Mitral-Valve Segmentation Across CT, TEE, and Surgical Video

Harshit Agrawal[1]

AIATELLA Oy., Helsinki, Finland
harshit@aiatella.com
Codabench username: hars25

Abstract. Mitral-valve assessment guides diagnosis, planning, and intraoperative care for structural heart disease across three visually distinct modalities: cardiac CT for valve geometry, 3D transesophageal echocardiography (TEE) for volumetric function, and surgical video for leaflet-motion confirmation. The MVAA 2026 challenge asks one team to segment the mitral valve in all three. We describe the system we submitted to the hidden test phase: for CT, an eight-member probability fusion of two nnU-Net v2 families, one anatomy-gated self-trained over 1,040 unlabeled volumes and one supervised with a large patch; for TEE, a single supervised nnU-Net v2 on 175 labeled cases; for video, an ensemble of three DINOv2-B/DPT networks initialized from surgical self-supervised weights and trained with consistency regularization over 1,379 unlabeled frames. On the hidden test set it reaches a Dice similarity coefficient of **0.844** on CT (Hausdorff 4.94 mm, average surface distance 0.416 mm) and **0.815** on video (243.3 px, 105.2 px), placing **3rd of 36** teams on CT and 6th overall; TEE is score-neutralized by the organizers because its data is public, so it is reported for completeness only. Our second contribution is a measured account of what did *not* transfer: two interventions the public validation board rejected improved the hidden test, as did a third our own oracle rejected; we show why that board is blind to the error mode – valve-presence mistakes on empty frames – that dominates the video score. Code is available at https://github.com/harshitAgr/mvaa2026.

Keywords: Mitral valve · Multi-modal segmentation · Semi-supervised learning · nnU-Net · Echocardiography · Surgical video

1 Introduction

Accurate mitral-valve delineation drives diagnosis, pre-procedural planning, and intraoperative guidance for structural heart disease, and no single modality provides complete information: *cardiac CT* resolves the static 3D annulus/leaflet geometry for device sizing, *3D transesophageal echocardiography (TEE)* gives intraoperative volumetric function, and *surgical video* gives the surgeon's real-time view of leaflet motion. All three still rely on manual segmentation.

The MVAA 2026 challenge [4,13,14] asks one team to segment the valve across all three. Tasks 1 (CT) and 3 (video) supply far more unlabeled than labeled data and therefore require semi-supervised learning (SSL); Task 2 (TEE) is fully supervised. Ranking is decided in a hidden test phase: participants publish a Docker image, the organizers pull it onto their own offline V100 and run it over unpublished data, and only task-level mean metrics are returned. That contract is looser than it first appeared – the enforced limit is a whole-run timeout, not the $\leq$10 s-per-case target we assumed while it was unpublished – and the resulting headroom allowed compact ensembles rather than one model per task.

Because each submitted image returns a score, the challenge also lets us ask a question that is usually unanswerable: *when a public validation split and a hidden test set disagree about an intervention, which should be believed?* We answer it with eight scored hidden-test runs that differ by one controlled change per task. Our contributions are (i) a deployed multi-modal system that runs offline in one container and scores DSC 0.844 (CT) and 0.815 (video) on the hidden test, ranking 3rd of 36 teams on CT and 6th overall; (ii) challenge-compliant SSL pipelines – anatomy-gated offline self-training for CT at 88.6% pseudo-label acceptance, and consistency-regularized teacher–student training for video; and (iii) a validation-to-test transfer study in which two interventions the public board rejected, and one our internal oracle rejected, each improved the hidden test, with the diagnostics that explain why. The components are established; the contribution is their cross-modal application under one budget and the measurement of what transfers.

2 Related Work

Our volumetric backbone is the self-configuring nnU-Net v2 [7]; a recent follow-up [8] argues that reported gains are often validation artifacts unless checked against an independent split – the discipline we apply to every intervention, and whose failure modes we quantify in Section 5. For the unlabeled-data-rich tasks we use consistency regularization with an exponential-moving-average teacher [21], following UniMatch V2 [25] for video. Encoder pretraining dominates the surgical-video literature: ImageNet features [20] and, more recently, surgery-specific pretraining [9] transfer better than from-scratch training on small intra-operative datasets. Our final model combines both lines – a DINOv2 [15] ViT-B/14 encoder carrying SurgeNetXL weights with a dense-prediction head [16] – and its adoption was the largest hidden-test gain we measured. Domain randomization [22] is the one robustness lever that survived our gates.

Our budget-first framing echoes the FLARE challenges [11,5], which favor a single efficient model over ensembling; MVAA 2026 [4] instead sets a whole-run timeout, which makes the opposite choice affordable, and differs from single-modality benchmarks such as surgical-video understanding [27] in requiring one team to satisfy three modalities at once – what makes the transfer question askable.

3 Method

The three tasks share a semi-supervised paradigm and a common deployment target, instantiated per modality below and summarized in Fig. S1.

3.1 Common Framework

For the volumetric tasks we adopt **nnU-Net v2** (v2.7.0) in its `3d_fullres` configuration, relying on its self-configuring pipeline – fingerprinting, preprocessing, and network/patch planning derived from the data – which transfers cleanly to both 3D modalities; inference uses sliding-window prediction with mirroring test-time augmentation [7]. The video task uses a 2D vision-transformer encoder with a dense-prediction head (Section 3.4). Tasks 1 and 3 share one teacher–student paradigm [21]: CT applies it *offline* (a supervised model pseudo-labels the unlabeled volumes, which are gated and folded back into training), video *online* (an EMA teacher supervises a student on strongly-augmented unlabeled frames). Task 2 has no unlabeled data and is fully supervised.

Deployment and validation protocol. The organizers pull our image from a public registry and execute it on one V100 (32 GB), offline, with a whole-run timeout of 6 h for the entire hidden set; all weights are baked in and every network is built with pretrained-weight downloads disabled. Because the timeout is per run rather than per case, ensembling is affordable – the deployed image uses about a tenth of the allowance. Our development GPU cannot execute a CUDA build supporting the V100, so every image was gated on exact CPU-versus-host mask parity (mean Dice $\geq$0.9998) instead of a local GPU smoke test. Each intervention passed through up to three surfaces: a leakage-free internal split (k-fold for 3D, leave-one-video-out for video) with thresholds fixed before the metric was opened; the public validation board; and the hidden test, of which at most ten scored attempts were available. Where the board and the oracle disagreed we followed the mechanism and the oracle, correctly twice; once we deployed *against* the oracle and were also right (Section 5).

3.2 Task 1: Cardiac CT

Task 1 segments the mitral valve in pre-procedural cardiac CT: 27 labeled volumes, 1,040 unlabeled volumes, 30 hidden validation and 70 hidden test cases. Volumes are supplied *pre-cropped to the mitral-valve ROI* (roughly 5–7 cm cubes) with sub-millimetre in-plane spacing and preserved Hounsfield units. The label is **binary**, forms exactly one connected component, and occupies only $\sim$1.2–2.7% of the cropped volume.

Self-training. A five-fold supervised ensemble trained on the 27 labeled cases pseudo-labels all 1,040 unlabeled volumes. Rather than a softmax-confidence threshold, acceptance uses a hard anatomical and acquisition filter whose every

bound is read off the labeled set rather than tuned: a single dominant component, foreground volume and fraction inside the labeled range, and a minimum through-plane resolution (full criteria in Supplementary S2). **921 of 1,040 volumes pass (88.6%)**, are largest-component-cleaned, and enter *only* the training half of each fold under a `PLBL_` prefix, with real cases oversampled so real:pseudo draws stay near 1:2.5. The fold assignment of the 27 real cases is copied verbatim from the supervised run, so pseudo cases never reach a held-out fold. One caveat: the teacher spans all 27 labeled cases, so a pseudo case in fold k's training set carries information from fold k's held-out cases; the *internal* estimate for this family is therefore optimistic and we do not use it to claim a self-training gain. No challenge validation or test case enters training by any path.

Two-family fusion. The deployed model averages **eight** nnU-Net members from two deliberately different families, in probability space, at equal *family* weight ($0.5\,\overline{A} + 0.5\,\overline{B}$), followed by a 0.5 threshold and a binary largest-connected-component step. Family A is five members trained on the self-training dataset with standard plans; family B is three supervised members trained on the 27 labeled cases with an enlarged patch of $112 \times 160 \times 192$.

The motivation is variance, not capacity. On the leakage-free out-of-fold split, ASD is nearly linear in Dice ($\mathrm{ASD} = 2.007\,(1 - \mathrm{DSC}) - 0.033$, $r = 0.934$); that line predicts 0.285 mm at our then-current hidden-test Dice of 0.8417, but we scored 0.466 mm. A uniformly harder cohort cannot produce that gap – a few catastrophic cases can, and member averaging is the canonical suppressor. A leakage-free two-family probe ($n = 21$) confirmed the signature: fusion is non-inferior to the better member on all three metrics and better on ASD, while the Hausdorff median moves by exactly zero, i.e. it does nothing on most cases and rescues a few. Two family-B members were then removed in separate steps, each selected *target-blind* from out-of-fold under-segmentation and consensus disagreement, and each confirmed on the hidden test before the next was attempted.

3.3 Task 2: 3D TEE

Task 2 segments the anterior and posterior mitral leaflets in intraoperative 3D TEE. It is the only fully supervised task, and it is **excluded from the competition ranking**: because the underlying data is public – the MVSeg 2023 challenge release, whose volumes were introduced by Carnahan et al. [3] – the organizers award every eligible team a flat normalized score on all three Task 2 metrics regardless of model quality. We therefore report it for completeness and spend no deployment trade-offs on it.

An MD5 check confirmed the MVAA data is the MVSeg 2023 train/val split verbatim, so the disjoint remainder of that public release adds 70 clean labeled cases, giving **175**. The deployed model is a single nnU-Net v2 `3d_fullres` trained on all 175 cases for 250 epochs with an enlarged patch of $128 \times 192 \times 224$ and one task-specific change: mirroring is kept on the left–right and superior–inferior

axes but **removed on the anterior–posterior axis**, along which a flip maps anterior-leaflet voxels into posterior-leaflet territory and poisons the chirally distinct labels. The expansion improved monotonically with training length (DSC 0.8527 / 0.8531 / 0.8571 at 100 / 200 / 250 epochs). At inference we retain the per-class largest connected component, which cuts Hausdorff from 12.69 to 10.87 mm at neutral DSC; mirroring test-time augmentation is disabled in the container, since it costs about 30% of the task's runtime and cannot affect the ranking. A boundary-targeted objective – soft-clDice [19] on the region+CE loss – moved DSC by only +0.0012, inside the ±0.008 five-fold noise floor, and was shelved. Raw hidden-test metrics are DSC 0.946, Hausdorff 3.15 mm, ASD 0.137 mm.

3.4 Task 3: Surgical Video

Per-frame mitral-valve segmentation in intraoperative endoscopic video: a 2D problem with far more unlabeled than labeled data, and the task on which the competition was decided. There are **180 labeled frames** from six videos (**118** containing the valve, 62 empty) plus **1,379 unlabeled frames**; one frame whose mask includes chamber tissue is excluded, leaving 179 eligible. The public validation split is 48 frames from two further videos, *all of which contain the valve*; the hidden test set, by design, also contains frames with no valve. Models operate on a 448×800 grid and every mask is written back at native 720×1280.

Deployed model. The final system is an **equal-weight probability ensemble of three DINOv2-B/DPT networks**: a DINOv2 ViT-B/14 encoder [15] carrying SurgeNetXL surgical self-supervised weights [9], with a dense-prediction transformer head [16]. Training follows a fixed UniMatch-V2-style recipe [25]: five decoder-only warm-up epochs, then full-backbone training with an EMA teacher, two strongly-augmented views, CutMix, and complementary dropout over all 1,379 unlabeled frames, for exactly 70 epochs. **No checkpoint selection was permitted** – the eligible state is the epoch-70 EMA weights, fixed in advance – and neither the pseudo-label nor the decision threshold was tuned on the public validation frames: both were fixed on the six-video internal split, which contains empty frames, precisely because the validation frames do not. Two members were trained with one video held out and the third on all six. At inference each member uses D4 test-time augmentation, the three probability maps are averaged at native resolution and thresholded at 0.45, and **no post-processing is applied**. Because the hidden score averages over frames with no valve, we tracked empty-frame false positives explicitly and vetoed any candidate that increased them (Supplementary S4).

How the recipe was reached. The deployed model is the last of seven Task 3 configurations scored on the hidden test; Table 2 lists the changes taken to a decision. We began with a convolutional ensemble: a UNet++ [26] on an ImageNet EfficientNet-B4 encoder [20] averaged with a U-Net [18] on a SurgeNetXL-initialized ConvNeXt V2-Tiny encoder [23], both trained with a DiceFocal objective (0.7 Dice [12] + 0.3 focal [10]) and an online mean teacher. Retraining

the first member under blur and photometric domain randomization [22], on the supervised branch only, was the first robustness lever to pass a six-fold leave-one-video-out gate against a held-out corruption oracle, and added +0.014 DSC on the hidden test; re-weighting it to 0.60/0.40 added +0.009; substituting the DINOv2-B/DPT members added +0.021; and dropping the component cleanup and then the last convolutional member each added a further increment. That last step contradicted our internal oracle by a wide margin (0.611 vs. 0.797 all-frame DSC) yet improved every hidden-test metric – a limitation of the oracle, not a vindicated prediction.

4 Experimental Setup

We use the official MVAA 2026 data [4,13]. The public splits are: Task 1, 27 labeled and 1,040 unlabeled volumes with 30 validation and 70 hidden test cases; Task 2, 105 labeled volumes with 20 validation cases, extended to 175 from the public source release (Section 3.3); Task 3, 180 labeled frames from six videos (118 with valve) and 1,379 unlabeled frames, with 48 validation frames from two further videos. Per-task training protocols are in Table S1.

Evaluation surfaces. Validation labels are scored on a public Codabench board, but the ranking is decided in the hidden test phase. We therefore report hidden-test metrics as headline results and validation numbers only as context. One caveat spans both surfaces: the organizers corrected the Task 3 distance formula between phases, so validation-phase video boundary metrics are not comparable to test-phase values.

Implementation and measured cost. Training ran in PyTorch 2.11 (CUDA 12.8) with nnU-Net v2 (v2.7.0) [7], MONAI 1.5 [2], and `segmentation_models.pytorch` 0.5 [6] on a single NVIDIA RTX PRO 6000; the container is pinned to PyTorch 2.5.1 (CUDA 12.4), the last build still emitting kernels for the V100. On the organizers' hardware the deployed image processes the entire hidden set in **2,151 s** of a 21,600 s allowance, and growing from one CT member to eight and two video members to three moved total wall time by under 30% (2,010 → 2,151 s, peak 2,593 s) – cost here is dominated by CPU-side resampling and I/O, not network evaluation. Per-task timings and memory are in Supplementary S3.

5 Results and Discussion

Table 1 reports the deployed system's hidden-test metrics. Cardiac CT is the most favourable target – DSC 0.844 with sub-half-millimetre average surface distance – ranking **3rd of 36** teams on the task and **1st on Hausdorff distance**. Surgical video reaches DSC 0.815; its boundary metrics are in pixels and two orders of magnitude larger, which is a property of the metric rather than of the masks (Section 5.2). The final board reports raw metrics rather than the normalized score, so standings are computed from published per-metric ranks

Table 1. Hidden-test results for the deployed container (submission 873932); per-metric rank in parentheses. HD/ASD are in mm for 3D and px for video. Baseline DSC is the official baseline reproduced on internal held-out splits (3 CT, 20 TEE, 39 frames), for context only. DSC ↑; HD, ASD ↓.

Task	DSC ↑	HD ↓	ASD ↓	Task rank	Baseline DSC
Cardiac CT	**0.8443** (4th)	**4.944 mm** (1st)	0.4163 mm (5th)	**3rd / 36**	0.796
3D TEE *(not ranked)*	0.9464	3.152 mm	0.1367 mm	—	0.701
Surgical video	0.8151 (12th)	243.3 px (17th)	105.2 px (21st)	16th / 36	0.632
Overall (mean per-metric rank; Task 2 flat for all teams)				**6th / 36**	

Table 2. Interventions taken to a decision. Internal surfaces are leave-one-video-out (video) or leakage-free out-of-fold (CT); hidden-test deltas are against the preceding scored run, and two runs each carried one T1 and one T3 change. Row ‡ changed two things within one task.

Intervention	Task	Internal	Validation board	Hidden test	Verdict
Domain randomization	T3	6-fold GO, blur recov. +0.61	+0.0036 DSC	+0.0138 DSC, −146 px HD	deployed
Fusion weight 0.60/0.40	T3	+0.0221 DSC, HD −18%	−0.0026 DSC (reject)	+0.0088 DSC, HD −13.9%	deployed
Two-family CT fusion	T1	ΔHD −0.113, non-inferior	+0.0009 DSC, HD +0.011 (reject)	+0.0019 DSC, HD −0.66	deployed
Surgical DINOv2 members	T3	—	—	+0.0205 DSC, HD −10.5 px	deployed
3rd member + no cleanup‡	T3	—	—	+0.0008 DSC, HD −4.9 px	deployed
All-transformer ensemble	T3	0.611 vs 0.797 DSC (reject)	—	+0.0038 DSC, HD −15.5 px	deployed
CT all-data variant	T1	—	HD gain	−0.0008 DSC, HD +0.49	reverted
SegFormer [24] 3rd member	T3	+0.0040 DSC (< gate)	—	+0.0005 DSC, HD +49.3 px	reverted

over the 36 teams with a successful run; on that basis the submission places **6th overall**.

5.1 What transferred, and what reversed

Because each Docker submission returns a score and we changed one thing per task at a time, the eight scored runs form an ablation measured on the ranked surface rather than on a proxy. Table 2 lists every intervention taken to a decision with its effect size on each surface; shelved interventions are in Table S2. The pattern is that internal oracles built to include the penalized error mode agreed with the hidden test, while the public board and internal oracles that excluded it did not.

5.2 Discussion

A small validation board can be structurally blind, not merely noisy. Three interventions were rejected before submission and improved the hidden test anyway (Table 2). For the video board the cause is composition, not sample size: it is 48 frames from two videos and *every frame contains the valve*, so it cannot express the false-positive-on-empty-frame error the hidden set penalizes. A leakage-free six-video oracle that did include empty frames predicted the fusion-weight result correctly in sign and magnitude (−18% HD, −37% ASD predicted; −13.9%, −21.1% delivered). A validation surface should be audited for *which failure modes it can represent* before its verdicts become gates.

Video boundary metrics are dominated by existence errors. Across the final leaderboard, Hausdorff distance is offset from average surface distance by a near-constant amount rather than scaling with it. Regressing one on the other over the 35 teams with a non-degenerate video run (one run with ASD >20,000 px excluded) gives slope 0.996 and intercept 175 px; equivalently, HD − ASD has median 141 px, interquartile range 126–167 px, and never falls below 110 px. Boundary sloppiness would scale the two together; a shared additive term will not. In the same field, mean ASD values of 290–480 px are geometrically impossible for a plausible valve mask in a 720×1280 frame. Both are signatures of per-frame existence errors – a mask where the truth is empty, or a missed valve – entering the mean with a large fixed penalty, so the video ranking is far more sensitive to valve-presence decisions than to contour quality.

Encoder pretraining still dominates on video. ImageNet pretraining of the EfficientNet-B4 encoder alone was worth 16.5 pp DSC over a from-scratch model, and replacing the convolutional pair with surgical-self-supervised DINOv2-B/DPT members gained a further +0.021 DSC – more than every other video intervention we scored. The same recipe with a *small* DINOv2 encoder was a decisive local failure (−0.056 DSC), so backbone scale and surgical-domain pretraining had to arrive together.

Component priors are modality-specific, and their sign can flip. Retaining the largest connected component is right for the 3D tasks, whose targets are single-component by construction (TEE Hausdorff 12.69 → 10.87 mm). On video it is wrong in the other direction, since the two leaflets legitimately appear as disjoint components; we had replaced it with a small-component cleanup – and on the hidden test, *removing that too* helped. Such priors are standard for compact, high-contrast, effectively single-component targets – projection-domain metal segmentation in cone-beam CT is the same shape of problem [1,17] – but the video leaflets satisfy none of those assumptions.

6 Conclusions

Our deployed MVAA 2026 system – one offline container segmenting the mitral valve in cardiac CT, 3D TEE, and surgical video – reaches hidden-test DSC **0.844** on CT (4.94 mm HD, 0.416 mm ASD) and **0.815** on video (243.3 px, 105.2 px): 3rd of 36 teams on CT, first on CT Hausdorff, 6th overall. Encoder pretraining remains the dominant lever on video, and self-configuring planning transfers across modalities without tuning. But three interventions our own validation evidence rejected improved the ranked score, each for a different reason: the video board contains no empty frames, the 30-case CT board cannot resolve rare catastrophic cases, and our own video oracle was no more reliable than the board. A challenge validation split should be audited for the failure modes it can express before its verdicts are used as gates.

References

1. Agrawal, H., Hietanen, A., Särkkä, S.: Deep learning based projection domain metal segmentation for metal artifact reduction in cone beam computed tomography. IEEE Access **11**, 100371–100382 (2023). https://doi.org/10.1109/ACCESS.2023.3314700
2. Cardoso, M.J., Li, W., Brown, R., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Carnahan, P., Moore, J., Bainbridge, D., Eskandari, M., Chen, E.C.S., Peters, T.M.: DeepMitral: Fully automatic 3D echocardiography segmentation for patient specific mitral valve modelling. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). LNCS, vol. 12904, pp. 459–468. Springer (2021), source of the MVSeg 3D TEE volumes
4. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J., Li, S.: Mitral valve anatomy analysis using multimodal imaging data. Zenodo, International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2026). https://doi.org/10.5281/zenodo.19726755
5. Huang, Z., Ye, J., Wang, H., Deng, Z., Li, T., He, J.: Exploiting pseudo-labeling and nnU-Netv2 inference acceleration for abdominal multi-organ and pan-cancer segmentation. In: Fast, Low-resource, and Accurate Organ and Pan-cancer Segmentation in Abdomen CT (FLARE 2023). LNCS, vol. 14544. Springer (2024)
6. Iakubovskii, P.: Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2019)
7. 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)
8. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). LNCS, vol. 15009, pp. 488–498. Springer (2024)
9. Jaspers, T.J.M., et al.: Scaling up self-supervised learning for improved surgical foundation models. Medical Image Analysis **108**, 103873 (2026). https://doi.org/10.1016/j.media.2025.103873, introduces the SurgeNet/SurgeNetXL surgical foundation models
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
11. Ma, J., Zhang, Y., Gu, S., et al.: Fast and low-GPU-memory abdomen CT organ segmentation: The FLARE challenge. Medical Image Analysis **82**, 102616 (2022)
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV). pp. 565–571 (2016)
13. MVAA Challenge Organizers: MVAA 2026: Mitral valve anatomy analysis using multimodal imaging data. https://www.codabench.org/competitions/15662/ (2026), mICCAI 2026 Challenge
14. MWM Workshop Organizers: 1st international workshop on medical world models (MWM). MICCAI 2026 Workshop (2026)
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) (2024), arXiv:2304.07193

16. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (2021)
17. Rautio, S., Meaney, A., Latva-Äijö, S.M., Agrawal, H., Brix, M., Jayakody, D., Siltanen, S.: Complex wavelet-based sinogram segmentation for metal artifact reduction in cone-beam CT. Physics in Medicine and Biology (2026). https://doi.org/10.1088/1361-6560/ae8fb3
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015)
19. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P.W., Bauer, U., Menze, B.H.: clDice - a novel topology-preserving loss function for tubular structure segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16560–16569 (2021)
20. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML). pp. 6105–6114 (2019)
21. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
22. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 23–30 (2017)
23. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142 (2023)
24. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
25. Yang, L., Zhao, Z., Zhao, H.: UniMatch V2: Pushing the limit of semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 3031–3048 (2025). https://doi.org/10.1109/TPAMI.2025.3528453
26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis (DLMIA). pp. 3–11. Springer (2018)
27. Zia, A., Berniker, M., Perreault, C., Nespolo, R., Jarc, A.: Surgical visual understanding challenge (SurgVU). Zenodo, Medical Image Computing and Computer Assisted Intervention (MICCAI) (2024). https://doi.org/10.5281/zenodo.14054184