U-ReLSE: Voxel-wise Evidential Uncertainty-Regularized MIL for Vertebral Metastasis Detection with Hybrid Supervision

Tatsuki Koike[1], Akimichi Ichinose[1], Akira Kudo[1,2], Yoshiro Kitamura[1], Mizuki Takei[1], Kazuma Kobayashi[3], Fumihiko Nakatani[3], Ryota Shibaki[3], Kimiteru Ito[3], Hiromi Matsumoto[3], Masatoshi Hori[2], Noriyuki Tomiyama[2], and Mototaka Miyake[3]

[1] Medical Systems Research & Development Center, FUJIFILM Corporation, Tokyo, Japan
[2] The University of Osaka Graduate School of Medicine, Suita, Japan
[3] National Cancer Center Hospital, Tokyo, Japan
tatsuki.koike@fujifilm.com

Abstract. In real-world CT lesion detection, reliable voxel-level masks are scarce, while vertebra-level labels mined from routine radiology reports are abundant but noisy, making hybrid supervision a practical challenge. We propose Uncertainty-Regularized Log-Sum-Exp (U-ReLSE) pooling, which converts voxel-level lesion predictions into vertebra-level probabilities by weighting the aggregation with voxel-wise Dirichlet evidential certainty. This design injects segmentation-supervised certainty into Multiple Instance Learning (MIL) pooling to suppress unreliable voxels, stabilizing learning under label noise and domain shift while retaining interpretable localization and uncertainty cues. We train a vertebral metastasis detection model on multi-institutional single-country data and evaluate it on two external, cross-national, multisite cohorts. U-ReLSE consistently improves vertebra-level detection over pooling baselines, specifically on Spine-Mets-CT-SEG (AUC 0.861 vs. 0.838–0.839; DeLong, $p<0.05$) and a US-private cohort (AUC 0.907 vs. 0.871–0.903). As U-ReLSE requires only replacing the pooling layer, it can be readily applied to other strong–weak supervised detection MIL tasks under label noise or domain shift.

Keywords: Bone Metastasis Detection · Hybrid Supervision · CT · Multiple Instance Learning · Uncertainty.

1 Introduction

Vertebral bone metastases are common in advanced cancer and can cause pain, paralysis, and fractures [3]. On CT, lesions may be subtle or obscured by artifacts, creating a strong demand for CAD. While deep learning approaches often focus on voxel-wise segmentation with dense lesion masks [8, 19, 11], clinical decisions are made at the vertebra level. Report-mined vertebra labels are

abundant but noisy due to omissions, vertebra-labeling errors, and report–image linkage failures. Moreover, although segmentation-assisted classification can improve detection and provide clinically intuitive explanations [15, 9], high-quality masks and curated vertebra-level labels are reliable but limited. Therefore, we target a hybrid setting that integrates (i) limited **Verified-Voxel** supervision (multimodality-confirmed masks), (ii) limited **Verified-Vertebra** supervision (multimodality-confirmed vertebra labels), and (iii) abundant **Report-Vertebra** supervision (noisy report labels).

Multiple instance learning (MIL) naturally fits this setting by aggregating instance predictions into vertebra-level decisions [6]. However, standard pooling methods often overlook reliability [13, 14, 2]. Existing uncertainty-aware MIL approaches primarily estimate uncertainty at the *instance/patch* level and incorporate it into pooling or instance selection [18, 10, 16, 5], mainly by learning evidence from weak bag- or instance-level supervision. In contrast, our hybrid setting provides limited but reliable *voxel-wise* lesion masks, enabling segmentation-supervised voxel-wise evidential learning [17, 20]. We then inject the resulting *Dirichlet certainty* as deterministic weights *within* a mask-aware log-sum-exp pooling operator (U-ReLSE), transferring certainty cues from strong supervision to stabilize pooling under noisy report labels and domain shift.

Most bone metastasis CAD studies are evaluated on in-house cohorts, which complicates systematic comparison across methods. Public datasets such as Spine-Mets-CT-SEG [4] could serve as external benchmarks but remain underused. Prior studies, including radiomics-based approaches [7], trained and evaluated their models on the same dataset, providing limited evidence on cross-dataset robustness under distribution shift.

Our contributions are: (1) **U-ReLSE** (Uncertainty-Regularized Log-Sum-Exp), an uncertainty-regularized MIL pooling strategy that uses segmentation-supervised *voxel-wise* evidential certainty to weight mask-aware log-sum-exp aggregation, yielding interpretable lesion and uncertainty maps. (2) a unified **hybrid training scheme** integrating voxel masks, reliable vertebra-level labels, and noisy report-derived labels into joint segmentation, evidential, and classification objectives. (3) external validation on **two cross-national external cohorts** (Spine-Mets-CT-SEG and US-private), demonstrating improved generalization under distribution shift.

2 Methods

2.1 Overview

We aim to detect vertebral bone metastases from CT scans by predicting a binary label $y_v \in \{0, 1\}$ for each vertebra v. Given a CT scan, we segment and index the vertebrae, crop a fixed-size 3D ROI around each vertebral centroid, and feed two windowed intensity channels into a shared 3D encoder based on a 3D Res U-Net-style architecture with four stages (channels: 32–64–128–128) and max pooling for downsampling. Two $1 \times 1 \times 1$ convolutional heads branch from the encoder: a

segmentation head that produces voxel-level logits and an evidential head that generates Dirichlet parameters α to quantify voxel-wise evidence. We aggregate voxel predictions into vertebra-level scores via the proposed U-ReLSE pooling, which down-weights uncertain regions. The resulting probability $\hat{p}_v$ is optimized using BCE with both Verified-Vertebra labels and Report-Vertebra labels mined from report impressions (e.g., "bone metastasis in T12" $\rightarrow y_{\mathrm{T12}}$=1), while Dice loss supervises segmentation where voxel masks are available. The segmentation logits and evidential maps further provide interpretable lesion localization cues (Figure 1).

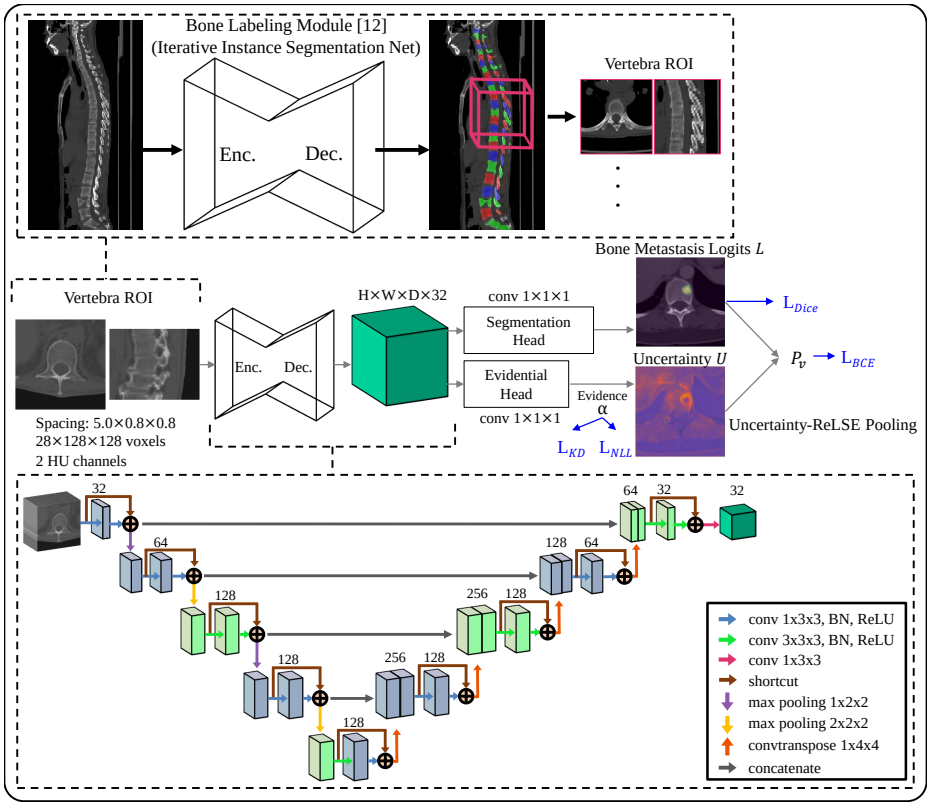

Fig. 1. Framework overview: vertebra ROI extraction, shared 3D encoder with segmentation/evidential heads, and U-ReLSE pooling for vertebra-level prediction.

2.2 Vertebra ROI and Input Representation

We obtain vertebral masks and indices using a pre-trained, fixed segmentation model [12] and resample each CT scan to a voxel spacing of $(z, y, x) =$

$(5.0, 0.8, 0.8)$ mm. For each vertebra, we crop a fixed-size ROI of $(28, 128, 128)$ voxels around the centroid with a margin covering both the vertebral body and posterior elements. The generated vertebral centroids and masks are used exclusively for ROI extraction, and the segmentation model was not fine-tuned on our datasets and is shared by all methods. Therefore, vertebra segmentation is a fixed preprocessing step and does not affect method comparisons. During evaluation, vertebral indices were curated by radiologists, and the ROI masks used for pooling were slightly dilated to reduce boundary sensitivity. Hounsfield Unit (HU) values are windowed into two channels: bone $[-300, 1200]$ and soft tissue $[-150, 500]$, normalized to $[0, 1]$.

2.3 Uncertainty-Regularized Log-Sum-Exp (U-ReLSE) Pooling

U-ReLSE aggregates voxel logits $\ell_{v,i}$ produced by the segmentation head into vertebra-level scores while down-weighting uncertain regions as estimated by the evidential head and restricting aggregation to the anatomical ROI, where i indexes voxel locations on the ROI output grid. To ensure anatomical relevance, we pool only over voxels within the vertebra mask. Let $m_{v,i} \in \{0, 1\}$ indicate whether voxel i lies within the vertebra–v mask (1 if inside, 0 otherwise), and let $N_v = \sum_i m_{v,i}$ denote the number of voxels inside the vertebra–v mask. The masked log-sum-exp (LSE) pooling is defined as:

$$s_v^{\mathrm{LSE}} = \frac{1}{\gamma} \log \Big(\frac{1}{N_v} \sum_i m_{v,i}\, e^{\gamma \ell_{v,i}} + \epsilon \Big), \quad \gamma > 0.$$

The evidential head outputs Dirichlet concentration parameters $\boldsymbol{\alpha}_{v,i}$, implemented as $\mathrm{softplus}(\cdot) + \mathbf{1}_K$ so that $\alpha_{v,i,c} \geq 1$, where $c \in \{1, \ldots, K\}$ indexes the class. We define the total evidence $S_{v,i}$ and uncertainty $U_{v,i}$ as:

$$S_{v,i} = \sum_c \alpha_{v,i,c}, \quad U_{v,i} = \frac{K}{S_{v,i}}, \quad (\text{here } K = 2 \text{ for binary classes}).$$

With $\alpha_{v,i,c} \geq 1$, we have $S_{v,i} \geq K$ and thus $U_{v,i} \in (0, 1]$. We compute certainty weights within the vertebral mask and normalize them over the ROI:

$$w_{v,i} = m_{v,i}(1 - U_{v,i}), \qquad W_v = \sum_j w_{v,j}, \qquad \tilde{w}_{v,i} = \frac{w_{v,i}}{W_v + \epsilon}.$$

Finally, U-ReLSE pooling computes the weighted LSE aggregation as:

$$s_v^{\text{U-ReLSE}} = \frac{1}{\gamma} \log \Big(\sum_i \tilde{w}_{v,i}\, e^{\gamma \ell_{v,i}} + \epsilon \Big), \qquad \hat{p}_v = \sigma\big(s_v^{\text{U-ReLSE}}\big).$$

In this formulation, voxels with high uncertainty $U_{v,i}$ are down-weighted via the $(1 - U_{v,i})$ term, while off-mask voxels are excluded via $m_{v,i}$. For numerical stability, we use log-sum-exp stabilization (max-shifted exponentiation), add a small ϵ to denominators and log arguments, and clamp masked-out logits to a large negative value. Additionally, γ is parameterized as a learnable scalar constrained to a finite range (here, $\gamma \in [4, 16]$). The output probability $\hat{p}_v$ is supervised with $\mathcal{L}_{\mathrm{bce}}$.

2.4 Hybrid Supervision and Loss

We train the model using heterogeneous supervision sources: (i) **Verified-Voxel** data (masks with multimodality-confirmed labels) to supervise the segmentation, evidential, and classification heads, and (ii) **Report-Vertebra** data (noisy report labels) to supervise classification only. Optionally, (iii) **Verified-Vertebra** data (multimodality-confirmed labels with or without masks) can be used to supervise classification when available.

Segmentation head ($\mathcal{L}_{\text{dice}}$). For vertebrae with voxel-wise masks, we optimize the segmentation head using soft Dice loss between the predicted probabilities $\hat{\mathbf{M}}_v$ and the ground truth $\mathbf{M}_v$:

$$\mathcal{L}_{\text{dice}} = 1 - \text{Dice}(\hat{\mathbf{M}}_v, \mathbf{M}_v).$$

Evidential head ($\mathcal{L}_{\text{edl}}$ and $\mathcal{L}_{\text{kd}}$). For voxels with Verified-Voxel masks (lesion or normal), the evidential head outputs Dirichlet parameters $\boldsymbol{\alpha}_{v,i}$, from which we compute voxel-wise expected class probabilities $\bar{\mathbf{p}}_{v,i}$ as $\bar{p}_{v,i,c} = \alpha_{v,i,c}/S_{v,i}$ (note that $\bar{\mathbf{p}}_{v,i}$ differs from the vertebra-level probability $\hat{p}_v$ computed via U-ReLSE pooling). The target label $y_{v,i,c}$ is defined only for Verified-Voxel data. The evidential head is trained using a data term on $\bar{\mathbf{p}}_{v,i}$:

$$\mathcal{L}_{\text{edl}} = \sum_{v,i} \sum_{c} \Big[(y_{v,i,c} - \bar{p}_{v,i,c})^2 + \frac{\bar{p}_{v,i,c}(1 - \bar{p}_{v,i,c})}{S_{v,i} + 1} \Big],$$

and a KL regularizer to a weak symmetric prior $\boldsymbol{\alpha}_0 = \mathbf{1}_K$ (here $K = 2$):

$$\mathcal{L}_{\text{kd}} = \sum_{v,i} \text{KL}\big(\text{Dir}(\boldsymbol{\alpha}_{v,i}) \,\|\, \text{Dir}(\boldsymbol{\alpha}_0)\big).$$

Vertebra classification ($\mathcal{L}_{\text{bce}}$ on U-ReLSE). Vertebra-level probability scores $\hat{p}_v$ obtained via U-ReLSE pooling are supervised by BCE with vertebra-level labels:

$$\mathcal{L}_{\text{bce}} = -\sum_{v} \big[y_v \log \hat{p}_v + (1 - y_v) \log(1 - \hat{p}_v) \big].$$

Total objective. The overall loss is given by:

$$\mathcal{L} = \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{edl}}\mathcal{L}_{\text{edl}} + \lambda_{\text{kd}}\mathcal{L}_{\text{kd}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}}.$$

We supervise segmentation with $\mathcal{L}_{\text{dice}}$, train the evidential head with $\mathcal{L}_{\text{edl}}$+$\mathcal{L}_{\text{kd}}$, and supervise vertebra-level classification with $\mathcal{L}_{\text{bce}}$ on U-ReLSE outputs.

3 Experiments

3.1 Datasets

We utilize a multi-institutional in-house dataset and two external datasets, each with three types of annotations: (i) **Verified-Voxel**: voxel-wise lesion masks

Table 1. Dataset statistics and characteristics. **Pt/Se/Vt**: Patients / Series / Vertebrae. **Annotation**: Annotation type available. **Usage**: Role in experimental splits.

Dataset	Size (Pt/Se/Vt)	Annotation	Usage
NCC (in-house)	174 / 198 / 765	Verified-Voxel	Train / Val / Test
Osaka Univ. (in-house)	253 / 305 / 1,166	Verified-Voxel	Train / Val
Osaka Univ. (in-house)	835 / 3,023 / 4,211	Report-Vertebra	Train
Spine-Mets [4]	55 / 55 / 185	Verified-Vertebra	Test
US-private	196 / 196 / 524	Report-Vertebra	Test

with multimodality-confirmed vertebra-level labels; (ii) **Verified-Vertebra**: confirmed vertebra-level labels; and (iii) **Report-Vertebra**: noisy vertebra-level labels derived from radiology reports. Table 1 summarizes the dataset statistics and usage. The in-house dataset includes CT scans from two hospitals. National Cancer Center Hospital (NCC) (Verified-Voxel) patients are split into 50 test cases and 10 validation cases, while The University of Osaka Hospital (Verified-Voxel) patients contribute 10 validation cases. The remaining 1,192 cases from both sites are used for training. The University of Osaka Hospital (Report-Vertebra) data are used solely for training. There is no patient overlap across train/val/test sets. External testing employs **Spine-Mets-CT-SEG** [4] (Verified-Vertebra) and **US-private** (Report-Vertebra). Note that the US-private labels, collected from 23 US sites via chronological random sampling, are report-derived and thus reflect a pragmatic, documentation-level clinical reference (i.e., *reported metastasis*), where subtle/small lesions may remain unlabeled. Ambiguous or non-vertebra-specific reports were excluded, and only vertebrae explicitly named in reports were assigned weak positive labels; unmentioned vertebrae were excluded from BCE rather than treated as negatives. In Verified-Voxel cases, vertebrae without positive masks were defined as negatives.

We define three training regimes for ablation analysis: (a) **Rep-Vert**: Trained on in-house Report-Vertebra data only. (b) **Vrf-VV**: Trained on in-house Verified-Voxel masks and Verified-Vertebra labels. (c) **Hybrid**: Trained on both (a) and (b). All regimes are evaluated on the same reliable validation and test sets.

3.2 Evaluation Metrics

We assess vertebra-level performance using ROC-AUC for metastasis detection, as well as threshold-based metrics including recall (sensitivity) and the average number of false-positive vertebrae per scan (FP/c). For datasets with weak, report-derived labels (notably US-private), these metrics quantify agreement with clinically documented findings rather than with exhaustive lesion ground truth. Nonetheless, this evaluation setting is highly relevant for assessing robustness and potential clinical impact at scale. For each model, we select the decision threshold on the in-house validation set such that FP/c is approximately 2.5, and then report recall and FP/c on each held-out test set using the fixed thresh-

old. We calculate 95% confidence intervals (CIs) for ROC-AUC via bootstrap resampling with 2,000 iterations.

3.3 Implementation Details

Each training sample consists of a vertebra ROI. The model is trained with Adam (batch size 14) for 20,000 steps, using a cosine-annealing learning rate schedule with a 1,000-step warm-up and a base learning rate of 1×10^{-4}. We set λ_{dice}=1.0, λ_{edl}=0.3, λ_{kd}=10^{-3}, and λ_{bce}=1.0. We select the checkpoint with the best validation accuracy. Mini-batches are constructed to include, as much as possible, a balanced mix of strongly labeled positives, strongly labeled negatives, and weakly labeled positive vertebrae. Standard 3D augmentations (random flips, small rotations, and intensity perturbations) are applied. Hyperparameters were selected based on the vertebra-level accuracy on the validation set and fixed before external testing.

3.4 Baseline Configurations

We compare four baselines: a 3D ROI classifier (no segmentation, global average pooling), segmentation-only (same encoder/hybrid supervision as ours, score by global max within mask), nGWP pooling, and Attention MIL pooling (same encoder/hybrid supervision as ours, without evidential outputs). Unless otherwise specified, all models utilize the same encoder and differ only in the pooling head and/or the presence of evidential outputs. Therefore, the performance differences primarily reflect the aggregation strategy and supervision regime.

4 Results and Discussion

4.1 External and In-house Evaluation

Table 2 summarizes the vertebra-level performance on two external test sets. Compared with the segmentation-only baseline, MIL-based pooling improves vertebra-level detection, and U-ReLSE achieves the highest ROC-AUC on both external datasets while reducing FP/c. The consistent gains on Spine-Mets-CT-SEG, which is distributionally further from our training cohort than US-private, indicate that voxel-wise evidential certainty regularizes MIL aggregation under domain shift. In our implementation, the evidential head adds only a single convolution (66 parameters), whereas Attention MIL introduces multiple fully connected layers (8.7k parameters excluding the shared encoder); thus, the gain is unlikely to be explained solely by increased model capacity.

We further ablate supervision regimes on Spine-Mets-CT-SEG and the in-house test set while keeping the backbone and pooling formulation fixed (Table 3). Hybrid supervision outperforms both Report-Vertebra-only supervision and Verified-Voxel/Verified-Vertebra supervision, and uncertainty-weighted pooling further improves vertebra-level ROC-AUC.

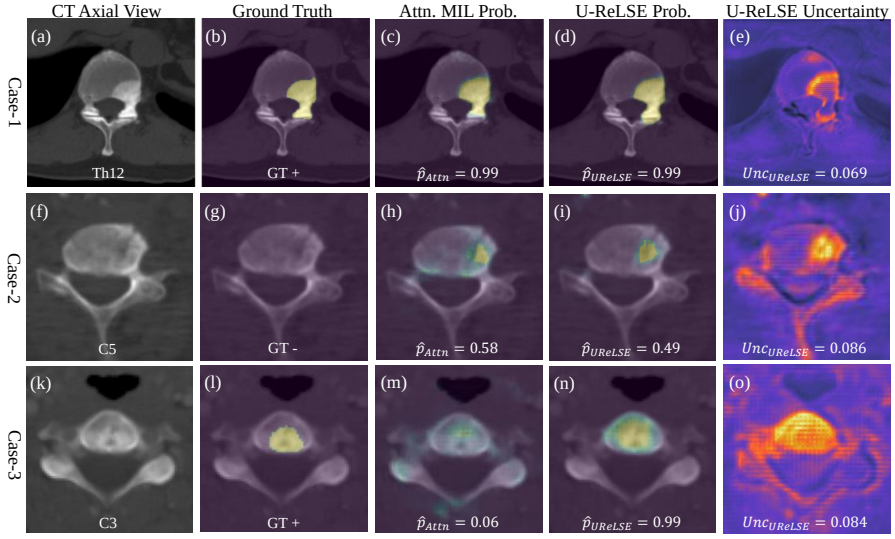

Fig. 2. Qualitative examples (Cases 1–3) showing U-ReLSE evidential uncertainty. Case 1: a clear positive with a focal lesion; Case 2: a hard negative with benign trabecular change; Case 3: a subtle low-contrast positive in which Attention MIL is diffuse but U-ReLSE concentrates on the lesion with higher evidence.

4.2 Evidential Pooling and Uncertainty

Qualitative examples (Figure 2) show that both Attention MIL and U-ReLSE localize strong activations on clear metastatic foci. For atypical hard negatives (benign trabecular change, Case 2), U-ReLSE yields low metastasis probability with high uncertainty, suggesting that uncertainty-based thresholding can reduce false positives compared with Attention MIL. In subtle positive cases (Case 3), where lesions are low-contrast and Attention MIL produces diffuse activations, U-ReLSE effectively concentrates on the lesion with high probability. Overall, voxel-level evidence estimation combined with uncertainty-regularized pooling improves robustness under heterogeneous supervision and domain shift.

5 Conclusion

We proposed a vertebra-level CAD framework for CT bone metastasis detection that combines voxel-level evidential learning with uncertainty-regularized log-sum-exp pooling (U-ReLSE), integrating Verified-Voxel masks, Verified-Vertebra labels, and noisy Report-Vertebra labels. The method consistently improved vertebra-level detection on a multi-institutional in-house cohort and two external datasets, indicating that instance-level uncertainty stabilizes MIL aggregation under label noise and domain shift. The segmentation logits and evidential maps provide interpretable localization cues that may support clinical adoption,

Table 2. Vertebra-level detection on two external test sets. We report ROC-AUC and threshold-based recall and FP/c (false-positive vertebrae per scan). Asterisk (*) indicates statistically significant improvement over all baselines (DeLong's test, $p<0.05$).

	Spine-Mets-CT-SEG			US-private		
Composition	AUC (95% CI)	Recall	FP/c	AUC (95% CI)	Recall	FP/c
3D ROI Classifier	0.773(0.730–0.816)	0.670	2.62	0.809(0.789–0.830)	0.798	3.96
Seg-Head only	0.809(0.776–0.841)	0.708	2.89	0.885(0.871–0.898)	0.865	2.76
nGWP [1]	0.838(0.803–0.872)	0.757	2.71	0.903(0.889–0.917)	0.870	2.62
Attn. MIL [6]	0.839(0.806–0.874)	0.751	2.70	0.871(0.854–0.887)	**0.887**	3.34
U-ReLSE (Ours)	**0.861***(0.829–0.891)	**0.778**	**2.43**	**0.907**(0.892–0.921)	**0.887**	**2.53**

Table 3. Spine-Mets-CT-SEG and in-house ablation of supervision and pooling. Vrf-VV: verified voxel-and-vertebra labels. Rep-Vert: report-derived vertebra labels. For the hybrid setting, we also report ReLSE [2] as a non-uncertainty baseline.

		Spine-Mets-CT-SEG			NCC (In-house) Test Set		
Superv.	Pooling	AUC (95% CI)	Recall	FP/c	AUC (95% CI)	Recall	FP/c
Rep-Vert	U-ReLSE	0.815(0.781–0.848)	0.643	2.76	0.810(0.769–0.847)	0.657	2.76
Vrf-VV	U-ReLSE	0.834(0.799–0.871)	0.714	2.56	0.844(0.806–0.879)	0.740	2.64
Hybrid	ReLSE [2]	0.831(0.795–0.864)	0.741	2.56	0.839(0.804–0.873)	0.707	2.50
Hybrid	U-ReLSE	**0.861***(0.829–0.891)	**0.778**	**2.43**	**0.862**(0.828–0.893)	**0.751**	**2.48**

and the framework is applicable to other weakly supervised detection settings. Future work will examine multi-class lesion characterization, longitudinal follow-up integration, and prospective clinical evaluation.

Disclosure of Interests. The authors (Koike, T., Ichinose, A., Kudo, A., Kitamura, Y., Takei, M.) are employees of FUJIFILM Corporation.

References

1. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4253–4262 (2020)
2. Chen, C.C., Yeh, Y.C., Lin, M.M., Yeh, C.Y.: Pixel-precise lesion localization in wsis via weakly supervised streaming convolution with relse and adaptive self-training. bioRxiv (2025). https://doi.org/10.1101/2024.10.03.616445
3. Coleman, R.E.: Clinical features of metastatic bone disease and risk of skeletal morbidity. Clinical cancer research **12**(20), 6243s–6249s (2006)
4. Haouchine, N., Hackney, D.B., Pieper, S.D., Wells, I.W.M., Sanhinova, M., Balboni, T.A., Spektor, A., Huynh, M.A., Kozono, D.E., Doyle, P., et al.: An open annotated dataset and baseline machine learning model for segmentation of vertebrae with metastatic bone lesions from ct. medRxiv (2024)

5. Hong, S., Lee, S., Shin, J., Jeong, J., Yi, M.Y.: Diagnose like a real pathologist: An uncertainty-focused approach for trustworthy multi-resolution multiple instance learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6132–6141 (2026)
6. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
7. Innocente, C., Iaconinoto, L., Notarangelo, D., Scalcione, A., Sergi, R., Velardi, A., Marullo, G., Vezzetti, E., Ulrich, L.: An end-to-end radiomic framework for automatic vertebral lesion classification and 3d visualization. Eng **7**(1), 18 (2026)
8. Kim, D.H., Seo, J., Lee, J.H., Jeon, E.T., Jeong, D., Chae, H.D., Lee, E., Kang, J.H., Choi, Y.H., Kim, H.J., et al.: Automated detection and segmentation of bone metastases on spine mri using u-net: a multicenter study. Korean Journal of Radiology **25**(4), 363 (2024)
9. Kong, Z., He, M., Luo, Q., Huang, X., Wei, P., Cheng, Y., Chen, L., Liang, Y., Lu, Y., Li, X., et al.: Multi-task classification and segmentation for explicable capsule endoscopy diagnostics. Frontiers in molecular biosciences **8**, 614277 (2021)
10. Liu, P., Ji, L.: Weakly-supervised residual evidential learning for multi-instance uncertainty estimation. arXiv preprint arXiv:2405.04405 (2024)
11. Liu, S., Feng, M., Qiao, T., Cai, H., Xu, K., Yu, X., Jiang, W., Lv, Z., Wang, Y., Li, D.: Deep learning for the automatic diagnosis and analysis of bone metastasis on bone scintigrams. Cancer Management and Research pp. 51–65 (2023)
12. Masuzawa, N., Kitamura, Y., Nakamura, K., Iizuka, S., Simo-Serra, E.: Automatic segmentation, localization, and identification of vertebrae in 3d ct images using cascaded convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 681–690. Springer (2020)
13. Murray, N., Perronnin, F.: Generalized max pooling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2473–2480 (2014)
14. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1713–1721 (2015)
15. Ryu, H., Shin, S.Y., Lee, J.Y., Lee, K.M., Kang, H.j., Yi, J.: Joint segmentation and classification of hepatic lesions in ultrasound images using deep learning. European radiology **31**(11), 8733–8742 (2021)
16. Schmidt, A., Morales-Alvarez, P., Molina, R.: Probabilistic attention based on gaussian processes for deep multiple instance learning. IEEE Transactions on Neural Networks and Learning Systems **35**(8), 10909–10922 (2023)
17. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
18. de Vente, C., van Ginneken, B., Hoyng, C.B., Klaver, C.C., Sánchez, C.I.: Uncertainty-aware multiple-instance learning for reliable classification: Application to optical coherence tomography. Medical Image Analysis **97**, 103259 (2024)
19. Zhang, Y., Li, J., Yang, Q., Yin, S., Hou, J., Cao, X., Ma, S., Wang, B., Luo, M., Zhou, F., et al.: A clinically applicable ai system for detection and diagnosis of bone metastases using ct scans. Nature Communications **16**(1), 4444 (2025)
20. Zou, K., Yuan, X., Shen, X., Wang, M., Fu, H.: Tbrats: Trusted brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 503–513. Springer (2022)