Solution for MVAA challenge 2026: Ensemble, Tool-paste Augmentation and Tool-free View Generation

Shunsuke Kikuchi[0009−0008−3353−657X]

Jmees Inc., 5-4-6 Kashiwanoha, Kashiwa-city, Chiba, Japan
shunsuke.kikuchi@jmees-inc.com

Abstract. We describe our solution to the MICCAI 2026 Mitral Valve Anatomy Analysis (MVAA) challenge, which asks participants to segment the mitral valve across three heterogeneous imaging settings: (i) 3D cardiac CT under a semi-supervised regime, (ii) 3D transesophageal echocardiography (TEE) under full supervision, and (iii) 2D surgical endoscopic video under a semi-supervised regime. Rather than committing to a single architecture, we treat every task independently and combine self-configuring 3D nnU-Net models, a matched reimplementation of its plain-convolution U-Net, SwinUNETR and 2.5D ImageNet-pretrained encoders through probability-level ensembling with mirror test-time augmentation (TTA) and connected-component post-processing. For CT we grow this to an eight-source ensemble with leakage-free pseudo-labels over 973 unlabeled volumes. For TEE we exploit a non-overlapping external echocardiography corpus to enlarge the training set. For video we use a multi-architecture ensemble regularized by a Mean Teacher, refined with ensemble pseudo-labels, instrument copy-paste augmentation and instrument-free view generation. A recurring finding, quantified across many submissions, is a weak and sometimes inverted correlation between local cross-validation and the hidden test set. Our final system attains validation-leaderboard Dice of 0.863/0.863/0.823 and hidden-test Dice of 0.838/0.887/0.836, ranking at or near the top on all three tasks. All source code and model weights are available at `https://github.com/ShunsukeKikuchi/MVAA26-jmees-sol.git`

Keywords: Mitral valve segmentation · Semi-supervised learning · Model ensembling

1 Introduction

The mitral valve is imaged with very different modalities along a clinical pathway—volumetric CT for pre-operative anatomy, transesophageal echocardiography (TEE) for functional assessment, and endoscopic video for intra-operative visualization—so a system that segments it across all three is of practical value. The MICCAI 2026 Mitral Valve Anatomy Analysis (MVAA) challenge [4] formalizes this as three parallel tasks scored jointly by Dice (DSC) and boundary distances (HD,

ASD); we assume familiarity with its data and protocol. Each task carries a distinct difficulty: CT (Task 1) is semi-supervised with few labels and many unlabeled volumes; TEE (Task 2) is fully supervised but low-contrast; video (Task 3) has sparse key-frame labels, frequent instrument occlusion, and—as we found—systematic annotation gaps. Our guiding principle is empirical and modality-specific: we pick the strongest configuration per task and combine complementary models by probability-level ensembling. Our contributions are: (1) a three-task pipeline mixing self-configuring nnU-Net, custom plain-convolution U-Nets, SwinUNETR and 2.5D encoders under a shared test-time-augmentation and post-processing protocol; (2) domain interventions—a non-overlapping external echocardiography corpus for TEE, and (3) data augmentations focused on surgical instruments—tool copy-paste augmentation and rule-based tool-free view generation.

2 Methods

2.1 Common Framework

Cross-validation and folds. All models use a shared 5-fold split for CT and TEE, persisted as a versioned fold assignment file so that every experiment is comparable. For video we use a recording-wise leave-one-out scheme over the six labeled recordings; one recording is excluded from cross-validation aggregation (while retained for training) because its footage is visibly different in appearance from the other recordings.

Self-configuring baselines. For the volumetric tasks we adopt nnU-Net v2 [6] in its `3d_fullres` configuration, which infers patch size, spacing, normalization and network topology from a dataset fingerprint. In parallel we reimplement the nnU-Net design as a custom plain-convolution 3D U-Net [3] (denoted plain-conv) trained with our own loop, which gives full control over augmentation and pseudo-labeling while matching nnU-Net accuracy. Networks are trained with a combined Dice and cross-entropy loss [13, 10], mixed-precision (AMP), deep supervision where applicable, and cosine or polynomial learning-rate decay. Seeds are fixed and checkpoints are resumable.

Inference protocol. Volumetric predictions use sliding-window inference with 0.5 overlap. Unless stated otherwise we apply 8-way test-time augmentation (TTA) by mirroring along all three axes [18], average the resulting softmax probabilities, resample probabilities back to the original image grid by trilinear interpolation, take the argmax, and retain only the largest connected component (largest-CC). Largest-CC is essential for the boundary metrics: it reduces the (95th-percentile) Hausdorff distance from $\sim$7 mm to $\sim$4 mm (CT) and $\sim$37 mm to $\sim$10 mm (TEE) by removing isolated false-positive islands. Distances are computed in physical millimeters on the original image grid. The final CT system is the one exception: with its ensemble in place, mirror TTA no longer changed the result, so its 3D members run without it—a variant we verified on the hidden test set

(−0.0003 DSC at 1.61× the throughput). The full system totals ∼16 GB of weights (40 CT, 15 TEE and 15 video networks) and ran within the challenge's single-GPU Docker budget; a reader wanting a lighter system can keep the four-source CT ensemble, which already reaches LB 0.863.

All headline numbers are recomputed in the original image space with true spacing and largest-CC; in-training metrics (which omit spacing and post-processing) are used only for early stopping.

2.2 Task 1 – CT Segmentation

Ensemble of eight sources. Our final CT model averages the probabilities of eight sources with equal weight, each averaged over the 5 folds (40 networks): nnU-Net `3d_fullres`; the plainconv U-Net, trained with the pseudo-labels below; two SwinUNETR [5] models, one from scratch and one from a BTCV-pretrained encoder [16]; three 2.5D UNet++ [21] models with an EfficientNet-B4 [15] encoder; and a hybrid with a 2D ResNet-RS encoder, parameter-free depth pooling and a light 3D decoder. Ensembling is what carries CT: the combination reaches out-of-fold DSC 0.868, HD 3.85 mm and ASD 0.236 mm, well beyond any of its members (Table 1).

2.5D members. These stack three adjacent slices as RGB channels for a 2D encoder and predict the centre slice, importing ImageNet-pretrained diversity no 3D member has. Since the mitral valve appears in a canonical orientation in these volumes, mirror flips are harmful: we drop flip augmentation and inject in-plane oblique tilts (±15°) instead, using rotation- rather than flip-based TTA, which is the largest single-model lever in this line (0.8367 → 0.8488 OOF). The sagittal plane was best of the three; the members are it, it with oblique tilts, and it with oblique tilts plus distortion. Removing flips from the *3D* members helps too, but only together with elastic warping (+0.0026 DSC, −0.94 mm HD); each alone is within noise, since 3-axis flipping multiplies the pose distribution by eight and swamps a sub-millimetre warp.

Leakage-free pseudo-labeling. CT offers 27 labeled and ∼1040 unlabeled volumes. Our first attempt concluded that pseudo-labeling does not transfer; that was an artifact of an *inverted* leave-one-fold-out, which for fold f averaged the models whose validation fold was *not* f—exactly the four that had trained on fold f's validation cases—and discarded the one clean model. The tell was that "removing" the leak raised CV by +0.006 instead of lowering it. Corrected, the teacher for fold f is the average of the fold-f models of each architecture, all of which held fold f out; teacher accuracy is then directly measurable, so we pick it by measurement rather than assumption (plainconv + oblique 2.5D, 0.8632, adopted over a 5.5× costlier variant worth +0.0011). Pseudo-labels are the largest connected component of $p > 0.5$ over 973 volumes; the student gains +0.0058 DSC in 5/5 folds. Beyond this the task is saturated: no volume contains a missed region and 43% of ground-truth voxels are surface voxels, so the entire error budget is boundary placement, and threshold tuning, an agreement filter and case-adaptive windowing all moved the metric by <0.0005.

2.3 Task 2 – TEE Segmentation

TEE is fully supervised but small. Our final model is a three-source probability-average ensemble: nnU-Net (5 folds), the plainconv U-Net (5 folds), and a plainconv variant trained with 50 additional real, labeled 3D TEE volumes from the public MVSeg 2023 mitral-valve challenge corpus [1]. Overlap with MVAA was checked by file-content hashing: 20 MVSeg volumes are bit-identical to MVAA validation images (the two challenges share sources) and were discarded together with the rest of that split; the 50 retained volumes have no hash match against any MVAA training or validation image, and the exact case list ships with our code. The external cases are passed through the identical challenge preprocessing (spacing and normalization); no re-annotation was performed. As an ablation, the external data alone improves the plainconv model by +0.010 DSC, −0.9 mm HD and −0.06 mm ASD on CV and +0.003 DSC on the LB, and the three-way ensemble improves further. All members use 8-way TTA and largest-CC.

2.4 Task 3 – Endoscopic Video Segmentation

Backbone and ensemble. The strongest single model is a UNet++ [21] decoder with an EfficientNet-B7 [15] encoder pretrained with Noisy-Student weights [19], preferred over DeepLabV3+ [2] in preliminary trials. Frames are treated as 2D and predicted independently. Our final video model probability-averages three architecturally diverse encoders—EfficientNet-B7, MaxViT-L [17] and a ConvNeXt-L [9] encoder pretrained with DINOv3 [14]—each trained with the same semi-supervised recipe below and averaged over 5 folds. The transformer/ConvNeXt members are weaker alone but add diversity: the three-architecture ensemble improves the leaderboard over the B7 model (+0.0024 DSC), whereas mixing in a weaker generator (Task 3 pseudo variant) hurt it.

Semi-supervision. A supervised baseline is regularized by a Mean Teacher: an exponential-moving-average teacher supervises the student on unlabeled frames via a consistency loss. The teacher notably fixes a common failure on truly empty frames, reducing spurious false positives. We then move to a two-stage ensemble pseudo-labeling scheme: for each fold, the pseudo-labels of the unlabeled frames are produced by averaging the probabilities of the models trained on the other folds (leave-one-out), thresholded at 0.50, and the network is retrained on labeled plus pseudo-labeled frames. Probability-level ensembling of four teachers lowers the pseudo-label foreground rate from 87% to ∼72%, suppressing false positives, and yields +0.017 DSC over the Mean Teacher.

Instrument copy-paste augmentation. Surgical instruments frequently occlude the valve. We therefore add a copy-paste augmentation ("ToolPaste", Fig. 1) built from a library of 296 alpha-matted surgical-instrument cut-outs harvested from two external robotic-surgery datasets, SAR-RARP50 [12] (161 instances)

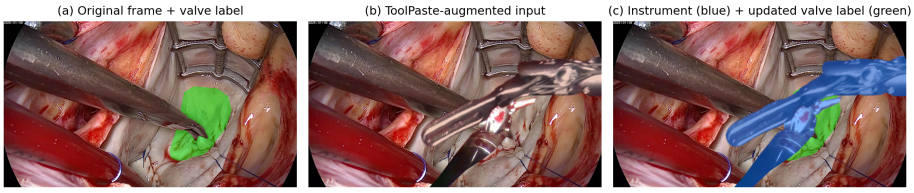

Fig. 1. ToolPaste augmentation on a real labeled video frame. (a) Original frame with the annotated mitral valve (green). (b) The ToolPaste-augmented frame that the network actually receives as input: a color-matched surgical-instrument cut-out is pasted so that it occludes the valve. (c) The same augmented frame with the pasted instrument overlaid in blue and the updated valve label in green (occluded pixels removed); the color matching otherwise makes the pasted instrument hard to spot. This forces the network to segment the valve under realistic tool occlusion.

and SurgToolLoc [22] (135 instances). Each cut-out uses the dataset's ground-truth instrument mask as its alpha channel (lightly eroded and feathered); for SAR-RARP50, which lacks instance masks, we take connected components over the union of instrument-part classes (excluding needle, thread and catheter), whereas SurgToolLoc provides instance masks directly. Instances are chosen in two stages: an automatic score keeps a few well-framed, moderately sized, sharp and hash-deduplicated frames per clip, after which we visually review contact sheets and retain only clean, representative instances.

Because instruments enter from a frame edge (not a corner) and point inward, we normalize each cut-out's pose: we take its base as the frame edge with the longest mask contact and its tip as the deepest interior pixels, and store the resulting base→tip axis. At paste time we sample one or two cut-outs, rotate and scale each so that it enters from a randomly chosen edge with its tip reaching a randomly chosen valve pixel, color-match it to the local region (luminance and white-balance gain), and alpha-blend it in. Crucially, wherever the opaque tool covers the valve we remove those pixels from the mask, reproducing the true occlusion relationship rather than asking the network to hallucinate through the instrument.

Tool-free view synthesis. We also synthesize instrument-free views of the labeled frames with our previously proposed stitching-based algorithm [7] (Fig. 2). Because instruments move between frames, anatomy occluded at frame i is often visible at a neighbor j: we match keyframes with learned local features (ALIKED [20] + LightGlue [8]), build a shared canvas excluding instrument pixels, warp it back into each frame so the instrument region is filled with borrowed anatomy, and composite with Poisson blending [11]. This fills 92% of instrument pixels on average (per-recording means 0.85–0.98); the remaining ∼8%—tools that never move, e.g. a static retractor—are not hallucinated but keep the original instrument pixels, so no special loss masking is needed. Isolating the two instrument interventions on a common fold-0 supervised baseline (0.880): Tool-

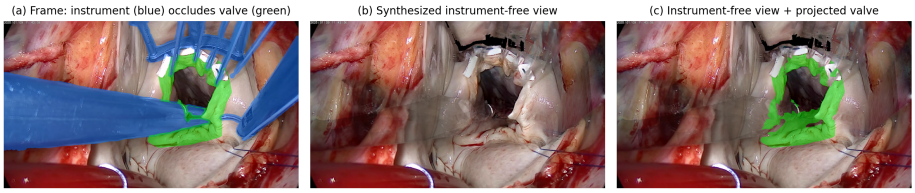

Fig. 2. Tool-free view synthesis on a real labeled frame. (a) Original frame: surgical instruments (blue) occlude the valve, whose annotation (green) partly falls under the tools. (b) Synthesized instrument-free view: instrument pixels are filled with anatomy warped from neighboring frames via feature matching and Poisson blending; unfillable regions remain as holes. (c) The instrument-free view with the valve label projected from neighboring frames (green), consistent with the synthesized anatomy.

Paste alone +0.013 DSC / −5.4 px HD, tool-free views alone −0.007, yet adding the tool-free model to the ensemble is +0.003 over ToolPaste alone—individually weak but complementary.

Annotation-quality finding. During error analysis we observed label-quality issues in the video annotations. Many masks carry spurious thin slivers of valve label that cling to the contours of surgical instruments crossing the valve—false-positive fragments tracing tool edges rather than true valve tissue. Because these artifacts sit exactly where instruments occlude the valve, they teach the network to paint the valve label along tool boundaries, and they inflate the boundary-distance metrics. Our instrument copy-paste augmentation, which removes the mask under a pasted tool, partly counteracts this bias by supplying occlusion examples with correct (tool-free) valve labels.

3 Experiments and Results

We report 5-fold CV for CT and TEE and recording-wise leave-one-out for video, all with the original-space, largest-CC pipeline; LB represents CodaBench validation-phase scores.

Table 1 collects results for all three tasks. For CT, single models plateau around DSC 0.857–0.863 (SwinUNETR and higher-resolution variants match plainconv on Dice but are worse on HD), so architectural diversity helps only through ensembling; doubling the ensemble from four to eight sources is the strongest lever we found. For TEE, the external MVSeg cases improve the single model on every metric and the three-way ensemble reaches LB 0.863 (its LB exceeds the CV probe because the submitted members use 5 folds). For video, semi-supervision and ensemble pseudo-labeling give large, monotone CV gains; the ToolPaste+ToolFree+pseudo variant reaches LB 0.823, our best video result.

Qualitative results. Fig. 3 shows representative validation predictions against ground truth: on CT the thin curvilinear leaflets are recovered with tight bound-

Table 1. Per-task results. CV is 5-fold (CT, TEE) or recording-wise leave-one-out (task3); LB is the CodaBench validation leaderboard, and "–" marks configurations not submitted alone during the validation phase (the full CT ensemble went only to the test phase, Section 3.1). All entries use 8-way TTA (video: 4-way) and, for CT/TEE, largest-CC in the original image space. Distances are in mm for CT/TEE and pixels for video.

	CV			LB		
Model	DSC	HD	ASD	DSC	HD	ASD
Task 1 – CT (distances in mm)						
plainconv (5-fold)	0.857	4.06	0.255	0.857	4.64	0.281
nnU-Net (5-fold)	0.861	4.14	0.247	–	–	–
plainconv + leak-free pseudo (5-fold)	0.863	3.85	0.244	–	–	–
Ensemble (nnU+plainconv+2×Swin)	0.863	4.08	0.245	0.863	4.42	0.254
+ two 2.5D models	0.866	3.97	0.239	**0.8634**	**4.42**	0.255
Full ensemble + pseudo (final)	**0.868**	**3.85**	**0.236**	–	–	–
Task 2 – TEE (distances in mm)						
plainconv (5-fold)	0.836	10.22	0.636	0.854	9.99	0.566
plainconv + external (5-fold)	0.846	9.33	0.580	0.857	10.01	0.557
Ensemble (nnU+plainconv+ext)	0.853	9.65	0.498	**0.863**	**9.59**	**0.526**
Task 3 – Video (distances in px)						
supervised (UNet++/B7)	0.867	47.8	17.8	–	–	–
+ Mean Teacher (semi)	0.873	47.5	17.9	–	–	–
+ ensemble pseudo	0.890	50.2	23.1	0.809	71.2	11.8
+ ToolPaste + ToolFree + pseudo (B7)	0.890	50.7	23.3	**0.823**	**67.9**	**11.2**

ary agreement; on TEE the annulus ring is delineated despite low contrast; on video the valve is segmented even under partial instrument occlusion.

3.1 Cross-Validation vs. Leaderboard, and Final System

Across submissions we repeatedly observed a weak or inverted CV→LB relationship. Over the configurations for which we hold both scores, the within-task rank correlation between CV and LB DSC was slightly negative (Spearman $\rho = -0.30$ for CT, $n = 5$; $\rho = -0.20$ for video, $n = 4$; neither significant)—CV rank carries essentially no signal about LB rank once configurations were within a few 10^{-2} of each other. For CT, adding two 2.5D members that clearly raised out-of-fold DSC edged the CT LB to only 0.8634. We attribute the mismatch to small validation sets and wide confidence intervals; our recommendation, and the policy we followed, is to use CV only as a coarse filter (discard clear regressions), and to rank surviving candidates exclusively by controlled one-variable-at-a-time submissions. The final system—the CT, TEE and three-architecture video ensembles—attains LB DSC 0.863/0.863/0.823, ranking first on video and within a few ten-thousandths of the leaders on CT and TEE, and hidden-test DSC 0.838/0.887/0.836. On the test phase we changed one variable per slot,

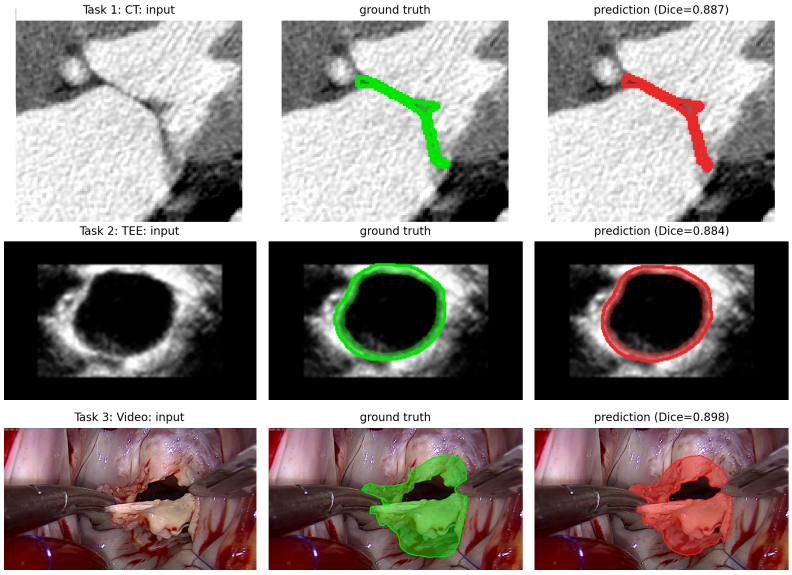

Fig. 3. Qualitative prediction-vs-ground-truth comparison on held-out validation data, one row per task: input (left), ground truth (green), and our prediction (red) with per-case Dice. Rows: Task 1 CT axial slice, Task 2 TEE slice, Task 3 endoscopic frame.

keeping the TEE and video components byte-identical so that their reproduction to six decimals attributes the CT differences to CT. Video confirmed there that its CV ranks configurations wrongly (CV-best 0.899 scored 0.803, against 0.836 for the LB-preferred one).

4 Discussion and Conclusion

We presented a modality-specific, ensemble-centric solution to the MVAA 2026 challenge. Two lessons emerge. First, heterogeneous-encoder ensembling with a disciplined TTA and CC protocol reliably gains both overlap and boundary accuracy across different modalities, and pseudo-labeling and instrument augmentations help where annotations are scarce—provided the leave-one-fold-out separation is correct, since an inverted one silently maximizes the leak it is meant to remove. Second, and most operationally important, local cross-validation was an unreliable predictor of the hidden test set, over-estimating positive effects while preserving their sign; we therefore selected models by transfer behavior and controlled, one-variable-at-a-time submissions rather than by CV rank.

Disclosure of Interests. The author have no competing interests to declare that are relevant to the content of this article.

References

1. 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: 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. pp. 459–468. Springer International Publishing, Cham (2021)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 833–851 (2018)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9901, pp. 424–432. Springer (2016)
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 2026 (MICCAI) (2026). https://doi.org/10.5281/zenodo.19726755
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop (BrainLes). LNCS, vol. 12962, pp. 272–284. Springer (2022)
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)
7. Kikuchi, S., Kouno, A., Matsuzaki, H.: Stitch-inferencer: Enhance endoscopic video segmentation and tracking via panoramic reconstruction (2026), `https://arxiv.org/abs/2607.14968`
8. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17627–17638 (2023)
9. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
10. 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. IEEE (2016)
11. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing **22**(3) (2003)
12. Psychogyios, D., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge (2024), `https://arxiv.org/abs/2401.00496`
13. 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)
14. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, et al.: DINOv3 (2025), `https://arxiv.org/abs/2508.10104`
15. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML). PMLR, vol. 97, pp. 6105–6114 (2019)

16. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of Swin transformers for 3D medical image analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20730–20740 (2022)
17. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer (2022)
18. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
19. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with Noisy Student improves ImageNet classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10687–10698 (2020)
20. Zhao, X., Wu, X., Chen, W., Chen, P.C.Y., Xu, Q., Li, Z.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. vol. 72, pp. 1–16 (2023)
21. 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). LNCS, vol. 11045, pp. 3–11. Springer (2018)
22. Zia, A., et al.: Intuitive surgical surgtoolloc challenge results: 2022-2023 (2025), https://arxiv.org/abs/2305.07152