Contrastive Alignment and CT-Guided Distillation for Ultrasound-Only Intracerebral Hemorrhage Screening

Jesus David Gonzalez[1,2], Renato Simonetti[1], Pere Canals[1,3], Marc Ribo[1,2], and David Rodriguez-Luna[1,2]

[1] Stroke Unit, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron University Hospital, Barcelona, Spain
[2] Facultat de Medicina, Universitat Autònoma de Barcelona, Barcelona, Spain
[3] Stanford University, Stanford, CA, USA
{jesus.gonzalez.riveros,renato.simonetti,pere.canals, marc.ribo}@vhir.org

Abstract. Rapid differentiation between intracerebral hemorrhage (ICH) and ischemic stroke is important in hyper-acute care, yet non-contrast CT (NCCT), the diagnostic gold standard, is unavailable during prehospital transport. Recent trials demonstrate that ultra-early interventions (blood pressure lowering and hemostatic therapy) improve ICH outcomes when initiated within minutes, creating an urgent need for prehospital stroke differentiation. We propose a cross-modal framework that uses paired NCCT–ultrasound data during training to improve ultrasound-only ICH screening at deployment. We collected 100 transtemporal transcranial B-mode ultrasound examinations from acute stroke patients, including 91 with paired NCCT, using a standardized 5-second sweep protocol. Our model combines an ultrasound pathway built on a pretrained ultrasound foundation model (USFM) with attention-based multiple-instance learning (MIL), a 3D ResNet NCCT encoder pretrained on the RSNA ICH dataset, and cross-modal objectives: contrastive feature alignment and CT-guided knowledge distillation. At inference, only ultrasound is required; multi-sweep MAX-pool aggregation yields a single patient-level prediction. Evaluated with patient-disjoint 5-fold cross-validation, the framework achieves AUROC 0.92±0.07, sensitivity 0.90±0.13, and specificity 0.86±0.14 at the Youden-optimal operating point, outperforming CT-only (AUROC 0.86±0.07) and ultrasound-only (AUROC 0.85±0.14) baselines. Ablation studies show that the two cross-modal losses provide complementary gains, with distillation acting as both a knowledge transfer mechanism and an implicit regularizer. Subgroup analyses show no significant performance differences across operators of varying expertise (p=0.59) or acoustic window quality (p=0.92), supporting the robustness and deployment feasibility of ultrasound-based ICH screening in the prehospital setting.

Keywords: Cross-modal learning · Knowledge distillation · Intracerebral hemorrhage · Transcranial ultrasound · Prehospital screening

1 Introduction

Intracerebral hemorrhage (ICH) accounts for 28.8% of all strokes and carries the highest acute mortality, reaching 45.6% of all stroke-related deaths [1]. Hematoma expansion occurs in approximately one-third of patients within the first hours and is the primary determinant of early neurological deterioration [2]. Ultra-early intervention can alter this trajectory. The INTERACT4 trial and subsequent pooled analyses demonstrated that intensive ambulance-delivered treatment significantly improved functional outcomes in hemorrhagic stroke patients, with benefit strongly dependent on early initiation [3, 4]. Likewise, the FASTEST trial [5] showed that recombinant Factor VIIa (rFVIIa) administered within 2 hours of ICH onset reduced hematoma growth, in patients treated within 90 minutes and those with CT angiography spot sign [6]. This evidence establishes that delays in ICH identification substantially affect treatment efficacy, yet stroke type differentiation in the prehospital setting currently requires non-contrast CT (NCCT) [7], equipment confined to hospitals or costly mobile stroke units.

Transcranial ultrasound (US) offers a viable path to ultra-early stroke differentiation. The technology is portable, radiation-free, and increasingly deployed in ambulances. B-mode transcranial imaging through temporal windows can visualize intracranial structures in real-time, but spatial resolution is inherently limited compared to CT, and automated ultrasound-only approaches for hemorrhage detection achieve modest accuracy [8]. The core challenge is that ultrasound and CT encode hemorrhage through fundamentally different physical mechanisms (acoustic impedance vs. X-ray attenuation) and share no direct spatial correspondence.

Cross-modal learning addresses this by using paired CT–ultrasound data during training only, so that at deployment only the portable device is needed. This paradigm has been applied to MRI-to-MRI distillation for brain tumor segmentation [10], cross-dataset organ segmentation [11], and CT-to-US transfer for kidney segmentation [12], but never to brain pathology classification nor across the extreme gap between 3D volumetric CT and 2D temporal ultrasound sweeps.

Our framework addresses these challenges through three design choices: (i) a CT teacher pretrained on the large-scale RSNA ICH dataset [13] ($\sim$19,500 studies) to provide robust hemorrhage features despite a small paired cohort; (ii) a multi-loss objective combining classification, contrastive alignment (InfoNCE [14]), and knowledge distillation (KL divergence [15]); and (iii) multi-sweep MAX-pool inference exploiting the temporal redundancy of the acquisition protocol. Because transcranial ultrasound for ICH detection is not collected routinely and no public paired datasets exist, data was obtained through a dedicated prospective protocol [9], motivating our focus on data-efficient cross-modal transfer.

Our contributions are:

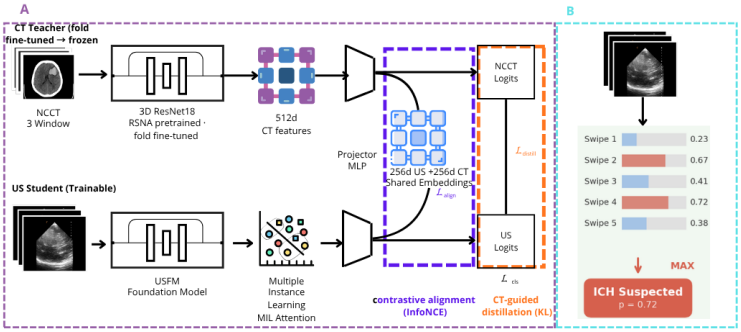

Fig. 1. Framework overview. (A) Training. The CT teacher (3D ResNet-18) is pretrained on RSNA and subsequently fine-tuned on the training patients of each cross-validation fold; its weights are frozen only for the cross-modal stage depicted here. The USFM backbone is likewise frozen; MIL attention, projectors and classifier heads are trainable. Both branches are jointly optimised through classification ($\mathcal{L}_{\text{cls}}$), contrastive alignment ($\mathcal{L}_{\text{align}}$) and CT-guided distillation ($\mathcal{L}_{\text{distill}}$, teacher→student, stop-gradient) in a shared 256-d embedding space. **(B)** Inference: ultrasound only; per-sweep probabilities are MAX-pooled to a patient-level prediction.

1. A cross-modal framework distilling NCCT knowledge into a transcranial ultrasound encoder for ICH detection, enabling screening with portable equipment where CT is unavailable.
2. A multi-loss training objective combining classification, contrastive alignment, and knowledge distillation.
3. Validation on 91 paired examinations with patient-disjoint 5-fold cross-validation, achieving AUROC 0.92±0.07 and sensitivity 0.90±0.13, substantially improving over the ultrasound-only baseline, with robustness across operators and acoustic window quality.

2 Methods

2.1 CT Teacher Encoder

The CT branch (Top of Fig. **??**, block A) extracts hemorrhage-discriminative features from multi-window NCCT volumes. Following clinical reading practice, each volume is rendered at three Hounsfield unit windows (brain: W80/L40, subdural: W200/L80, bone: W2800/L600) and stacked as a 3-channel input. The encoder g_ϕ is a 3D ResNet-18 [16] with 2D kernels inflated to 3D for volumetric processing.

To compensate for the small paired cohort, we pretrain g_ϕ on the RSNA Intracranial Hemorrhage Detection dataset [13] (∼19,500 CT studies). Without this pretraining, the CT encoder achieves only AUROC 0.61±0.07 on our institutional data, insufficient for effective knowledge transfer. With RSNA pretraining, the CT teacher reaches AUROC 0.86±0.07 on our cohort (Table 1).

While comparable to the ultrasound-only baseline, it encodes complementary hemorrhage information through a different physical modality, enabling cross-modal knowledge transfer. Within each fold, g_ϕ is first pretrained on RSNA, then fine-tuned on the NCCT volumes of that fold's training patients only; test patients are never seen at any stage. This fold-wise fine-tuned encoder is the model reported in the CT row of Table 1. Its weights are frozen only for the subsequent cross-modal stage, where the 512-d feature vector is shared across all sweeps from the same patient.

2.2 US Student Encoder

The ultrasound branch processes individual sweeps of K=20 frames acquired using a standardized 5-second sweep protocol. Each frame is encoded by a pretrained Ultrasound Foundation Model (USFM) [17] backbone h_ψ. A gated attention mechanism following the Multiple-Instance Learning (MIL) framework [18] aggregates frame-level features into a single sweep-level representation $z^{\text{us}}_{i,s} \in R^{512}$:

$$z^{\text{us}}_{i,s} = \sum_{k=1}^{K} \alpha_k \cdot h_\psi(x^{\text{us}}_{i,s,k}), \quad \alpha_k = \frac{\exp(w^\top \tanh(V h_k))}{\sum_j \exp(w^\top \tanh(V h_j))} \tag{1}$$

where V, w are learnable attention parameters. The MIL formulation treats each sweep as a "bag" of frames, allowing the model to learn which frames carry the strongest hemorrhage signal without requiring frame-level annotations. The USFM backbone is frozen; only the MIL attention module, projection layers, and classifier heads are trained.

2.3 Cross-Modal Training Objective

Both modality features are projected into a shared 256-d space via modality-specific MLP projectors: $e^{\text{ct}} = \pi_{\text{ct}}(z^{\text{ct}}) \in R^{256}$ and $e^{\text{us}} = \pi_{\text{us}}(z^{\text{us}}) \in R^{256}$. Separate classification heads produce logits from these shared embeddings. The total loss combines three objectives:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} \tag{2}$$

Classification loss ($\mathcal{L}_{\text{cls}}$) applies cross-entropy to both modality logits: $\mathcal{L}_{\text{cls}} = \text{CE}(\hat{y}^{\text{us}}, y) + \text{CE}(\hat{y}^{\text{ct}}, y)$, ensuring the shared space is discriminative for both modalities.
Contrastive alignment ($\mathcal{L}_{\text{align}}$) uses symmetric InfoNCE [14] to encourage paired CT–US embeddings to be proximal while separating unpaired embeddings, with temperature τ=0.07.
Knowledge distillation ($\mathcal{L}_{\text{distill}}$) minimizes the KL divergence [15] between softened CT teacher and US student logits:

$$\mathcal{L}_{\text{distill}} = T^2 \cdot \text{KL}(\sigma(\hat{y}^{ct}/T) \parallel \sigma(\hat{y}^{us}/T)) \tag{3}$$

with distillation temperature T=4.

During training, CT features are randomly replaced with zeros with probability p_{drop}=0.3 (modality dropout), ensuring the US branch can function independently at inference.

2.4 Multi-Sweep Inference

At inference, only the US branch is retained. Each of the S_i sweeps from patient i yields a probability $p_{i,s}$. The patient-level prediction is obtained via MAX-pool: $\hat{p}_i = \max_s p_{i,s}$. This reflects the screening logic: if any temporal window captures a hemorrhage signal, the patient should be flagged for CT confirmation.

3 Experiments

3.1 Dataset and Setup

A sample of 100 transtemporal transcranial B-mode ultrasound examinations was collected from patients presenting with acute stroke symptoms at an anonymized institution, of whom 91 had paired same-day NCCT (30 ICH, 61 ischemic/non-ICH). Transcranial ultrasound acquisitions followed a standardized protocol: 5-second sweeps, acquired once patients were stable shortly after super-acute management to minimize deviation from the intended prehospital window in future practice. Because transcranial US is not routine in this setting, examinations were purpose-acquired under an approved research protocol.

Each patient underwent multiple sweeps (median 3, range 1–7; ∼350 sweep-level samples), each comprising 20 sampled frames. We performed 5-fold patient-level cross-validation with stratified splits (64/16/20% train/val/test), ensuring strict patient disjointness; metrics are reported on held-out test folds as mean±std across folds.

Both backbones were frozen; only MIL attention, projectors, and classifiers were trained using AdamW (lr 10^{-4}, weight decay 10^{-4}), cosine annealing, batch size 8, and early stopping (patience 20), with λ_{align}=0.1 and λ_{distill}=0.5.

3.2 Main Results

Table 1 summarizes performance across all methods. The US-only baseline uses the same USFM backbone with MIL attention, trained with cross-entropy only and without CT supervision. Our framework outperforms both the US-only baseline and the CT teacher across all metrics. At the Youden-optimal operating point, sensitivity reaches 0.90±0.13 with specificity 0.86±0.14, detecting nine of ten hemorrhages while maintaining clinically acceptable false-positive rates suitable for a screening tool intended to triage patients toward confirmatory CT.

The student surpassing the teacher is consistent with distillation literature [19, 20]: soft targets from a moderate teacher, combined with strong pretrained student initialization (USFM), provide complementary signals that exceed either

Table 1. Main results (5-fold patient-level CV, mean±std). Youden metrics use the per-fold optimal threshold. Δ: improvement over US-only baseline.

Method	Inference	AUROC	@ Youden threshold	
			Sensitivity	Specificity
US-only (USFM+MIL)	US only	0.85±0.14	0.52±0.07	0.87±0.12
CT encoder (RSNA-pretrained)	CT only	0.86±0.07	0.83±0.09	**0.95±0.04**
Ours (full)	**US only**	**0.92±0.07**	**0.90±0.13**	0.86±0.14
Δ vs. US-only		+0.07	+0.38	−0.01

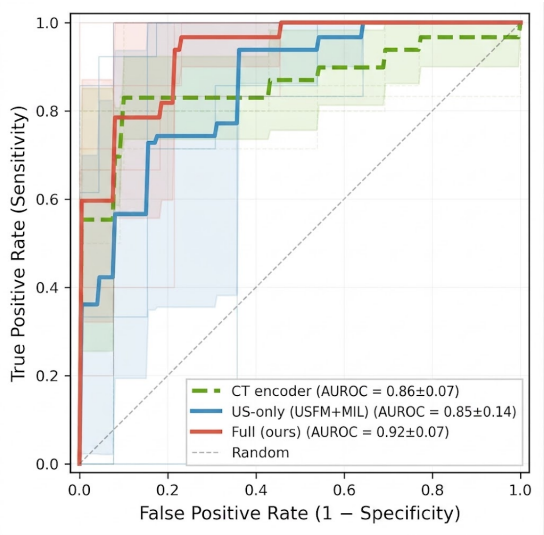

Fig. 2. AUROC curves (5-fold CV). Thin: per-fold. Bold: mean. Shaded: ±1 std.

source alone. Multi-sweep MAX-pool inference further amplifies sensitivity by providing multiple independent detection opportunities per patient.

Figure 2 shows AUROC curves across all folds, demonstrating consistent separation between our method and the US-only baseline.

3.3 Ablation Study

Table 2 isolates the contribution of each loss component. All rows use the same architecture with separate US and CT classifier heads, sweep-level training, and MAX-pool patient aggregation, differing only in which cross-modal loss terms are active.

Row (b) uses the cross-modal architecture with CT inputs but only classification loss, no explicit cross-modal transfer, yielding a modest +0.02 AUROC over the US-only baseline (a). The cross-modal losses provide complementary, additive improvements: contrastive alignment contributes +0.02 over the

Table 2. Ablation study: contribution of each loss component (5-fold CV, mean±std). Sensitivity and specificity at Youden-optimal threshold. Δ_{err}: percentage of remaining error reduced relative to the US-only baseline, computed as (AUROC − 0.85)/(1.0 − 0.85).

Configuration	AUROC	Sens	Spec	ΔAUROC	Δ_{err}(%)
(a) US-only baseline (no CT)	0.85±0.14	0.52±0.07	0.87±0.12	—	—
(b) $\mathcal{L}_{\text{cls}}$ only	0.87±0.08	0.84±0.15	0.88±0.11	+0.02	13.3
(c) $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{align}}$	0.89±0.08	0.83±0.15	0.89±0.11	+0.04	26.7
(d) $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{distill}}$	0.90±0.08	0.86±0.10	**0.90±0.10**	+0.05	33.3
(e) **Full (ours)**	**0.92±0.07**	**0.90±0.13**	0.86±0.14	**+0.07**	**46.7**

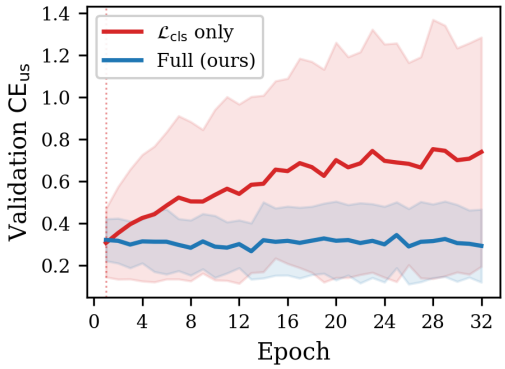

Fig. 3. Validation CE_{us} across epochs (5-fold mean ± std). Both configurations share the same cross-entropy on US logits. Without cross-modal losses, validation CE rises monotonically (2.4× initial value); the full objective maintains stable generalization, consistent with distillation acting as an implicit regularizer.

classification-only cross-modal baseline (c−b), while knowledge distillation contributes +0.03 (d−b), confirming distillation as the stronger individual transfer mechanism. Combining both (e) yields +0.07 AUROC over the US-only baseline, reducing the remaining error gap by 46.7%. Alignment constrains the embedding geometry, while distillation transfers the teacher's calibrated confidence through soft logit matching, additionally serving as an implicit regularizer [19] that prevents overfitting to hard label boundaries (Fig. 3), a critical advantage on small medical datasets.

3.4 Subgroup Analysis

Table 3 assesses robustness across operators and acoustic window quality. Three operators with varying experience performed the acquisitions; window quality was graded independently by an experienced sonographer.

No significant differences were observed across operators (p=0.59) or window quality (p=0.92), supporting deployment feasibility where expert sonographers

Table 3. Subgroup analysis across operators and acoustic window quality.

Subgroup	N	ICH%	AUROC	Sens	Spec	BalAcc
Operator 1	66	36.9	0.846	0.71	0.81	0.76
Operator 2	21	20.0	0.969	0.75	0.94	0.84
Operator 3	4	25.0	1.000†	1.00†	1.00†	1.00†
Kruskal–Wallis, p = 0.59						
Good	47	38.3	0.820	0.67	0.83	0.75
Average	33	19.4	0.960	0.83	0.84	0.84
Poor	11	45.5	1.000†	0.80	1.00†	0.90
Kruskal–Wallis, p = 0.92						

†Small sample; interpret with caution.

may not be available. The robustness to poorer windows likely reflects the MIL attention mechanism, which upweights informative frames regardless of overall quality, combined with MAX-pool aggregation across multiple sweeps.

4 Discussion and Conclusion

Ultra-early ICH identification is a high-value target because it determines immediate management and destination triage [3, 5]. In the prehospital phase, this distinction is rarely available without CT. Transcranial B-mode ultrasound is portable and fast, but operator-dependent acquisition and interpretation currently limit its use as a field screening tool.

Here, the main benefit of cross-modal distillation is not just higher discrimination but improved decision behavior. The US-only baseline ranks cases reasonably (AUROC 0.85) yet misses many hemorrhages at a practical threshold (sensitivity 0.52). Distillation shifts scores toward clearer positives, increasing sensitivity to 0.90. In deployment, the operating point would likely be set below the Youden threshold to further prioritize sensitivity, since the cost of missed hemorrhage outweighs unnecessary CT triage.

Table 1 does not imply that ultrasound is a better ICH detector than CT: NCCT defines our reference labels, and the CT row measures the quality of the available teacher signal under a small paired cohort, not the diagnostic ceiling of CT. The student also aggregates several sweeps per patient whereas the teacher sees a single volume, and soft targets from an imperfect teacher additionally act as a regularizer.

We therefore combine foundation-model initialization with CT-to-US distillation from a CT teacher trained on a large public dataset. Ablations suggest contrastive alignment reduces the modality gap, while distillation improves robustness and regularizes against classification-only overfitting (Fig. 3); together they reduce the error gap by 46.7%.

Limitations remain. The cohort is single-center (n=91 paired) and requires external validation. Six ICH cases were consistently misclassified, suggesting a

potential detection boundary for certain presentations; characterization is ongoing. Knowledge transfer is teacher-limited, consistent with the CT encoder dropping to AUROC 0.61 without RSNA pretraining (Sec. 2.1); stronger CT teachers plus prospective multicenter evaluation and clinical-variable integration are natural next steps.

References

1. Parry-Jones, A., Krishnamurthi, R., Ziai, W., Shoamanesh, A., Wu, S., Martins, S., Anderson, C.: World Stroke Organization (WSO): Global intracerebral hemorrhage factsheet 2025. International Journal of Stroke **20**(2), 145–150 (2025)
2. Steiner, T., Purrucker, J., de Sousa, D., Apostolaki-Hansson, et al: European Stroke Organisation (ESO) and European Association of Neurosurgical Societies (EANS) guideline on stroke due to spontaneous intracerebral haemorrhage. European Stroke Journal **1**(80), (2025)
3. Li, G., Lin, Y., Yang, J., et al.: Intensive ambulance-delivered blood-pressure reduction in hyperacute stroke. The New England Journal of Medicine **390**(20), 1862–1872 (2024)
4. Delcourt, C., Wang, X., Song, L., et al.: Effects of blood pressure lowering in relation to time in acute intracerebral haemorrhage: a pooled analysis of the four INTERACT trials. Lancet Neurology **24**(7), 571–579 (2025)
5. Broderick, J.P., Naidech, A.M., Elm, J.J., et al.: Recombinant factor VIIa versus placebo for spontaneous intracerebral haemorrhage within 2 h of symptom onset (FASTEST): a multicentre, double-blind, randomised, placebo-controlled, phase 3 trial. Lancet (2026). https://doi.org/10.1016/S0140-6736(26)00097-8
6. AHA/ASA International Stroke Conference 2026, https://professional.heart.org/en/meetings/international-stroke-conference/programming/late-breaking-science, last accessed 2026/02/20
7. Morotti, A., Boulouis, G., Nawabi, J., et al.: Using Noncontrast Computed Tomography to Improve Prediction of Intracerebral Hemorrhage Expansion. Stroke **54**(2), 567–674 (2023)
8. Almubayyidh, M., Alghamdi, I., Parry-Jones, A., Jenkins, D.: Prehospital identification of intracerebral haemorrhage: a scoping review of early clinical features and portable devices. BMJ Open **14**(e079316), (2023)
9. Anonymous Authors. Reference withheld for double-blind review.
10. Wang, H., Ma, C., Zhang, J., et al.: Learnable Cross-modal Knowledge Distillation for Multi-modal Learning with Missing Modality. In: Greenspan, H., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, vol. 14223. Springer, Cham. (2023). https://doi.org/10.1007/978-3-031-43901-8_21
11. Ceausescu, C., Alexe, B.: Multi-Dataset Cross-Domain Knowledge Distillation for Medical Image Segmentation In: 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025), pp. 3007–3016. Procedia Computer Science, (2025). https://doi.org/10.1016/j.procs.2025.09.425
12. Song, Y., Zheng, J., Lei, L., Ni, Z., Zhao, B., Hu, Y.: CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. Ultrasonics **122**(106706), (2022)
13. Flanders, AE., Prevedello, LM., Shih, G., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiology: Artificial Intelligence **2**(3), (2020)

14. van den Oord, A., Li, Y., Vinyals, O.: Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748 (2019)
15. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. Las Vegas, NV, USA (2016)
17. Jiao, J., Zhou, J., Li, X., et al.: USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**(103202) (2024)
18. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning. PMLR pp. 2127–2136, (2018)
19. Li, Y., Tay, F.E.H., Li, G., Wang, T., Feng, J.: Revisiting Knowledge Distillation via Label Smoothing Regularization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3902–3910. Seattle, WA, USA (2020)
20. Wang, C., Yang, Q., Huang, R., Song, S., Huang, G.: Efficient knowledge distillation from model checkpoints. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 607–619. New Orleans, LA, USA (2022)