoupled Single-Mask Annotation Noise Detection via Cross-Sectional Patch Self-Consistency

Yinheng Zhu[1] and Xiaowei Xu[2(✉)]

[1] Tsinghua University, Beijing, China
zhuyinheng666@gmail.com
[2] Guangdong Provincial People's Hospital, Guangzhou, China
xiao.wei.xu@foxmail.com

Abstract. Vascular computed tomography datasets are commonly annotated only once per scan, yielding the pervasive yet under-addressed problem of *single-mask* annotation noise. Existing solutions either require costly multi-rater fusion or are coupled with network training, preventing explicit auditing of *where* and *why* labels fail. We introduce a *decoupled* framework for single-mask annotation noise detection that leverages *cross-sectional patch self-consistency* to produce *interpretable* and *auditable* noise evidence. Tubular anatomy exhibits strong cross-sectional recurrence: patches extracted orthogonally along vessel centrelines recur in appearance across locations and subjects. Thus, anatomically similar patches should have consistent masks, and disagreement signals unreliable annotation. Our method samples cross-sectional patches, retrieves intensity-equivalent neighbours via scalable vector search, and computes a *patch-level noise score* from statistical mask disagreement, yielding explicit image–mask evidence for every flagged region. Aggregating scores produces scan-level *quality maps* for dataset quality assessment or quality-weighted training. Experiments on the coronary CT dataset validate the detected noise for improving training robustness and reveal systematic annotation biases, i.e., transverse and oblique vessels exhibit 5.1× higher error rate than axis-aligned structures, with additional correlations to cross-sectional area and intensity. Code is available here.

Keywords: Annotation Noise Detection · Single-Rater Annotation · Cross-Sectional Patch Consistency · Quality Assessment · Vascular Segmentation.

1 Introduction

Vascular segmentation is a core component in many Computed Tomography (CT) analysis pipelines, yet high-quality vascular annotations are difficult to obtain due to thin *curve-network* anatomy and contrast agent propagation [25]. Consequently, real-world datasets often contain localized annotation noise, while being annotated only once per scan because repeated expert labeling is costly.

This yields a common but under-addressed setting: *a single noisy mask per image*, which complicates both model training [22] and dataset quality control [12].

Existing noise-handling strategies fall into two categories. (1) *Multi-rater Fusion* fuses several masks per image [16,11,9,17,13], but fundamentally requires multiple raters, making it impractical for vascular CT where datasets almost always contain a single mask per scan. (2) *Training-coupled robust learning* incorporates noise handling into network optimization through weighting, disagreement, or structure-aware correction [10,23,14,18,8,24,19,4]. While effective for improving robustness, these approaches are not decoupled from training and cannot provide auditable evidence of label unreliability. As a result, no existing method supports decoupled and auditable noise localization in the practical single-mask setting.

Our key observation is that tubular anatomy provides natural cross-sectional recurrence across locations and subjects. We therefore introduce a *cross-sectional patch self-consistency principle*: patches with highly similar appearance should exhibit consistent masks, and strong disagreement indicates unreliable annotation. To operationalize this idea, we sample orthogonal cross-sectional patches along vessel centrelines, retrieve intensity-equivalent neighbours across scans, and compute a *patch-level noise score* that is directly traceable to explicit patch-pair evidence. We further aggregate patch scores into a 3D *quality map* for quality assessment (QA) and for practical noise mitigation via quality-weighted training.

Our contributions are threefold:

- We introduce a decoupled framework for annotation noise localization that operates directly on the image–mask pair, enabling practical use in the ubiquitous single-rater setting of vascular CT.
- We establish a cross-sectional patch self-consistency principle for detecting local label inconsistencies, producing inherently interpretable and auditable evidence via explicit patch-pair comparisons.
- We demonstrate that the detected noise exposes systematic annotation biases and improves segmentation robustness through quality-weighted training on coronary CT angiography.

2 Methodology

Self-Consistency Principle and Overview: To make the logical starting point explicit, let I and M denote the CT image and its corresponding segmentation annotation(s). The similarity measures in image and mask spaces are denoted by d_I and d_M, respectively. We first recall the principle underlying classical multi-rater consistency, where segmentations obtained repeatedly from the same image are expected to remain stable [16]:

$$\text{(Multi-rater Consistency)} \qquad d_I(I_i, I_j) = 0 \quad \Rightarrow \quad d_M(M_i, M_j) \leq \epsilon_M. \tag{1}$$

However, Eq. (1) is not directly applicable in the single-rater setting, because exact repetition of *whole image* is rare. Considering the key property that vascular structures exhibit strong cross-sectional recurrence, this rationale can be

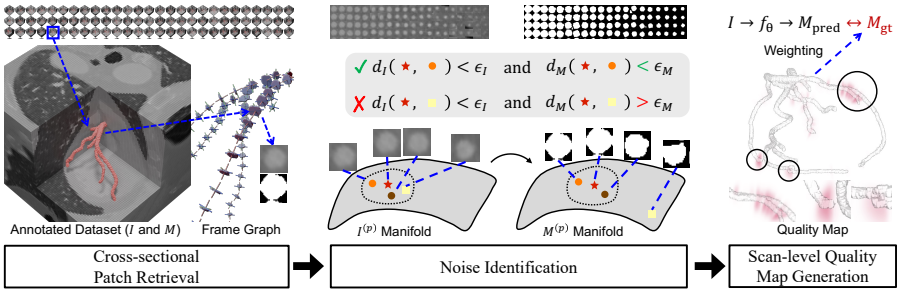

Fig. 1. Overview of the proposed method. Left: Cross-sectional patches are extracted along Bishop frames of vessel centrelines, and visually similar patches are retrieved across subjects. Middle: For each patch, its mask is compared with those of its intensity-equivalent neighbours. If any neighbour has a highly similar image but a substantially different mask, at least one annotation is potentially noisy. Right: Pair and patch level scores are aggregated into a scan-level quality map for weighted training.

transferred into *cross-sectional patch*. Specifically, when sampled orthogonally along vessel centrelines, cross-sectional image patches $(I^{(p)}, M^{(p)})$ frequently share highly similar appearances across nearby locations and even across different subjects. Therefore, local image–mask pairs can still be compared meaningfully across the dataset.

Building upon the above, we formulate the cross-sectional patch self-consistency principle as follows. For two patches $(I_i^{(p)}, M_i^{(p)})$ and $(I_j^{(p)}, M_j^{(p)})$,

$$\text{(Self-Consistency)} \qquad d_I(I_i^{(p)}, I_j^{(p)}) \leq \epsilon_I \quad \Rightarrow \quad d_M(M_i^{(p)}, M_j^{(p)}) \leq \epsilon_M. \qquad (2)$$

Accordingly, pairs that satisfy the antecedent but violate the consequent, i.e., high mask disagreement despite near-identical images, are flagged as potentially inconsistent. The overall pipeline is illustrated in Fig. 1, while the following paragraphs detail each stage.

Cross-sectional Patch Retrieval: **(1) Core idea**. To operationalize Eq. (2) effectively, we need to enlarge the number of near-identical image patch pairs, where the selection of local coordinate frames matters. We avoid the commonly seen Frenet–Serret frame [3] because its torsion term induces unstable rotations, especially in low-curvature segments, reducing the number of near-identical patch pairs and causing numerical instabilities. The Bishop frame [2] eliminates torsion and yields numerically stable, rotation-free orthonormal directions, making it ideal for consistent patch extraction. **(2) Detailed steps**. Given a volumetric image I and its segmentation mask M, we construct the graph of Bishop frames as follows. We first extract the graph of vessel centreline as described in [25]. The Bishop frame of each node is then propagated by parallel transport with zero-twist constraint along the centreline graph, starting from an arbitrary initial frame at the root node. Once the Bishop frame graph is constructed, we can sample cross-sectional patches, at fixed physical extent to cover the maximum expected vessel diameter, along the normal and binormal plane at each

node. Applying this process to all scans in the annotated dataset $\mathcal{D}$ yields the complete patch set $\mathcal{P} = \{(I_i^{(p)}, M_i^{(p)})\}$.

Noise Identification: **(1) Nearest-neighbour retrieval.** For each $P_i \in \mathcal{P}$, we define its neighbourhood as $\mathcal{N}_i = \{P_\ell \in \mathcal{P} \mid d_I(I_i^{(p)}, I_\ell^{(p)}) < \epsilon_I,\ i \neq \ell\}$, computed efficiently via vector search [5]. We set d_I as mean squared error (MSE) and d_M as $1 - \text{IoU}$ (Intersection over Union). MSE is adopted for scalability: with $\sim 3{\times}10^6$ patches the retrieval entails $\sim 10^{12}$ pairwise comparisons, for which structural metrics such as SSIM are computationally prohibitive and incompatible with scalable vector search; the resulting sensitivity to rotation and window-level shifts is analyzed in the failure cases (Fig. 6). For each neighbour pair (i, j), we compute the distance in image and mask domain, denoted as $d_I(I_i^{(p)}, I_j^{(p)})$, $d_M(M_i^{(p)}, M_j^{(p)})$ respectively. Aggregating across the dataset, these samples characterize the conditional distribution of mask distance given image distance, i.e., $p(d_M \mid d_I)$. **(2) Pair-level residual.** We summarize $p(d_M \mid d_I)$ by estimating the conditional mean and variance of the mask distance d_M as a function of the image distance d_I, and identify pairs whose d_M is unusually high given their d_I (i.e., masks that disagree despite similar images). Concretely, we partition $[0, \epsilon_I)$ into K equal-width bins $\{B^k\}_{k=1}^K$ according to d_I. For each bin B^k, we estimate the conditional baseline of d_M by the sample mean and standard deviation $\mu^k \approx \mathbb{E}[d_M \mid d_I \in B^k]$, $\sigma^k \approx \sqrt{\text{Var}[d_M \mid d_I \in B^k]}$. Fig. 2 illustrates this calibration. Under our definition $d_M = 1 - \text{IoU}$, unusually *large* d_M corresponds to the *lower tail* of IoU at the same d_I level. Given a pair (i, j) with $d_I(I_i^{(p)}, I_j^{(p)}) \in B^k$, we define a similarity-conditioned residual

$$r_{ij} = \frac{\mu^k - d_M(M_i^{(p)}, M_j^{(p)})}{\sigma^k + \varepsilon}, \tag{3}$$

where ε is a small constant for numerical stability. Equivalently, r_{ij} is a signed z-score analogous to t-statistic: a strongly negative value indicates that the observed mask disagreement is anomalously large given the image similarity, i.e., a statistically significant violation of Eq. (2). Bins with insufficient samples to estimate (μ^k, σ^k) reliably are excluded. **(3) Patch-level score.** Since r_{ij} is a pair-level measure, it does not directly indicate which patch is noisy. We therefore aggregate via the median, asking whether the *majority* of neighbours signal anomalous disagreement:

$$R_i = \text{median}(\{r_{ij} \mid j \in \mathcal{N}_i\}). \tag{4}$$

Under H_0 (patch i is clean), residuals scatter symmetrically around zero. If patch i is noisy, its mask consistently disagrees with neighbours, driving $R_i \ll 0$.

Scan-level Quality Map Generation: With noise scores computed for all patches, we aggregate them to the scan level to produce a spatially coherent quality map for downstream applications. We first propagate patch scores to all voxels via Voronoi labeling: each voxel $\mathbf{v}$ is assigned the noise score R_i of its nearest centreline node, yielding a per-voxel residual field $R(\mathbf{v})$. We then convert $R(\mathbf{v})$ to a voxel-wise quality score via the sigmoid function $q(\mathbf{v}) = \frac{1}{1+e^{-R(\mathbf{v})}}$,

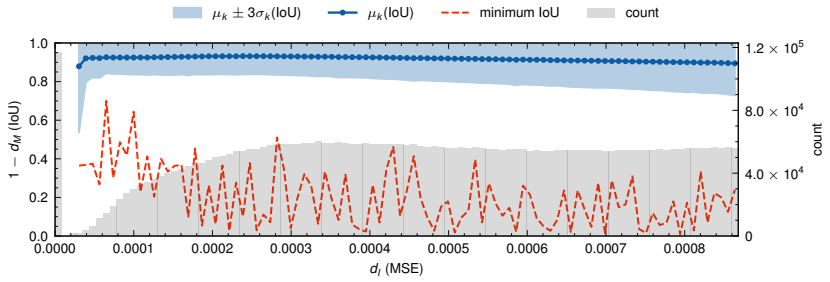

Fig. 2. Conditional distribution $p(d_M|d_I)$ and statistics of pairwise residual r_{ij}. The blue curve shows the bin-wise mean IoU, the shaded band indicates ±3 standard deviations, the red dashed curve plots the bin-wise minimum IoU, and the gray histogram reports the sample count per bin.

and define the final quality map as $Q(\mathbf{v}) = 1 - (1 - q(\mathbf{v}))\, w(\mathbf{v})$, where $w(\mathbf{v})$ is a distance-to-centreline weight that decays smoothly away from the vessel. By construction, $Q(\mathbf{v}) \in [0, 1]$ decreases monotonically with $R(\mathbf{v})$, while the attenuation term $w(\mathbf{v})$ limits the effect to vessel-adjacent regions to avoid down-weighting background voxels during training. An example quality map is shown in Fig. 1(right). Given any segmentation network f_θ producing per-voxel predictions $\hat{y}(\mathbf{v})$, we incorporate $Q(\mathbf{v})$ by reweighting a standard voxel-wise loss $\ell(\hat{y}(\mathbf{v}), y(\mathbf{v}))$: $\mathcal{L}_{\text{qw}} = \frac{1}{\sum_{\mathbf{v}} Q(\mathbf{v})} \sum_{\mathbf{v}} Q(\mathbf{v})\, \ell(\hat{y}(\mathbf{v}), y(\mathbf{v}))$, so that voxels with lower estimated annotation quality contribute less to optimization.

3 Experiments and Results

Experiment Setup: We evaluate the proposed framework on the ImageCAS dataset [21], which contains 1000 coronary CT angiography scans with expert annotations. Following the proposed pipeline above, we extract approximately 3×10^6 cross-sectional patches with size of 24×24 at pixel spacing of 0.125 mm, compute patch-level noise scores $\{R_i\}$ with $\epsilon_I = 10^{-3}$ (MSE), corresponding to peak signal-to-noise ratio (PSNR) of 30 dB (visually imperceptible), using $K = 100$ equal-width residual bins (a trade-off between per-bin variance-estimation stability and bin locality), and construct scan-level quality maps $Q(\mathbf{v})$ that reweight the training loss. We adopt nnUNet [6] with default automatic configuration as the baseline segmentation model and compare standard training with quality-weighted training. The dataset is split into training and test sets at an 8:2 ratio. Segmentation performance is evaluated using both volumetric and vessel-specific metrics. In addition to scan-level Dice score (DSC), we report Dice score on curved planar reformation [7] view (CPR-DSC) to measure cross-sectional agreement along the centreline, Average Surface Distance (ASD) and Hausdorff Distance at 95th percentile (HD-95) to assess boundary accuracy.

Table 1. Quantitative segmentation results with and without quality weighting. All metrics are reported as mean with the 95% CI in brackets.

	DSC ↑	CPR-DSC ↑	ASD (mm) ↓	HD-95 (mm) ↓
w/o	0.812 [0.804, 0.822]	0.801 [0.794, 0.809]	1.63 [1.41, 1.85]	10.27 [8.60, 11.95]
w/	**0.814** [0.805, 0.822]	**0.815** [0.807, 0.822]	**1.57** [1.36, 1.77]	**9.85** [8.23, 11.48]

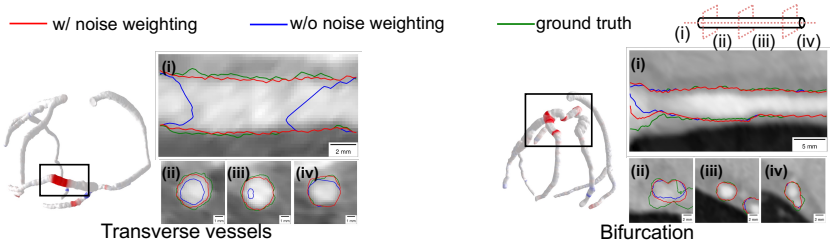

Fig. 3. Qualitative comparison on transverse vessels (left) and bifurcations (right), with CPR views (i) and cross-sections (ii–iv).

Overall Results: As shown in Table 1, the benefits of training with the noise-downweighted dataset are most evident on boundary-sensitive metrics, including CPR-DSC, ASD, and HD-95, indicating improved cross-sectional fidelity and surface regularity. These gains are particularly relevant for vascular segmentation, where geometric accuracy along thin structures and boundaries is often more informative than volumetric overlap alone. Meanwhile, the overall DSC remains comparable, which is expected since DSC primarily reflects foreground volumetric overlap and is less sensitive to localized boundary discrepancies. Fig. 3 provides qualitative comparisons in two challenging scenarios. For transverse vessels, whose axes are nearly parallel to the axial imaging plane, quality-weighted training yields contours that align more closely with the reference in both curved planar reformation and cross-sectional views. For bifurcations, where multiple branches converge and boundaries are prone to ambiguity, quality weighting produces smoother and more anatomically plausible delineations, which is particularly relevant for downstream analyses such as stenosis assessment and hemodynamic simulation.

Computation time: The pipeline splits into two stages. Precomputation, which performs top-k nearest-neighbour search over all $\sim(3 \times 10^6)^2$ patch pairs via FAISS [5] on an NVIDIA RTX 4090, takes approximately 6 hours and is performed once for the entire dataset. After the index is built, noise identification for a single scan of $\sim$3000 patches completes in under one minute.

Identified Noise and Analysis: The histogram of R_i (Fig. 4) exhibits a one-sided fat-tailed distribution, with a small but critical fraction extending to extreme negative values, corresponding to the most inconsistent patches in the dataset. Representative patch pairs sampled across the R_i spectrum confirm that the score reliably ranks annotation quality: at extreme negative scores

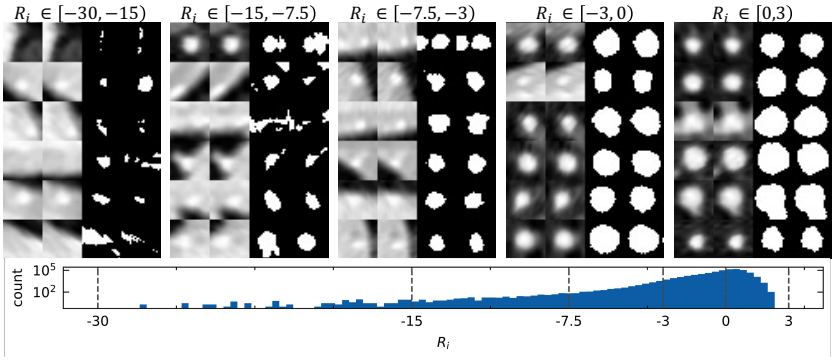

Fig. 4. Detected annotation noise across the R_i spectrum. Bottom: Histogram of R_i (log-scaled count). Top: Representative patch pairs from five R_i intervals.

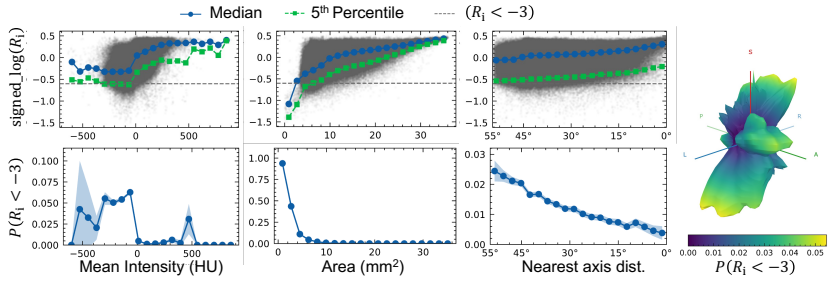

Fig. 5. Conditional distribution of R_i versus local attributes. The signed-log transform $\text{signed\_log}(R_i) = \text{sign}(R_i) \cdot \log_{10}(1 + |R_i|)$ compresses the fat-tailed distribution for visualization. Shaded bands denote 95% confidence intervals, and wide bands at intensity extremes reflect sparse samples. In the spherical histogram, both radius and color encode $P(R_i < -3)$.

($R_i \in [-30, -15)$), image patches are visually near-identical yet their masks differ drastically, constituting a clear self-consistency violation. As R_i increases toward -7.5, discrepancies become subtler, manifesting as local boundary shifts rather than wholesale segmentation disagreement. Patches near zero or positive R_i show well-matched masks, confirming that the score correctly identifies clean annotations as well as noisy ones.

We further investigate *where* annotation noise concentrates (Fig. 5). The probability of extreme errors ($R_i < -3.0$) peaks in the low-HU range (< 0 HU, approximately 5–6%) and drops in the typical lumen range (50–200 HU), consistent with reduced contrast between vessel wall and background. Small vessels ($< 2\,\text{mm}^2$) exhibit markedly elevated noise rates, rapidly declining beyond $5\,\text{mm}^2$, likely because thin vessels occupy only a few pixels in axial slices, limiting annotation precision. Both trends reflect the same underlying condition: reduced boundary discriminability in the annotation interface. Vessel orienta-

Fig. 6. Representative undetected annotation noise. Left: patches near-identical up to rotation. Right: patches near-identical up to window-level shift.

tion presents a qualitatively different, directional dependence: we quantify it as the angular distance θ to the nearest imaging axis (0° for axis-aligned, 55° for maximally oblique). Spearman correlation yields $\rho = -0.2$ ($p < 0.001$), and the Cochran–Armitage trend test [1] confirms a monotonic increase in extreme error rate ($z = -33.31$, $p < 0.001$), rising from 0.44% at $\theta \approx 0°$ to 2.24% at $\theta \approx 55°$ (a **5.1-fold** increase). The spherical histogram demonstrates this pattern: error rates are lowest near each of the six axis-aligned poles and peak in the inter-axis regions, where the vessel axis is far from all three imaging axes simultaneously. The mechanism is geometric: axis-aligned vessels appear as circular cross-sections in their corresponding imaging plane, whereas oblique vessels appear as elongated ellipses with diffuse boundaries, making consistent delineation inherently harder.

Failure Cases: The dominant failure pattern is false negatives: noise goes undetected when no sufficiently similar counterpart is retrieved. As shown in Fig. 6, this occurs when patches are near-identical only up to rotation or window-level shift, where raw MSE retrieval is sensitive, leaving the inconsistency undetected. Centreline and Bishop-frame errors likewise reduce retrieval *recall*, not *precision*, which our design prioritizes. Recall is recoverable by denser sampling of rotations and shifts at higher computational cost.

Discussion: **(1)** The key conceptual innovation is grounding noise detection in *cross-sectional self-consistency* rather than model behavior. Training-coupled single-mask methods infer noise indirectly from optimization dynamics such as loss or feature divergence (e.g., deep self-cleansing [4]), leaving it ambiguous whether a flagged region is mislabeled or merely a hard sample, and they consume this signal by excluding or replacing labels inside the training loop. In contrast, our criterion is computed directly from the image–mask pair, so each flagged region traces to a verifiable patch-pair violation and is *auditable*: a suspicious patch is surfaced together with its near-identical neighbours as explicit evidence for human QA, without multi-rater annotation or training interference. The closest network-free analogue is the multi-rater iSTAPLE [11], which similarly emphasizes a consistency criterion rather than a benchmark comparison. **(2)** This enables the discovery of systematic biases otherwise hidden: the dominant orientation dependence follows directly from the slice-by-slice paradigm of tools such as ITK-SNAP [20], where oblique vessels appear as irregular elongated shapes. Hemodynamic simulation workflows corroborate this by annotating directly on centreline cross-sections [15]. **(3)** The method targets tubular anatomy and detects rather than corrects noise, while extension to other tubu-

lar structures is straightforward. Our evaluation is confined to a single dataset (ImageCAS), and the cross-sectional recurrence assumption may hold less well on more heterogeneous data; broader cross-dataset validation and explicit label correction are left to future work.

4 Conclusion

We present a *decoupled*, *interpretable*, and *auditable* framework for localizing annotation noise from a *single* vascular mask. The method exploits cross-sectional self-consistency: near-identical patches should have near-identical masks, and violations are flagged as noise. Each flagged region is directly traceable to concrete image–mask patch-pair evidence. On ImageCAS, quality-weighted training improves CPR-DSC by 1.4% and reduces HD-95 by 4.1%. More importantly, the analysis reveals that **annotation noise is not random**, and systematic biases exist, correlated with vessel orientation, area, and intensity.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62276071), Guangdong Special Support Program-Science and Technology Innovation Talent Project (No. 0620220211), Guangzhou Key Research and Development Program (No. 2025B01J3011).
Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agresti, A.: Categorical Data Analysis. Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, 3rd edn. (2013)
2. Bishop, R.L.: There is more than one way to frame a curve. The American Mathematical Monthly **82**(3), 246–251 (1975). https://doi.org/10.2307/2319846
3. do Carmo, M.P.: Differential Geometry of Curves and Surfaces. Prentice-Hall, Englewood Cliffs, NJ (1976)
4. Dong, J., Zhang, Y., Wang, Q., Tong, R., Ying, S., Gong, S., Zhang, X., Lin, L., Chen, Y.W., Zhou, S.K.: Deep self-cleansing for medical image segmentation with noisy labels (Sep 2024)
5. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
7. Kanitsar, A., Fleischmann, D., Wegenkittl, R., Felkel, P., Gröller, M.E.: CPR – curved planar reformation. In: Proceedings of the IEEE Visualization (VIS). pp. 37–44. IEEE (2002). https://doi.org/10.1109/VISUAL.2002.1183754
8. Li, S., Gao, Z., He, X.: Superpixel-guided iterative learning from noisy labels for medical image segmentation (Jul 2021)

9. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation (Mar 2022). https://doi.org/10.48550/arXiv.2111.13410
10. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2596–2606. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.00263
11. Liu, X., Montillo, A., Tan, E.T., Schenck, J.F.: istaple: improved label fusion for segmentation by combining staple with image intensity. In: Ourselin, S., Haynor, D.R. (eds.) Medical Imaging 2013: Image Processing. vol. 8669, p. 86692O. SPIE, Lake Buena Vista (Orlando Area), Florida, USA (Mar 2013). https://doi.org/10.1117/12.2006447
12. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2021)
13. Rädsch, T., Reinke, A., Weru, V., Tizabi, M.D., Heller, N., Isensee, F., Kopp-Schneider, A., Maier-Hein, L.: Quality assured: rethinking annotation strategies in imaging ai. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, vol. 15136, pp. 52–69. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73229-4_4
14. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation (May 2023)
15. Updegrove, A., Wilson, N.M., Merkow, J., Lan, H., Marsden, A.L., Shadden, S.C.: Simvascular: an open source pipeline for cardiovascular simulation. Annals of biomedical engineering **45**(3), 525–541 (2017)
16. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (Jul 2004). https://doi.org/10.1109/TMI.2004.828354
17. Wu, N., Sun, Z., Yan, Z., Yu, L.: Feda3i: Annotation quality-aware aggregation for federated medical image segmentation against heterogeneous annotation noise. Proceedings of the AAAI Conference on Artificial Intelligence **38**(14), 15943–15951 (Mar 2024). https://doi.org/10.1609/aaai.v38i14.29525
18. Xu, Z., Lu, D., Luo, J., Wang, Y., Yan, J., Ma, K., Zheng, Y., Tong, R.K.Y.: Anti-interference from noisy labels: mean-teacher-assisted confident learning for medical image segmentation. IEEE Transactions on Medical Imaging **41**(11), 3062–3073 (Nov 2022). https://doi.org/10.1109/TMI.2022.3176915
19. Yao, J., Zhang, Y., Zheng, S., Goswami, M., Prasanna, P., Chen, C.: Learning to segment from noisy annotations: a spatial correction approach (Jul 2023). https://doi.org/10.48550/arXiv.2308.02498
20. Yushkevich, P.A., Gao, Y., Gerig, G.: Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 3342–3345. IEEE (2016)
21. Zeng, A., Wu, C., Lin, G., Xie, W., Hong, J., Huang, M., Zhuang, J., Bi, S., Pan, D., Ullah, N., Khan, K.N., Wang, T., Shi, Y., et al.: ImageCAS: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography

angiography images. Computerized Medical Imaging and Graphics **109**, 102287 (2023). https://doi.org/10.1016/j.compmedimag.2023.102287
22. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization (Feb 2017). https://doi.org/10.48550/arXiv.1611.03530
23. Zhang, F., Liu, H., Wang, J., Lyu, J., Cai, Q., Li, H., Dong, J., Zhang, D.: Cross co-teaching for semi-supervised medical image segmentation. Pattern Recognition **152**, 110426 (Aug 2024). https://doi.org/10.1016/j.patcog.2024.110426
24. Zhou, Y., Yu, H., Shi, H.: Study group learning: improving retinal vessel segmentation trained with noisy labels. In: De Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, vol. 12901, pp. 57–67. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_6
25. Zhu, Y., Wang, Y., Di, C., Liu, H., Liao, F., Ma, S.: Sparse and transferable three-dimensional dynamic vascular reconstruction for instantaneous diagnosis. Nature Machine Intelligence pp. 1–13 (2025)