DOVE: Domain-Robust Ordinal Evidence Learning with 2.5D Multiple Instance Learning for MRI-Based Liver Fibrosis Staging

Zhourui Liu[#], Feijian Xu[#], Zhao Yao[*], Zeyu Liu, Jiazheng Wang, Yue He, and Min Liu

Hunan University, Changsha, China
yaozhao24@hnu.edu.cn

Abstract. MRI-based liver fibrosis staging must generalize across scanners, respect the ordered disease continuum, and infer a patient diagnosis from a volume in which individual slices are only weakly informative. We present *DOVE*, a domain-robust ordinal volumetric evidence framework that addresses these three properties jointly. A liver-focused pipeline first localizes the organ with a parameter-efficient 2.5D segmenter. The staging network then applies MRI-specific domain randomization to adjacent-slice inputs, aggregates slice-level features with gated attention at the patient level, and introduces a Gaussian soft-label ordinal objective that distributes supervision across neighboring fibrosis stages before forming cumulative targets. On the CARE2026 cohort of 454 patients from three acquisition centers spanning two vendors and three scanner systems, random cross-validation gives an optimistic cirrhosis AUC of approximately 0.87, whereas center-held-out development exposes a substantially lower baseline of 0.727. The complete five-model DOVE ensemble reaches a macro center-held-out AUC of 0.854 for cirrhosis and 0.730 for substantial fibrosis; the lowest baseline center improves from 0.621 to 0.841. A small annotated segmentation test set yields a Dice score of 0.933. These results support a task-aligned design in which acquisition variation, ordinal ambiguity, and volumetric evidence are modeled explicitly rather than absorbed by a conventional slice-level classifier.

Keywords: Liver fibrosis staging · Domain robustness · Soft-label ordinal learning · Multiple-instance learning · 2.5D MRI

1 Introduction

Liver fibrosis is a progressive response to chronic injury and may culminate in cirrhosis, portal hypertension, and liver failure. Although biopsy remains a reference standard, it is invasive and subject to sampling variability; MRI offers a non-invasive view of the entire organ. The CARE-Liver benchmark formulates

[#] Zhourui Liu and Feijian Xu contributed equally.
[*] Corresponding author.

MRI-based staging as four ordered categories, S1–S4, and evaluates cirrhosis (S4 vs. S1–S3) and substantial fibrosis (S2–S4 vs. S1) [1,2]. This setting presents three coupled challenges. First, scanner vendor, field strength, reconstruction, and intensity statistics induce substantial acquisition shift, which random folds can conceal [5,6,7,8,9]. Second, S1–S4 discretize a continuous disease process, so treating them as unrelated classes ignores neighboring-stage ambiguity and hard targets overstate boundary certainty [10,11]. Third, fibrosis is diagnosed at the patient level from a three-dimensional examination; copying that label to every axial slice creates noisy supervision because slices contain unequal diagnostic evidence [14,15].

To address these challenges, we introduce DOVE, a *D*omain-robust *O*rdinal *V*olumetric *E*vidence learning framework. A parameter-efficient segmenter first localizes the liver on the reliably annotated enhanced phase. For cross-domain robustness, training randomizes MRI-specific appearance factors, including low-frequency intensity inhomogeneity, contrast, and sharpness. For volumetric reasoning, each axial location and its two neighbors form a 2.5D instance, and gated attention aggregates all instances into a patient representation without assigning slice-level fibrosis labels. For ordinal prediction, we compare hard cumulative targets with a Gaussian-soft alternative over S1–S4 that preserves neighboring-stage similarity and boundary ambiguity; both objectives also provide complementary ensemble members. Unlike previous CARE systems centered on multimodal fusion or registration-assisted semi-supervision [3,4], DOVE couples one quality-controlled liver ROI with explicit acquisition, ordering, and patient-level inductive biases.

Figure 2 illustrates that center-dependent differences remain visible after identical ROI extraction and intensity preprocessing, in both the gray-level and frequency domains. These differences motivate explicit appearance randomization rather than assuming that preprocessing alone removes scanner effects.

Our contributions are threefold:

- a task-aligned staging formulation that couples scanner-aware domain randomization, 2.5D patient-level MIL, and Gaussian soft-label cumulative ordinal learning;
- a center-held-out analysis showing that random cross-validation materially overestimates cross-center performance and identifying the weakest acquisition domain; and
- an end-to-end LiSeg–LiFS pipeline with explicit ROI quality checks, center-crop fallback, and probability outputs for both challenge tasks.

2 Method

2.1 Problem setup and overview

For patient i, the input is an MRI volume X_i and the stage label is $y_i \in \{0, 1, 2, 3\}$, corresponding to S1–S4. The system must output a probability vector $\mathbf{p}_i \in \mathbb{R}^4$. The challenge derives Task 1 from $p_{i,3} = P(\text{S4})$ and Task 2 from

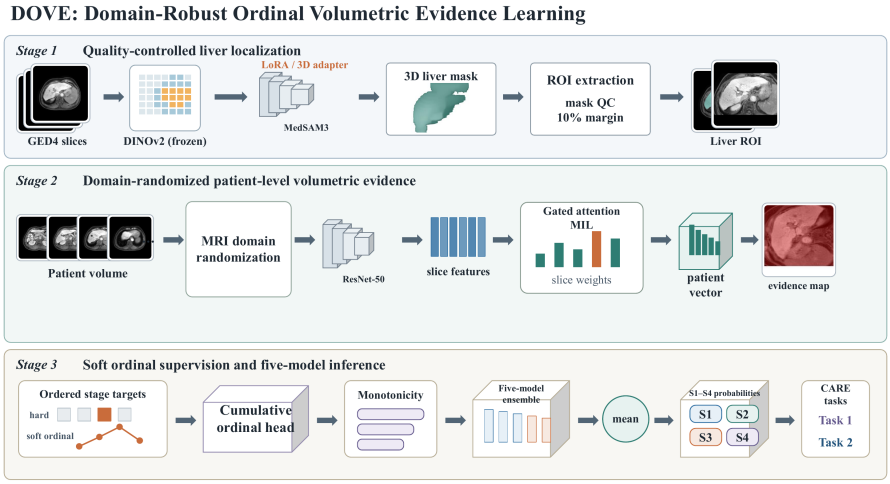

Fig. 1. Overview of DOVE. A quality-controlled MedSAM3/DINOv2 liver ROI feeds domain-randomized slice encoding and gated patient-level MIL; cumulative ordinal prediction, monotonic projection, and a five-model ensemble produce S1–S4 probabilities and the two CARE task scores. Classification uses patient-level labels only.

$1 - p_{i,0} = P(\text{S2:S4})$. Figure 1 summarizes the pipeline. We use the GED4 phase because it provides the most reliable liver localization in the available annotations and avoids error propagation from poorly aligned or missing modalities.

2.2 Liver localization

Only about 30 training cases contain a manual GED4 liver mask. We adapt a MedSAM3 image model [20,19,24,25] using low-rank updates (LoRA) [21] and three-slice adapters inspired by MA-SAM [22]. For the center slice, its preceding and following slices provide through-plane context, with boundary slices replicated. A frozen DINOv2 ViT-B/14 [18] additionally supplies a semantic prior: the feature channels of each patch token are averaged to form a 14×14 activation map, which is resized and concatenated as a fourth input channel. The added patch-embedding channel is initialized to zero. DINOv2 is evaluated during both training and inference; maps are cached within each inference pass. The 3D output is reduced to its largest connected component, hole-filled, and median-filtered.

For staging, a valid predicted mask must contain at least 50 voxels, occupy no more than 60% of the volume, and yield a non-empty bounding box. Its box is expanded by 10% in each dimension. If these checks fail, a central crop covering 60% of each spatial dimension is used. This fallback guarantees a defined classifier input without interpreting a failed mask as anatomy.

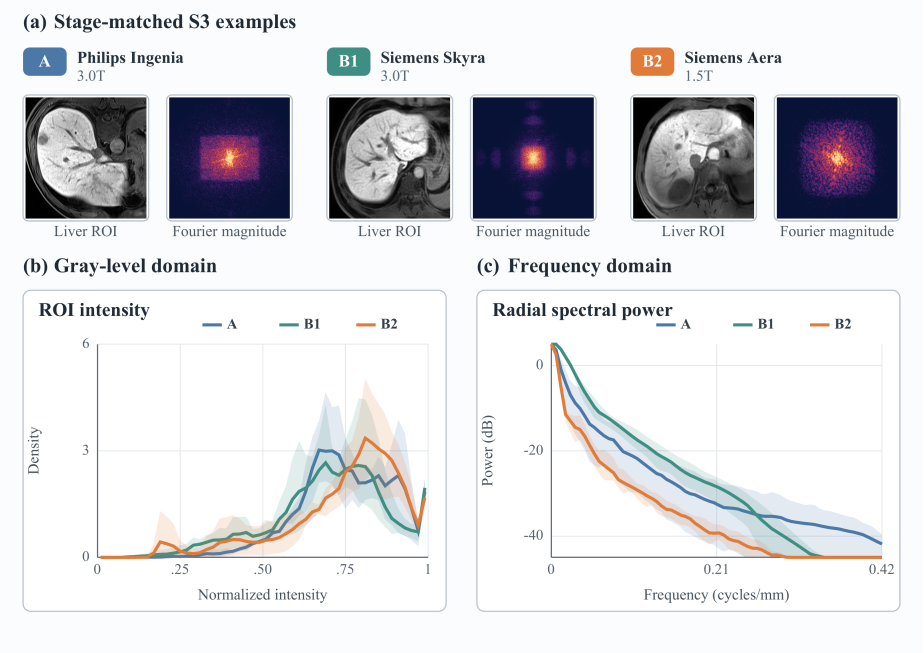

Fig. 2. Representative acquisition-domain differences in GED4. Each center contributes one stage-matched S2, S3, and S4 case (nine patients total). Panel (a) shows the S3 liver ROI and its log-magnitude Fourier spectrum; panels (b) and (c) show the center-wise mean and case-wise range over all three matched cases. Intensities follow the classifier preprocessing (1st–99th percentile clipping and scaling to $[0, 1]$); spectra use Hann-windowed ROI crops and physical in-plane spacing. A, B1, and B2 correspond to Philips Ingenia 3.0T, Siemens Skyra 3.0T, and Siemens Aera 1.5T, respectively. This small representative comparison motivates appearance randomization and is not a full-cohort population estimate.

2.3 Domain-randomized 2.5D representation

For each ROI location z, we follow the established 2.5D principle of stacking contextual slices as channels [16]. Specifically, we stack slices $[X_{z-1}, X_z, X_{z+1}]$ as channels, replicating the nearest valid slice at volume boundaries. Each channel is clipped at its 1st and 99th percentiles, normalized to $[0, 1]$, and resized to 224×224. The first and last 5% of ROI slices are discarded because they are dominated by crop margins.

Training applies a shared spatial transform to the three channels: horizontal and vertical flips, rotation, translation, and cutout, followed by intensity jitter and Gaussian noise. Three MRI-specific transformations then randomize the acquisition appearance: (i) a smooth multiplicative bias field generated from a low-resolution random grid, (ii) gamma adjustment, and (iii) Gaussian blur. In the selected setting, their probabilities are 0.7, 0.6, and 0.3, respectively;

the bias amplitude is 0.3, gamma is sampled from $[0.7, 1.5]$, and blur σ from $[0.4, 1.2]$. These operators alter low-frequency intensity non-uniformity, contrast, and sharpness without requiring target-domain images.

2.4 Patient-level volumetric evidence

A ResNet-50 encoder [17] maps each 2.5D input to a feature $\mathbf{h}_s \in \mathbb{R}^d$. During training, a patient bag contains at most 32 randomly sampled slices; all eligible slices are retained at validation and inference. Following gated attention MIL [14], the importance of slice s is

$$a_s = \frac{\exp\{\mathbf{w}^\top[\tanh(\mathbf{V}\mathbf{h}_s) \odot \sigma(\mathbf{U}\mathbf{h}_s)]\}}{\sum_j \exp\{\mathbf{w}^\top[\tanh(\mathbf{V}\mathbf{h}_j) \odot \sigma(\mathbf{U}\mathbf{h}_j)]\}}, \qquad \mathbf{h}_i = \sum_s a_s \mathbf{h}_s, \tag{1}$$

where $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{d_a \times d}$ are learnable projections for the content and gate branches, respectively, $\mathbf{w} \in \mathbb{R}^{d_a}$ maps their element-wise interaction to a scalar attention logit, and $\odot$ denotes element-wise multiplication. We use $d_a = 128$. The softmax normalization makes $a_s \geq 0$ and $\sum_s a_s = 1$, so $\mathbf{h}_i$ is a patient-level weighted combination of its slice instances. Only this patient representation is supervised. This avoids assigning the fibrosis label to every slice and provides an interpretable distribution of evidence over the volume.

2.5 Soft ordinal label distribution and loss

A hard stage label assumes an exact boundary between adjacent fibrosis stages, although S1–S4 are discrete observations of a continuous pathological process. We therefore evaluate Gaussian soft labels as an alternative ordinal supervision target. For patient i, a discrete Gaussian centered at the annotated stage y_i assigns non-zero probability to its ordered neighbors [13]:

$$q_{i,c} = \frac{\exp[-(c - y_i)^2/(2\sigma_y^2)]}{\sum_{r=0}^{3} \exp[-(r - y_i)^2/(2\sigma_y^2)]}, \qquad c \in \{0, 1, 2, 3\}, \quad \sigma_y = 0.7. \tag{2}$$

The ordinal head outputs logits $\mathbf{z}_i = (z_{i,0}, z_{i,1}, z_{i,2})$ for the three events $y_i > k$. We convert the stage distribution into cumulative soft targets

$$t_{i,k}^{\text{soft}} = P(y_i > k) = \sum_{c=k+1}^{3} q_{i,c}, \qquad k \in \{0, 1, 2\}. \tag{3}$$

Training loss. Let n_c denote the number of training patients at stage c, and define the normalized inverse-frequency weight as $\alpha_c = (1/n_c)/[\frac{1}{4}\sum_{r=0}^{3}(1/n_r)]$. The classification network is optimized by the following single patient-level loss:

$$\text{Loss}_{\text{soft}} = -\frac{1}{B}\sum_{i=1}^{B} \alpha_{y_i} \frac{1}{3}\sum_{k=0}^{2} \left[t_{i,k}^{\text{soft}} \log \sigma(z_{i,k}) + (1 - t_{i,k}^{\text{soft}}) \log(1 - \sigma(z_{i,k}))\right]. \tag{4}$$

This is the class-weighted mean of three soft binary cross-entropies for every patient bag. No additional slice-level classification or attention loss is used; its gradients train both the ordinal head and attention-MIL encoder end to end. As $\sigma_y \to 0$, $q_{i,c}$ becomes one-hot and $t_{i,k}^{\text{soft}}$ reduces to $\mathbb{1}[y_i > k]$. The hard-target models therefore use the same loss with $t_{i,k}^{\text{soft}}$ replaced by $\mathbb{1}[y_i > k]$. The proposed soft loss propagates supervision to neighboring stages while preserving their order through cumulative thresholds. This independent cumulative-threshold objective differs from the conditional formulation of CORN [12].

At inference, $s_{i,k} = \sigma(z_{i,k})$ is projected to a non-increasing sequence by $\tilde{s}_{i,k} = \min_{j \le k} s_{i,j}$. The stage probabilities are

$$\mathbf{p}_i = [1 - \tilde{s}_{i,0},\ \tilde{s}_{i,0} - \tilde{s}_{i,1},\ \tilde{s}_{i,1} - \tilde{s}_{i,2},\ \tilde{s}_{i,2}], \tag{5}$$

which are non-negative and sum to one. The final prediction averages three hard-target models and two soft-label models trained with different seeds; the latter use $\text{Loss}_{\text{soft}}$ to retain boundary-aware supervision.

3 Experiments and Results

3.1 Dataset, protocols, and implementation

The training cohort contains 454 patients from centers A, B1, and B2, with stage counts 99/75/48/232 for S1/S2/S3/S4. These centers use Philips Ingenia 3.0T, Siemens Skyra 3.0T, and Siemens Aera 1.5T systems, respectively; B1 and B2 therefore belong to the same scanner vendor, and B2 is the only 1.5T source [1]. The available validation cohort contains 60 additional patients. For classification experiments, 453 patients have a usable mask-derived GED4 ROI: 157 from A, 223 from B1, and 73 from B2. We report the missing case explicitly rather than silently changing the cohort denominator.

We compare patient-stratified random five-fold cross-validation with leave-one-center-out (LOCO) development. Task 1 AUC uses $P(\text{S4})$; Task 2 AUC uses $1 - P(\text{S1})$. The current LOCO checkpoints were selected by mean Task 1/Task 2 AUC on the held-out center; accordingly, these experiments quantify a *center-held-out development gap*, not a strictly target-agnostic generalization estimate. Source-only model selection is required for that stronger claim (Sec. 3.5).

The classifier uses an ImageNet-pretrained ResNet-50 and is optimized with AdamW [23] for at most 25 epochs using a learning rate of 3×10^{-4}, weight decay 5×10^{-4}, cosine scheduling, patient-level inverse-frequency class weights, batch size 8 bags, dropout 0.5, and early stopping patience 5. The segmentation models use the same 24/3/3 annotated train/validation/test split, batch size 1 with gradient accumulation 8, and learning rate 10^{-5}. Because its test split contains only three cases (132 slices), the segmentation numbers are preliminary rather than a population estimate.

Table 1. GED4 liver segmentation on the fixed three-case annotated test split. AP averages IoU thresholds 0.50:0.95.

Model	Dice	AP	AP_{50}	AP_{75}
MedSAM3 + LoRA (2D)	0.902	0.776	0.962	0.881
+ 3D adapter	0.928	0.785	**0.980**	0.874
+ DINOv2 prior	**0.933**	**0.816**	0.978	**0.881**

Table 2. Component results for center-held-out cirrhosis AUC. D: domain randomization; O: Gaussian soft ordinal supervision; V: 2.5D patient-level attention-MIL. The two single-MIL rows form a matched seed-0 control in which only the ordinal target changes; macro values use unrounded center scores.

Configuration	D	O	V	A	B1	B2	Macro
Slice classifier, mask ROI				0.744	0.814	0.621	0.727
Domain-randomized slice classifier	✓			0.841	0.777	0.717	0.778
Hard/soft slice ensemble	✓	✓		0.821	0.822	0.725	0.789
Single hard attention-MIL	✓		✓	0.794	0.815	0.728	0.779
Single soft attention-MIL	✓	✓	✓	0.825	0.814	0.821	0.820
DOVE, hard×3/soft×2	✓	✓	✓	**0.857**	**0.865**	**0.841**	**0.854**

3.2 Liver segmentation

Table 1 shows the controlled segmentation comparison. Adding through-plane context improves Dice from 0.902 to 0.928, and the DINOv2 prior raises it to 0.933. The same ordering is observed for COCO-style mAP. Given the three-case test set, we interpret this table as architecture selection evidence, not as a claim of broadly established segmentation generalization.

3.3 Cross-center staging

Random five-fold evaluation yields approximately 0.87 Task 1 AUC, but the slice-level baseline falls to 0.727 when an entire center is held out. This 0.14 gap demonstrates that random splitting does not measure the acquisition shift that motivates the task. Table 2 traces the development pipeline through a component-wise checkmark ablation, where D, O, and V denote domain randomization, Gaussian soft ordinal supervision, and 2.5D patient-level attention-MIL, respectively. MRI domain randomization provides the largest first improvement. In a matched seed-0 MIL comparison, changing only the ordinal target from hard to soft raises macro Task 1 AUC from 0.779 to 0.820. The complete multi-seed ensemble reaches 0.854, and its weakest baseline fold, B2, rises from 0.621 to 0.841.

For Task 2, the final ensemble obtains AUCs of 0.777, 0.796, and 0.618 on A, B1, and B2, respectively (macro 0.730). The B2 point estimate should not be over-interpreted because that fold contains only five S1 negatives.

Table 3. Matched single-model attention-MIL control (seed 0). Entries are AUCs macro-averaged across centers; Mean averages the two tasks. All settings except the ordinal target are identical.

Ordinal target	Task 1	Task 2	Mean
Hard cumulative	0.779	**0.744**	**0.761**
Gaussian soft cumulative	**0.820**	0.697	0.759
Soft minus hard	+0.041	−0.047	−0.003

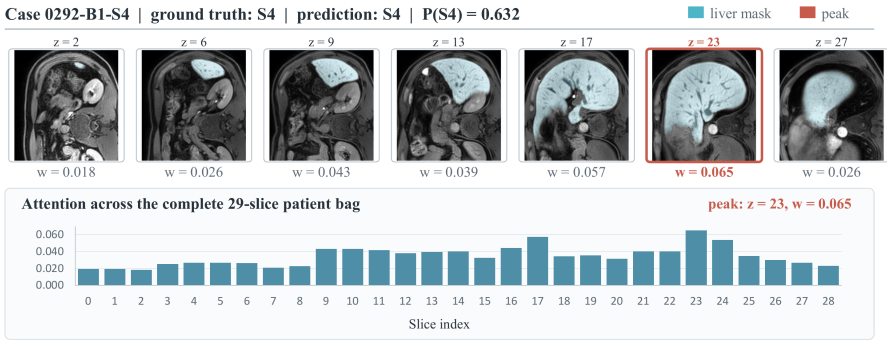

Fig. 3. Patient-level attention for a correctly classified S4 GED4 case. Seven slices sample the volume and use a translucent liver-mask overlay for anatomical context; the bar chart reports normalized gated-attention weights for all 29 instances. The maximum occurs at $z = 23$ ($w = 0.065$). The mask is used for ROI localization, whereas the attention weights are learned only from patient-level staging labels. The visualization is qualitative and is not a causal explanation.

3.4 Ablation results

Table 2 reports the component results supported by available checkpoints and the new matched control. Domain randomization improves Task 1 macro AUC from 0.727 to 0.778. With architecture, data, augmentation, optimization, early stopping, and seed fixed, Gaussian soft targets improve Task 1 from 0.779 to 0.820, driven mainly by B2. However, Table 3 shows the opposite direction on Task 2: soft targets reduce its macro AUC from 0.744 to 0.697, leaving the mean across tasks nearly unchanged (0.761 versus 0.759). Thus the controlled evidence supports a task-dependent trade-off and ensemble diversity, not a uniform soft-label benefit; the full ensemble gain cannot be attributed to the ordinal objective alone.

The attention visualization in Fig. 3 provides a qualitative sanity check. The complete weight profile is non-uniform: the representative patient has a clear maximum at slice $z = 23$, while peripheral and several intermediate slices receive substantially less weight. Attention is not a causal explanation, but the pattern confirms that the patient representation is not formed by uniform averaging.

3.5 Limitations and validity boundary

Three limitations define the current evidence boundary. First, the center-held-out checkpoints use the held-out center for early stopping; a strict domain-generalization evaluation must select epochs and hyperparameters only from source centers, then evaluate the target center once. The present numbers are therefore development estimates. Second, the fixed segmenter is trained with annotated cases from all available centers. The LOCO experiment isolates the staging classifier, not the entire segmentation-to-staging system. End-to-end LOCO would require a source-only segmenter for each outer fold. Third, segmentation is assessed on only three cases, and Task 2 on B2 has only five negative patients. Larger independently acquired cohorts are required for precise estimates.

The current implementation also imposes a practical boundary: DINOv2 is executed at inference, so its computational cost is part of the segmentation stage. The system is packaged with local model code and weights, but an offline container claim should be made only after an end-to-end no-network run and output-equivalence check.

4 Conclusion

DOVE organizes MRI-based fibrosis staging around the actual structure of the problem: scanner appearance is randomized, neighboring slices provide compact through-plane context, patient-level attention replaces noisy slice labels, and cumulative targets respect the ordered stage continuum. The matched control shows that Gaussian soft targets improve cirrhosis AUC but reduce substantial-fibrosis AUC, so they provide a task-dependent trade-off rather than a uniform gain. On the available CARE2026 development cohort, the complete ensemble raises macro center-held-out cirrhosis AUC from 0.727 to 0.854. The study also exposes which claims require stricter validation: source-only model selection, end-to-end center holdout, and a larger segmentation test. These steps are necessary for a defensible unseen-domain conclusion.

Acknowledgments

We thank the CARE 2026 organizers for providing the anonymized CARE-Liver benchmark used in this study. All analyzed patient data were supplied through the CARE-Liver track; no external patient cohort was used.

References

1. Liu, Y., Shi, N., Zhang, Z., et al.: How Far Has AI Come in Liver Fibrosis Staging? A Large-Scale Real-World Dataset and Benchmark. Medical Image Analysis (2026). arXiv:2605.25595
2. Liu, Y., et al.: Liver Fibrosis Quantification and Analysis: The LiQA Dataset and Baseline Method. arXiv:2512.07651 (2025)

3. Zhou, Y., Yuan, K., Wei, Y., Chen, J.: Multi-modal Liver Segmentation and Fibrosis Staging Using Real-world MRI Images. In: MICCAI CARE Challenge (2025). arXiv:2509.26061
4. Wang, B., Li, R., Chen, C., Chen, X.: Semi-supervised Liver Segmentation and Patch-based Fibrosis Staging with Registration-aided Multi-parametric MRI. In: MICCAI CARE Challenge (2025). arXiv:2602.09686
5. 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: IROS, pp. 23–30 (2017)
6. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain Generalization: A Survey. IEEE TPAMI 45(4), 4396–4415 (2023)
7. Li, H., Wang, Y., Wan, R., Wang, S., Li, T.Q., Kot, A.: Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization. In: NeurIPS, vol. 33, pp. 3118–3129 (2020)
8. Gulrajani, I., Lopez-Paz, D.: In Search of Lost Domain Generalization. In: ICLR (2021)
9. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain Generalization with MixStyle. In: ICLR (2021)
10. Diaz, R., Marathe, A.: Soft Labels for Ordinal Regression. In: CVPR, pp. 4738–4747 (2019)
11. Cao, W., Mirjalili, V., Raschka, S.: Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation. Pattern Recognition Letters 140, 325–331 (2020)
12. Shi, X., Cao, W., Raschka, S.: Deep Neural Networks for Rank-Consistent Ordinal Regression Based on Conditional Probabilities. Pattern Analysis and Applications 26, 941–955 (2023)
13. Geng, X.: Label Distribution Learning. IEEE TKDE 28(7), 1734–1748 (2016)
14. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based Deep Multiple Instance Learning. In: ICML, pp. 2127–2136 (2018)
15. Campanella, G., Hanna, M.G., Geneslaw, L., et al.: Clinical-grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. Nature Medicine 25, 1301–1309 (2019)
16. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations. In: MICCAI, pp. 520–527 (2014)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR, pp. 770–778 (2016)
18. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2024)
19. Carion, N., Gustafson, L., Hu, Y.T., et al.: SAM 3: Segment Anything with Concepts. In: ICLR (2026). arXiv:2511.16719
20. Liu, A., Xue, R., Cao, X.R., et al.: MedSAM3: Delving into Segment Anything with Medical Concepts. arXiv:2511.19046 (2025)
21. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022)
22. Chen, C., Miao, J., Wu, D., et al.: MA-SAM: Modality-Agnostic SAM Adaptation for 3D Medical Image Segmentation. Medical Image Analysis 98, 103310 (2024)
23. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)

24. Kirillov, A., Mintun, E., Ravi, N., et al.: Segment Anything. In: ICCV, pp. 4015–4026 (2023)
25. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images. Nature Communications 15, 654 (2024)