From Promise to Practice: Benchmarking Vision Foundation Models for Semantic Segmentation of Industrial Defects

Christof Weickhardt*, Gregor Hammer*, Susanne Suter, and Moritz Kirschmann

School of Computer Science, University of Applied Sciences and Arts Northwestern Switzerland, Brugg-Windisch, Switzerland
gregor.hammer@students.fhnw.ch
christof.weickhardt@students.fhnw.ch

Abstract. In computer vision, machine learning extracts image information, and semantic segmentation labels each pixel to localize industrial defects. Training requires expert-annotated masks, while foundation models promise label-efficient transfer. Across nine domains, zero-shot use, domain-specific models at nested label budgets, and one jointly trained model tested how that promise translates into practice. A frozen foundation encoder with a trained head and low-rank adaptation were compared with fully fine-tuned convolutional models using Defect-IoU, which measures predicted–annotated defect overlap. Zero-shot use detected no defect pixels, while additional labels yielded domain-dependent Validation gains. In the capped specialist–joint comparison, joint training increased the eight-domain Test mean for both foundation-model adaptations but reduced it for convolutional models. Separately, the best fully fine-tuned convolutional specialist achieved 49.1% Validation Defect-IoU on proprietary Injection Molding with 45 labeled training images. Under the evaluated conditions, pretraining did not guarantee superior segmentation; outcomes depended on adaptation and domain-specific supervision.

Keywords: Semantic segmentation; vision foundation models; industrial defect detection; smart manufacturing; multi-domain training; data efficiency

1 Introduction

Industrial inspection needs both defect detection and localization. Image-level classification reports whether a product is defective, whereas semantic segmentation labels each pixel and captures defect extent and shape. This matters for small or irregular regions that one image label cannot localize.

Pixel-level supervision is demanding because defect boundaries must be outlined and verified. Pretrained visual models may reuse learned representations with fewer task-specific labels, but their label efficiency in industrial inspection must be measured.

Broad pretraining may not transfer reliably to small, fine-grained, and domain-specific defects, so target labels and adaptation may remain necessary. Pooling domains adds observations but may also introduce conflicting products, textures, and defect morphologies. Both adaptation and training scope therefore remain empirical questions.

The benchmark asks how strategies with different amounts and locations of adaptation behave as target labels increase, and whether joint training helps relative to one specialist per domain. All nine domains share the binary labels `defect` and `background`, enabling visual-transfer comparisons without taxonomy reconciliation [1]. Eight are 3CAD

* These authors contributed equally.

product families from computer, communication, and consumer electronics manufacturing [2]; one is a proprietary injection-molding domain.

Four adaptation regimes are represented by five configurations because full fine-tuning uses two convolutional neural networks (CNNs). Fixed-prompt SAM3 provides zero-shot segmentation [3]; a frozen self-supervised DINOv3 encoder supplies features to a learned head [4]; SAM3 uses low-rank adaptation (LoRA) [5]; and DeepLabv3+ and SegNeXt-T represent full fine-tuning [6, 7]. Nested specialist budgets and domain-specific versus joint training provide label-efficiency curves and paired held-out Test evidence. A fixed-epoch pooled SAM3 follow-up examines adaptation depth.

2 Methodology

A common harness makes the evaluated configurations comparable across domains, label budgets, and training scopes.

Benchmark workflow and evaluation pipeline. One pipeline evaluates every configuration. It fixes the binary ontology, group-aware splits, nested budgets, tile geometry, threshold, and metrics. Its primary metric is foreground intersection over union (Defect-IoU) [8]. Each model produces one defect logit per pixel, and the evaluator reconstructs native-resolution masks. Validation selects checkpoints. Public tests remain sealed until the protocol is fixed. Hash-addressed configurations and run records provide provenance (Fig. 1).

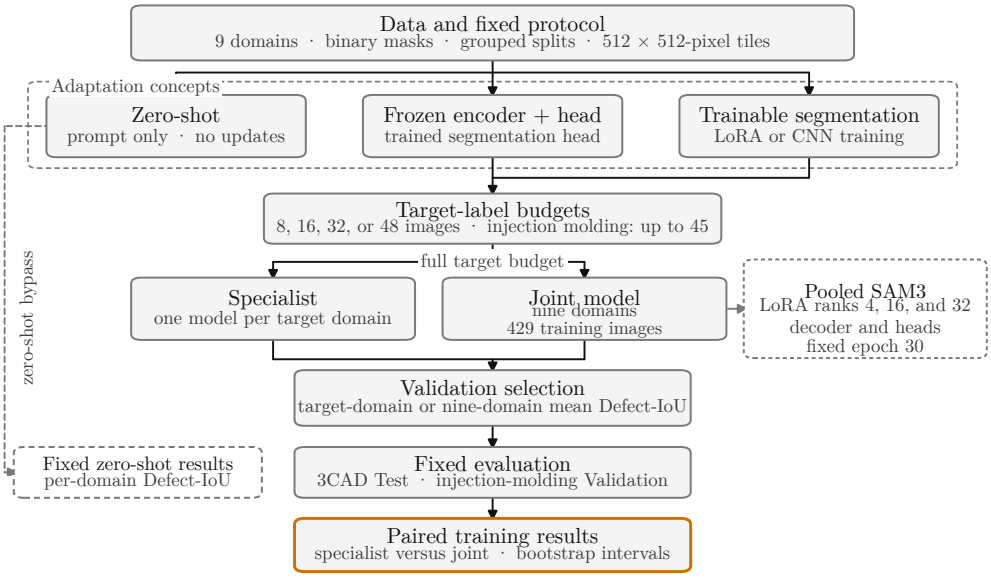


Fig. 1. Benchmark workflow. Nested budgets precede the full-budget specialist–joint comparison; zero-shot bypasses training. Injection molding is Validation-only, while pooled SAM3 also reports fixed epoch-30 3CAD Test results.

Domains and splits. Eight public 3CAD product families and one proprietary injection-molding line form nine binary defect domains (Table 1). Anomaly labels map to

`defect`, and verified negatives map to `background`. The injection-molding domain contains 45 training images from series A and C and 15 Validation images from series B. It has no independent Test set. With seed 42 and group-constrained stratification, we split the official 3CAD test images once into 60% training, 20% Validation, and 20% sealed Test subsets. Images labeled `good` in the official training split remained training-only negatives. Near-duplicate groups never crossed splits.

Table 1. Nine domains. Available supervised images include labeled positives and verified negatives. Evaluation images are sealed Test data for 3CAD and Validation data for injection molding. Median defect share is the median foreground-pixel proportion among positive images; full target budgets are 48 and 45 images, respectively.

Code	Product family	Available	Evaluation	Median defect
3CAD–ACC	Aluminum Camera Cover	2,230	289	0.96%
3CAD–AI	Aluminum iPad	4,143	409	0.03%
3CAD–AMF	Aluminum Middle Frame	3,027	295	0.09%
3CAD–ANMF	Aluminum New Middle Frame	2,478	281	0.15%
3CAD–ANI	Aluminum New iPad	7,169	987	0.14%
3CAD–AP	Aluminum component	4,859	632	0.09%
3CAD–CS	Copper Stator	1,368	191	0.35%
3CAD–IS	Iron Stator	1,765	222	0.82%
Injection	Injection molding (proprietary)	60	15	1.19%

Adaptation and training. The four core regimes yield five configurations. SAM3 zero-shot uses prompt `"defect"` and threshold 0.5. DINOv3 freezes its small Vision Transformer with 16×16-pixel patches and trains a linear head. SAM3 LoRA freezes its base and trains rank-4 adapters. DeepLabv3+ and SegNeXt-T are fully fine-tuned.

All trainable models use the Adam optimizer with decoupled weight decay [9], focal plus soft Dice loss [10, 11], an effective batch size of eight, no augmentation, and Validation every five epochs. LoRA also uses image-level defect-presence loss. Predefined early stopping applies within caps of 50 epochs for DINOv3 and DeepLabv3+, 30 for LoRA, and 100 for SegNeXt-T.

Checkpoint selection depends on scope. Specialists select the highest domain Validation score. Joint models select the highest scheduled nine-domain Validation mean and use the worst domain as a tie-break. Training uses deterministic 512×512-pixel tiles. Evaluation reconstructs native masks with a 384-pixel stride, overlap fusion, and threshold 0.5.

Nested target-label budgets. With seed 42, specialist curves use 8, 16, 32, and the full training cohort (48 images; 45 for injection molding). Each larger manifest contains the smaller ones. All budgets share the training configuration and evaluator. This analysis is Validation-only. The 36-run decoder-and-head specialist extension spans the same nine domains and four budgets. It selects checkpoints on Validation within 30 epochs and is reported descriptively in Fig. 2.

Specialist versus joint. At full budget, each 3CAD specialist uses 48 images, and the injection-molding specialist uses 45. For each target, the *nine-domain joint model* combines those images with every other domain. It uses 429 unique source images under domain-balanced sampling. Budgets count source images, not tiles. Every checkpoint emits one binary output without domain-specific heads or thresholds. For target d, the

paired effect is

$$\Delta_d = \text{Defect-IoU}_d(\text{joint}) - \text{Defect-IoU}_d(\text{specialist}). \quad (1)$$

For each seed-42 3CAD checkpoint pair, 2,000 bootstrap resamples at the image-group level provide percentile 95% intervals [12]. They quantify fixed-checkpoint evaluation variation, not retraining variation.

Pooled SAM3 follow-up. At seed 42, a separate full-pool joint experiment compares LoRA ranks 4, 16, and 32 with decoder-and-head tuning at a common 30-epoch endpoint. All variants share data, sampler, loss, prompt, evaluator, and pretrained checkpoint. They use 20,468 training images and 76,770 optimizer updates. Direct tuning updates the transformer decoder, segmentation head, and dot-product scoring; encoders and encoder-side fusion remain frozen. The four epoch-30 checkpoints are evaluated on eight held-out 3CAD Test domains and separately on injection-molding Validation. For the decoder arm, epoch 30 also maximizes the scheduled five-epoch Validation measurements within the 30-epoch budget. This comparison has no matched full-pool specialist arm; the decoder specialists in Fig. 2 belong to the nested-budget extension.

Metrics and provenance. Defect-IoU divides the predicted and reference defect pixels in common by their union. Counts are pooled over each complete split. A prediction without foreground scores zero when positive reference pixels exist. Background is excluded because median positive-image foreground fractions are only 0.03–1.19%. Pixel accuracy can therefore remain high when defects are missed. Diagnostic metrics include precision, recall, Dice, generalized Dice, boundary IoU, 95th-percentile Hausdorff distance, and average symmetric surface distance [13–15]. These diagnostics describe complementary errors but do not select checkpoints. Hashes link each result to its split, budget, tiles, configuration, checkpoint, and evaluator. The campaign contains 140 core, 36 decoder-specialist, one capped-joint decoder, and three full-pool SAM3 runs.

3 Benchmark Results Across Training Settings

Results move from zero-shot and specialist Validation to the full-budget specialist–joint Test comparison and pooled SAM3 follow-up.

3.1 The Tested SAM3 Zero-Shot Setting Detected No Defect Pixels

With prompt `"defect"` and threshold 0.5, SAM3 predicted no defect pixels in both evaluation settings: eight 3CAD Test domains and injection- molding Validation. Defect-IoU was therefore 0.0% throughout. This result applies to the tested prompt–threshold pair. It does not estimate the best zero-shot performance obtainable through prompt or threshold tuning.

3.2 Validation Across Specialist and Joint Training

Figure 2 brings together the available seed-42 Validation pathways for all nine domains. The specialist region retains the actual target-domain budgets. The separate joint region reports the capped 429-image pool and the full 20,468-image pool. Outcomes vary by model and domain, and moving to joint training can help or harm a target. The figure is therefore a descriptive comparison of training regimes rather than one continuous sample-size experiment. The core checkpoints follow their original protocol selection. Decoder

specialists use the best checkpoint within 30 epochs. Both joint decoder points use epoch 30: capped is fixed, whereas full pool is Validation-selected. Their connector compares independent regimes (1,620 versus 76,770 updates), not continued training or a data-only effect.

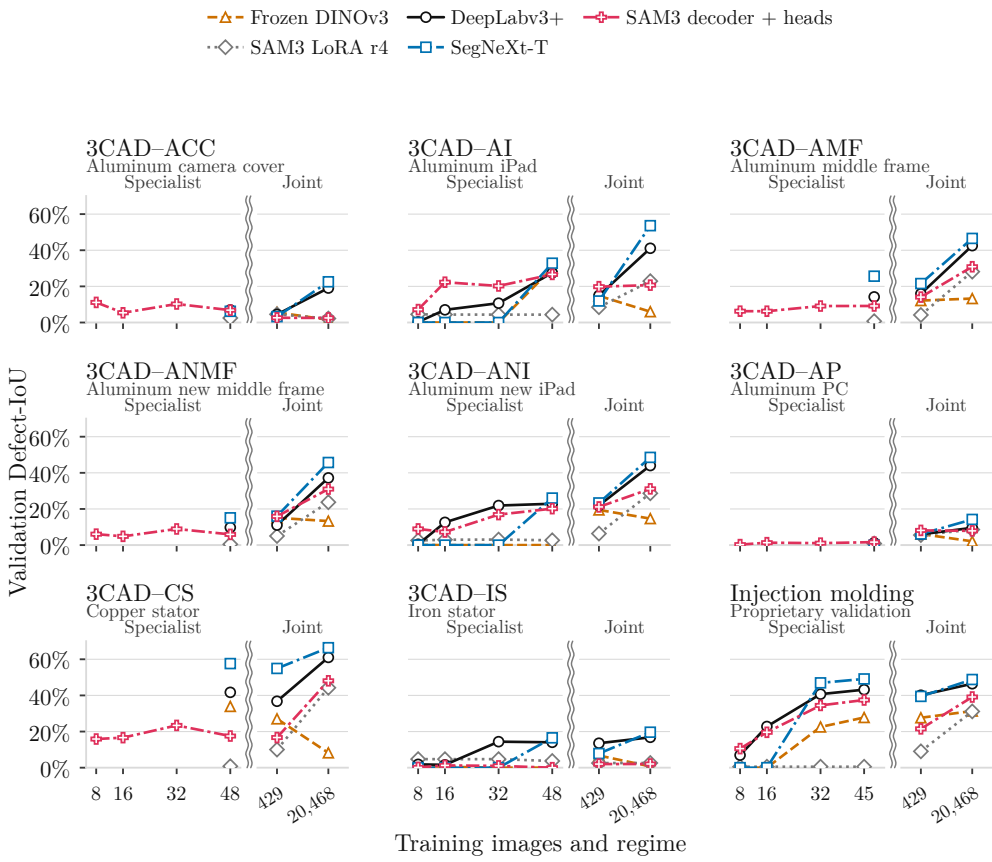


Fig. 2. Seed-42 Validation pathways for all nine domains. Specialists use 8, 16, 32, and 48 target images (45 for injection molding); joint subaxes show 429- and 20,468-image pools. The double wave changes scope and x-scale, so lines do not form a continuous dose-response. Core-model 8/16/32 runs exist only for 3CAD–AI, 3CAD–ANI, 3CAD–IS, and injection molding; blank cells elsewhere were not run. Both decoder joint points use epoch 30, with different selection rules described in the text.

3.3 Joint-Training Effects Differ Across Strategies

At full target budget, joint-training effects depended on strategy (Table 2; Fig. 3). Joint training increased the 3CAD Test mean for frozen DINOv3 and SAM3 LoRA but reduced it for both CNNs. SegNeXt-T specialists retained the highest absolute mean at 20.4%, while LoRA improved all eight domain point estimates despite low absolute quality.

Table 2. Full-budget specialist and nine-domain joint results. Values are Defect-IoU; change is joint minus specialist in percentage points. The 3CAD Test columns are unweighted means over eight held-out domains, whereas Injection molding is reported separately on Validation.

	3CAD Test mean			Injection-molding Validation		
Model	Specialist	Joint	Change	Specialist	Joint	Change
DINOv3 frozen	7.4%	14.9%	+7.5	27.8%	27.6%	−0.1
SAM3 LoRA	1.8%	6.0%	+4.2	0.5%	9.1%	+8.5
DeepLabv3+	17.2%	14.8%	−2.4	43.2%	40.1%	−3.1
SegNeXt-T	20.4%	18.2%	−2.2	49.1%	39.5%	−9.7

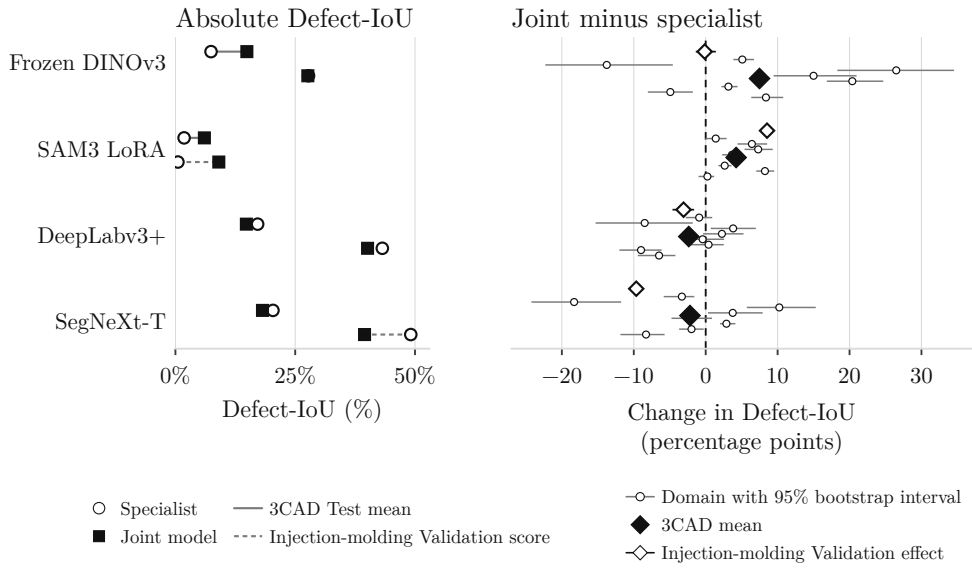


Fig. 3. Full-budget specialists versus the corresponding nine-domain joint model for each strategy. Left: absolute Defect-IoU values in percent for the 3CAD Test mean and the separate injection-molding Validation score. Right: joint-minus-specialist changes in percentage points with grouped 95% intervals; for example, a change from 10% to 15% is 5 percentage points. Table 1 expands the 3CAD codes.

The grouped 95% intervals qualified this pattern. Six of eight were entirely above zero for both DINOv3 and LoRA. The other two were below zero for DINOv3 and included zero for LoRA, while the CNN intervals were mixed. On injection-molding Validation, LoRA produced the largest joint gain at 8.5 percentage points, but the SegNeXt-T specialist retained the highest absolute score at 49.1%.

3.4 Pooled SAM3 Adaptation Depth

At the common 30-epoch endpoint, the held-out 3CAD Test mean did not improve monotonically with LoRA rank (Table 3). Rank 16 exceeded rank 4, while rank 32 fell below rank 16. Decoder-and-head fine-tuning remained below rank 16 on the 3CAD Test mean

but achieved the highest injection-molding Validation score. For this decoder arm, epoch 30 was also the best scheduled nine-domain Validation checkpoint within the budget. These single-seed results compare complete configurations, not an isolated effect of parameter count.

Table 3. Full-pool joint SAM3 results at the common 30-epoch comparison point (random seed 42). For decoder and heads, epoch 30 was also the best scheduled nine-domain Validation checkpoint within the budget. The 3CAD Test mean is unweighted across eight held-out domains; injection molding is reported separately on Validation. Trainable share is relative to total parameters. Both score columns report Defect-IoU.

Adaptation	Trainable	Share	3CAD Test mean	Injection Validation
LoRA rank 4	258,048	0.031%	22.5%	32.1%
LoRA rank 16	1,032,192	0.123%	25.3%	34.4%
LoRA rank 32	2,064,384	0.245%	23.2%	33.9%
Decoder and heads	14,924,081	1.776%	22.5%	39.1%

4 Discussion

Target labels changed the preferred configuration. At the eight-image budget, all four early-budget DeepLabv3+ runs produced nonzero Validation overlap. SegNeXt-T improved later and led the full-budget specialist 3CAD Test mean at 20.4%. Thus low-label and endpoint performance favored different configurations, without consistent gains across budgets and domains.

Joint effects depended on configuration: means rose for frozen feature reuse and capped SAM3 LoRA but fell for both CNNs. Restricted adaptation may constrain specialization and full tuning may permit interference, but differences in pretraining, architecture, capacity, optimization, and duration prevent causal attribution.

Relative gains did not imply high absolute overlap. Capped 429-image joint SAM3 LoRA gained 4.2 percentage points over its specialists but reached 6.0% Test Defect-IoU. This is distinct from the full-pool experiment.

Trainable capacity did not order full-pool SAM3 quality. At the common epoch-30 comparison point, which was also the decoder arm's best scheduled Validation checkpoint within the budget, LoRA rank 16 led the 3CAD Test mean, rank 32 did not improve it, and decoder-and-head tuning led only on injection-molding Validation. These single-seed runs used 20,468 images and 76,770 updates versus 429 and 1,620 in the capped comparison, confounding adaptation, data, and optimizer exposure.

Figure 4 shows four prespecified examples with misses and partial overlap. They illustrate error modes but not their prevalence or causes.

Among the five profiled core configurations, neither trainable share nor supported floating-point-operation (FLOP) lower bounds ordered quality (Fig. 5).

Evidence limits. Both comparisons use one seed; fixed-checkpoint bootstraps omit retraining variation, and 24 seed-43/44 Validation-only repeats are excluded. Joint runs change data and updates. Injection molding lacks independent Test evidence; all Test domains are 3CAD. Scope covers one ontology, pipeline, threshold, and zero-shot prompt–threshold pair.

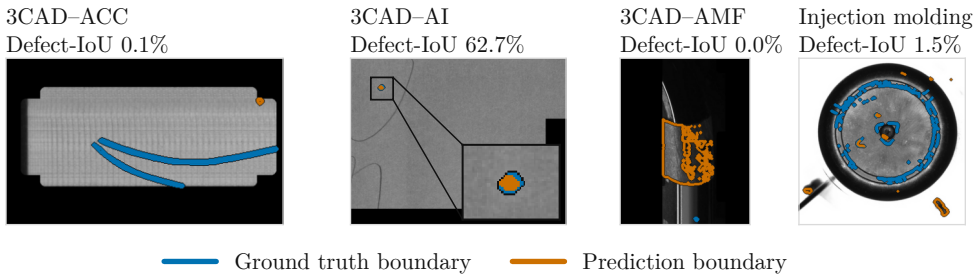


Fig. 4. Prespecified seed-42 SAM3 LoRA specialist Validation examples at epoch 30. Blue/orange contours mark ground truth/prediction; labels give per-image Defect-IoU. Images are nearest each domain's 33rd defect-area percentile; 3CAD–AI has a four-times inset.

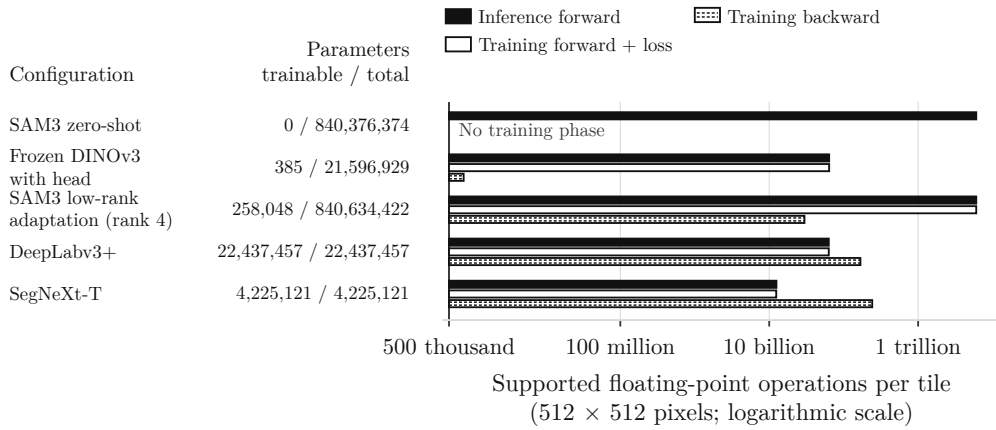


Fig. 5. Trainable/total parameters and supported-operation FLOP lower bounds per 512×512-pixel tile. Solid/open/dashed bars show inference, training forward plus loss, and backward on a shared logarithmic axis; lengths are not additive. Unsupported operations and system work are excluded. SAM3 uses 1008×1008-pixel inputs; pooled variants lack profiles.

5 Conclusion

Across nine domains, a shared protocol compared zero-shot use, nested-budget specialists, and joint training to test how pretrained vision moves from promise to practice. Zero-shot did not localize defects; specialists achieved the strongest endpoints. Capped joint training helped frozen and LoRA configurations but reduced CNN Test means, while extra full-pool SAM3 capacity gave no consistent gain. A fully fine-tuned CNN specialist led Injection Molding (49.1% Validation Defect-IoU; 45 labels). Practical choice therefore depends on labels, adaptation, scope, and domain.

Acknowledgements. We thank Katerina Snejdrlova, Philipp Krzikalla, and Christian Rytka (Institute of Polymer Engineering), and Melanie Völlmin (Institute for Data Science) for providing the injection-molding dataset. Innosuisse project 121.394 IP-ICT supported this work.

References

1. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: A composite dataset for multi-domain semantic segmentation. In: Conference on Computer Vision and Pattern Recognition. (2020) 2879–2888
2. Yang, E., et al.: 3CAD: A large-scale real-world 3C product dataset for unsupervised anomaly detection. arXiv preprint arXiv:2502.05761 (2025)
3. Carion, N., Gustafson, L., Hu, Y.T., et al.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
4. Siméoni, O., Vo, H.V., Seitzer, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations. (2022)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision. (2018)
7. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: SegNeXt: Rethinking convolutional attention design for semantic segmentation. In: Advances in Neural Information Processing Systems. (2022)
8. Wang, Z., Berman, M., Rannen-Triki, A., Torr, P.H.S., Tuia, D., Tuytelaars, T., Van Gool, L., Yu, J., Blaschko, M.B.: Revisiting evaluation metrics for semantic segmentation: Optimization and evaluation of fine-grained intersection over union. In: Advances in Neural Information Processing Systems. (2023)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations. (2019)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: International Conference on Computer Vision. (2017) 2980–2988
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. (2016) 565–571
12. Efron, B.: Bootstrap methods: Another look at the jackknife. The Annals of Statistics **7**(1) (1979) 1–26
13. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. (2017) 240–248
14. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary IoU: Improving object-centric image segmentation evaluation. In: Conference on Computer Vision and Pattern Recognition. (2021)
15. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool. BMC Medical Imaging **15** (2015) 29