

Opportunistic Osteoporosis Diagnosis via Texture-Preserving Self-Supervision, Mixture of Experts and Multi-Task Integration

Jiaxing Huang^{1,5,†}, Heng Guo^{2,6,†}, Le Lu², Fan Yang³, Minfeng Xu^{2,6},
Ge Yang^{1,5,‡}, and Wei Luo^{4,‡}

¹ Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

² DAMO Academy, Alibaba Group

³ Union Hospital, Huazhong University of Science and Technology, China

⁴ Xiangya Hospital, Central South University, China

⁵ School of Artificial Intelligence, University of Chinese Academy of Sciences, China

⁶ Hupan Lab, Hangzhou 310023, China

ge.yang@ia.ac.cn, luowei0928@126.com

Abstract. Osteoporosis, characterized by reduced bone mineral density (BMD) and compromised bone microstructure, increases fracture risk in aging populations. While dual-energy X-ray absorptiometry (DXA) is the clinical standard for BMD assessment, its limited accessibility hinders diagnosis in resource-limited regions. Opportunistic computed tomography (CT) analysis has emerged as a promising alternative for osteoporosis diagnosis using existing imaging data. Current approaches, however, face three limitations: (1) underutilization of unlabeled vertebral data, (2) systematic bias from device-specific DXA discrepancies, and (3) insufficient integration of clinical knowledge such as spatial BMD distribution patterns. To address these, we propose a unified deep learning framework with three innovations. First, a self-supervised learning method using radiomic representations to leverage unlabeled CT data and preserve bone texture. Second, a Mixture of Experts (MoE) architecture with learned gating mechanisms to enhance cross-device adaptability. Third, a multi-task learning framework integrating osteoporosis diagnosis, BMD regression, and vertebra location prediction. Validated across three clinical sites and an external hospital, our approach demonstrates superior generalizability and accuracy over existing methods for opportunistic osteoporosis screening and diagnosis.

Keywords: Osteoporosis Diagnosis · Self-supervised Learning · Mixture of Experts · Multi-task Learning.

1 Introduction

Osteoporosis, marked by reduced bone mineral density (BMD) and deteriorated bone microstructure, predisposes affected individuals to increased fracture sus-

[†] Equal Contribution, [‡] Corresponding Author.

ceptibility [18]. With global aging populations, both the prevalence of osteoporosis and fracture-related socioeconomic burdens are projected to ascend substantially [1]. Current clinical guidelines for fracture risk stratification integrate quantitative BMD assessment via dual-energy X-ray absorptiometry (DXA) with established clinical risk parameters [14,9]. However, significant diagnostic disparities persist in developing countries due to limited DXA accessibility [22,9]. This disparity underscores the urgency to develop accessible alternatives using existing imaging modalities. Notably, opportunistic computed tomography (CT) analysis, leveraging existing imaging data acquired for alternative clinical indications, has emerged as a promising strategy for osteoporosis diagnosis without requiring additional radiation exposure [20,19].

Previous studies have demonstrated the feasibility of opportunistic osteoporosis screening and BMD quantification on clinically acquired CT imaging wherein the lumbar vertebrae are key regions suitable for diagnosis of osteoporosis [16,24,3]. Existing methodologies exhibit three critical limitations that impede real-world clinical generalization. First, previous approaches do not adequately leverage unlabeled images. In Figure 1(A), a subset of vertebrae in CT scans have the corresponding DXA measurements, typically L1-L4 [5]. Spine imaging data from chest and abdominal CT scans are abundant in hospital databases, but most unlabeled vertebral CT data remain underutilized [28]. Moreover, since osteoporosis manifests itself through changes in bone texture, there is an urgent need for a texture-preserving self-supervised learning algorithm to fully harvest the unlabeled vertebral dataset. Second, systematic bias arises from device-specific discrepancies in DXA measurements. Inter-manufacturer variability among DXA devices can yield divergent results for the same patients [7], and these discrepancies persist despite standardization efforts [6] as demonstrated in Figure 1(A). Current AI models for simulating DXA measurements from CT scans mostly focus on single-device scenarios, neglecting the inherent heterogeneity and complexity of practical environments. Consequently, models trained on single-device DXA labeled data exhibit compromised generalizability in multi-device clinical settings. Third, conventional architectures do not incorporate the clinical priors essential for accurate osteoporosis diagnosis. Notably, positional variations in BMD distribution across vertebrae [23] – a critical diagnostic feature – are not yet incorporated into existing models. These limitations underscore the need for more robust methodologies to improve real-world applicability.

To overcome these limitations, we propose a unified deep learning framework with three novel contributions. First, we develop a texture-preserving self-supervised learning method based on SwAV [2]. Unlike conventional random cropping operations that may compromise structural integrity, our approach uses a sliding window mechanism to extract radiomic features from the original CT images. This modification effectively employs unlabeled data for model pre-training. Second, we develop a Mixture of Experts (MoE) [29] architecture in the decoding part to enhance the model’s adaptability across multiple devices. This design enables dynamic selection of predictors via learned gating mechanisms. Third, given the clinical importance of spatial bone mineral dis-

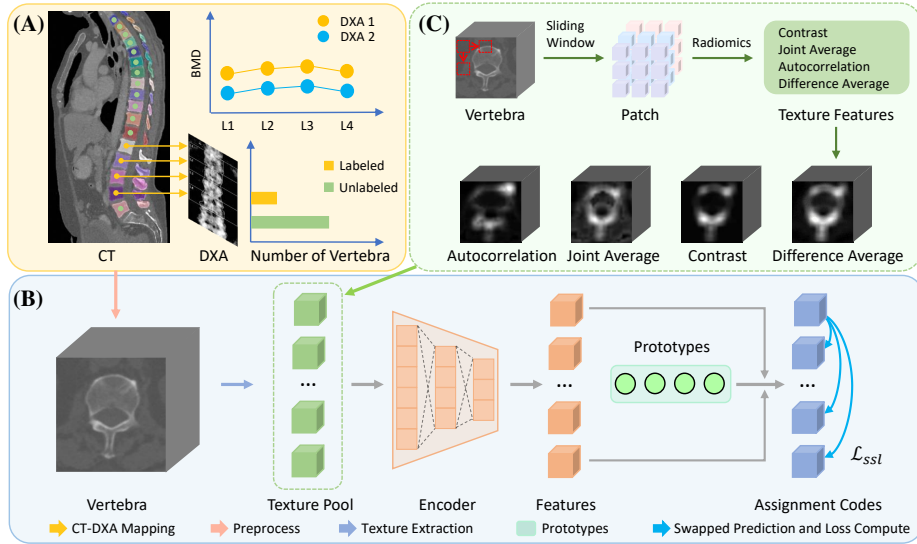


Fig. 1. (A) Data Acquisition: A localization model identified lumbar regions on CT scans corresponding to paired DXA measurements, correlating vertebra-level CT and DXA data. Notably, different DXA devices may produce varying BMD values for the same vertebra of a patient. (B) Overview of the texture-preserving self-supervised learning framework. (C) Workflow for constructing the texture feature pool.

tribution patterns and quantitative BMD values in osteoporosis assessment, a multi-task learning loss is used, synergistically integrating differential diagnosis, BMD regression and anatomical location prediction. This tripartite architecture improves the precision of osteoporosis diagnosis through complementary supervision. The proposed framework is validated in three independent clinical sites and an external hospital, demonstrating superior generalizability compared to existing methods.

2 Method

2.1 Texture-preserving Self-supervised Learning Framework

Self-supervised learning (SSL) has shown great potential in medical image analysis to leverage unlabeled data [26,15,27]. However, its application to osteoporosis diagnosis faces a critical challenge: texture features, which are pivotal for characterizing bone mineral density patterns, may be compromised by conventional random cropping operations in standard SSL frameworks. To this end, we propose a texture-preserving SSL method (TP-SwAV), as in Figure 1(B). Our approach consists of three key components. Firstly, we establish a texture pool containing both the original image and its texture features including autocorrelation, joint average, contrast, and difference average. We extract these spatially lo-

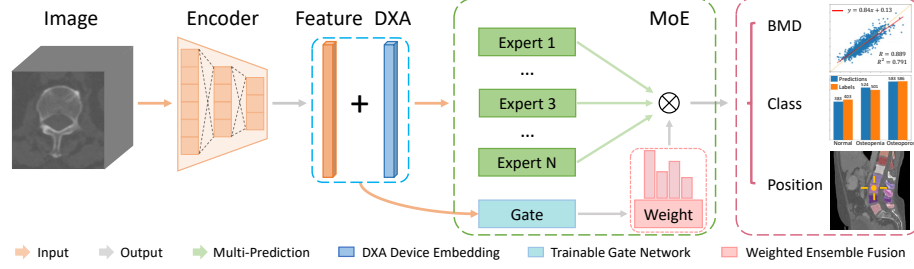


Fig. 2. Overflow of MoE-enhanced multi-task learning framework.

calized texture descriptors $\{\mathbf{TF}_1, \dots, \mathbf{TF}_M\}$ through a sliding window mechanism with adaptive interpolation [13] as shown in Figure 1(C) via the Pyradiomics toolkit [25]. This ensures dimensional consistency between the radiomic features and the original image. Secondly, the texture features $\{\mathbf{TF}_1, \dots, \mathbf{TF}_M\}$ are encoded into latent embeddings $\{\mathbf{P}_1, \dots, \mathbf{P}_M\}$ through a general encoder. These embeddings will be assigned to a set of K learnable prototypes $\{\mathbf{C}_1, \dots, \mathbf{C}_K\}$. For example, for $\mathbf{P}_t$ and $\mathbf{P}_s$, generated by two different views of an image, we compute their prototype assignment distributions $\mathbf{Q}_t$ and $\mathbf{Q}_s$ via optimal transport matching [4, 2]. Finally, a swapped prediction loss enforces cross-view consistency:

$$\mathcal{L}_{ssl} = - \sum_k \mathbf{Q}_s^{(k)} \log \frac{\exp(\frac{1}{\tau} \mathbf{P}_t^\top \mathbf{C}_k)}{\sum_{k'} \exp(\frac{1}{\tau} \mathbf{P}_t^\top \mathbf{C}_{k'})} - \sum_k \mathbf{Q}_t^{(k)} \log \frac{\exp(\frac{1}{\tau} \mathbf{P}_s^\top \mathbf{C}_k)}{\sum_{k'} \exp(\frac{1}{\tau} \mathbf{P}_s^\top \mathbf{C}_{k'})} \quad (1)$$

where τ is a temperature parameter. Simultaneously, a clustering regularization term ensures prototype diversity.

2.2 Mixture of Experts Architecture for Cross-Device Adaptability

To address the inherent heterogeneity across DXA devices, we propose a Mixture of Experts (MoE) architecture integrated into the decoding stage. As shown in Figure 2, the framework comprises two core components: (1) Device-specific expert modules: A set of N ($N = 3$) specialized expert heads $\{E_1, \dots, E_N\}$ that capture device-specific features. (2) Adaptive fusion mechanism: A dynamic gating network that generates adaptive attention weights based on the sum of image embeddings $\mathbf{P}_I$ and the corresponding DXA device embeddings $\mathbf{D}_I$ of image I . The gating operation is formulated as:

$$\alpha_i = \text{Softmax}(f_g(\mathbf{P}_I + \mathbf{D}_I)), \quad \sum_{i=1}^N \alpha_i = 1 \quad (2)$$

Table 1. Dataset composition across four clinical sites. D* represents the external site.

Site	DXA Devices	Patient			Vertebra w/o GT	Vertebra w/ GT		
		Train	Valid	Test		Train	Valid	Test
A	Hologic	1330	226	669	18308	2980	510	1490
B	GE Lunar	683	112	325	9232	1220	200	610
C	MEDIX	217	36	106	2716	801	129	398
D*	Hologic	0	0	193	0	0	0	271
ALL	Mix	2230	374	1293	30256	5001	839	2769

where $f_g(\cdot)$ denotes the parameterized gating function. The final prediction synthesizes expert outputs through weighted summation:

$$\text{Output} = \sum_{i=1}^N \alpha_i \cdot E_i \quad (3)$$

Introducing MoE structure at the decoding stage provides two key benefits to the model: (1) **Adaptive Feature Fusion:** The gating network automatically activates relevant experts via attention weights derived from input image embeddings and corresponding DXA device embeddings, enabling dynamic adaptation to different devices. (2) **Cross-Device Regularization:** Joint optimization of both expert-specific parameters and shared gating parameters leads to an implicit regularization effect, which prevents the model overfitting to any single device while maintaining discriminative power.

2.3 Multi-task Learning with Complementary Supervision

Our fine-tuning framework integrates three complementary supervision signals through multi-task learning, as illustrated in Figure 2. First, a BMD regression branch directly predicts continuous bone mineral density values from encoded image features, ensuring the model captures biologically meaningful patterns critical for osteoporosis assessment. Second, a classification branch models vertebral anatomical landmarks (T1-T12, L1-L6) as ordinal labels (1-18) to induce position-related regularization, thereby enabling the model to perceive vertebral positions, as vertebra at different positions exhibit distinct BMD distributions. Third, the primary diagnostic task employs cross-entropy loss for osteoporosis detection, enhanced by shared representations from auxiliary tasks. The unified objective function combines these components:

$$\mathcal{L}_{unified} = \alpha \mathcal{L}_{class} + \beta \mathcal{L}_{BMD} + \gamma \mathcal{L}_{position} \quad (4)$$

where $\mathcal{L}_{class}$ and $\mathcal{L}_{position}$ both utilize CE loss and $\mathcal{L}_{BMD}$ employs MSE loss. Hyperparameters ($\alpha = 1.0$, $\beta = 1.0$, $\gamma = 0.5$) balance task-specific gradients, enabling synergistic learning of diagnostic features through shared weights.

Table 2. Vertebra-level quantitative results of osteoporosis diagnosis.

Model	Pre-train	Loss	Site A					Site B				
			ACC	SEN	SPE	F1	P	ACC	SEN	SPE	F1	P
ResNet34	/	$\mathcal{L}_{class}$	83.89	76.86	89.24	80.49	***	86.56	62.24	91.21	59.80	***
Swin-UNTER	/	$\mathcal{L}_{class}$	84.70	83.52	85.36	79.64	***	88.69	76.06	90.35	61.02	***
ViT	MAE	$\mathcal{L}_{class}$	81.34	70.21	92.99	79.38	***	82.62	50.00	92.64	59.23	*
nnFormer	Alice	$\mathcal{L}_{class}$	85.50	79.00	90.38	82.35	***	88.52	70.93	91.41	63.54	***
SwinT	SwinMM	$\mathcal{L}_{class}$	85.37	86.36	84.86	80.04	***	87.38	63.81	92.28	63.51	***
ResNet34	VOCO	$\mathcal{L}_{class}$	83.83	79.39	86.71	79.45	***	87.21	64.89	91.28	61.00	***
ResNet34	SwAV	$\mathcal{L}_{class}$	83.36	78.84	86.28	78.84	***	87.05	66.67	90.17	57.75	***
ResNet34	TP-SwAV	$\mathcal{L}_{class}$	84.43	80.94	86.60	79.97	/	86.72	62.89	91.23	60.10	/
ResNet34	TP-SwAV	$\mathcal{L}_{unified}$	84.76	81.44	86.83	80.38	/	88.19	67.00	92.35	65.05	/
ResNet34+MoE	TP-SwAV	$\mathcal{L}_{unified}$	86.64	83.19	88.86	82.98	Base	89.02	69.31	92.92	67.63	Base
Model	Pre-train	Loss	Site C					Site D				
			ACC	SEN	SPE	F1	P	ACC	SEN	SPE	F1	P
ResNet34	/	$\mathcal{L}_{class}$	89.20	71.43	92.11	65.04	*	73.80	78.70	70.55	70.54	***
Swin-UNTER	/	$\mathcal{L}_{class}$	89.95	81.40	90.99	63.64	***	73.43	92.96	66.50	64.71	***
ViT	MAE	$\mathcal{L}_{class}$	87.18	59.09	93.86	67.10	***	78.59	76.22	81.25	78.99	**
nnFormer	Alice	$\mathcal{L}_{class}$	91.21	88.10	91.57	67.89	***	79.70	90.62	73.71	75.98	**
SwinT	SwinMM	$\mathcal{L}_{class}$	89.45	87.88	89.59	58.00	**	80.07	92.47	73.60	76.11	***
ResNet34	VOCO	$\mathcal{L}_{class}$	88.44	69.81	91.30	61.67	*	74.54	81.37	70.41	70.64	**
ResNet34	SwAV	$\mathcal{L}_{class}$	88.44	68.42	91.79	62.90	*	79.70	93.33	72.93	75.34	*
ResNet34	TP-SwAV	$\mathcal{L}_{class}$	90.70	77.78	92.73	69.42	/	78.97	94.19	71.89	73.97	/
ResNet34	TP-SwAV	$\mathcal{L}_{unified}$	91.21	76.67	93.78	72.24	/	79.33	87.83	74.40	76.27	/
ResNet34+MoE	TP-SwAV	$\mathcal{L}_{unified}$	90.95	74.60	94.02	72.31	Base	82.65	85.83	80.13	81.42	Base

3 Experiments

3.1 Dataset and Evaluation Metrics

We conducted a multicenter retrospective study utilizing data from three tertiary hospitals (Sites A-C, March 2018-December 2023) for model development and one external testing site (Site D, April 2022-December 2023). Inclusion criteria required paired non-contrast chest/abdominal CT and dedicated DXA scans within six months. Exclusion criteria included: 1) poor image quality; 2) metal implants; 3) vertebral compression fractures; 4) failed automated vertebrae identification and 5) missing lumbar spine data.

CT scans were obtained using standard non-contrast protocols across multiple vendors (Siemens, GE, Philips, etc.). DXA measurements (L1-L4 vertebrae) were performed using Hologic, GE Lunar, and MEDIX systems by certified technicians following standardized protocols. The Med-Query algorithm [8] established spatial CT-DXA correlations via automated vertebral localization and rotated ROI extraction, enabling precise anatomical mapping and corresponding BMD quantification. As shown in Table 1, the final dataset comprised 2230 patients (5001 vertebrae) for training, 374 (839 vertebrae) for validation, 1100 patients (2498 vertebrae) for internal testing and 193 patients (271 vertebrae) for external testing on Site D. Additionally, 30256 unlabeled vertebrae were used for pre-training. For comparative analysis, we selected two classification models trained from scratch (ResNet34 [12] and Swin-UNETR [10]) as baseline methods, along with several representative self-supervised learning (SSL) approaches. These include two general-purpose SSL methods (MAE [11] and SwAV [2]) and

Table 3. Patient-level quantitative results of osteoporosis diagnosis.

Model	Pre-train	Loss	Site A				Site B			
			ACC	SEN	SPE	F1	ACC	SEN	SPE	F1
ResNet34	/	$\mathcal{L}_{class}$	83.89	77.18	89.55	81.43	84.65	76.09	89.81	78.85
Swin-UNTER	/	$\mathcal{L}_{class}$	85.44	85.22	85.57	81.44	85.92	84.85	86.38	78.32
ViT	MAE	$\mathcal{L}_{class}$	83.86	76.79	90.99	82.69	81.85	75.55	90.61	69.86
nnFormer	Alice	$\mathcal{L}_{class}$	86.08	82.61	88.55	83.12	86.38	82.02	88.56	80.05
SwinT	SwinMM	$\mathcal{L}_{class}$	85.80	89.39	83.73	82.18	87.69	71.43	91.60	69.23
ResNet34	VOCO	$\mathcal{L}_{class}$	83.51	79.32	86.50	80.06	84.38	78.10	87.67	77.49
ResNet34	SwAV	$\mathcal{L}_{class}$	83.76	78.92	87.39	80.62	84.38	77.66	87.99	77.66
ResNet34	TP-SwAV	$\mathcal{L}_{class}$	85.82	82.91	87.83	82.65	86.01	80.80	88.71	79.74
ResNet34	TP-SwAV	$\mathcal{L}_{unified}$	84.60	79.94	88.61	82.75	92.00	81.54	94.62	80.30
ResNet34+MoE	TP-SwAV	$\mathcal{L}_{unified}$	87.74	83.10	91.88	86.33	93.23	86.89	94.70	82.81
Model	Pre-train	Loss	Site C				Site D			
			ACC	SEN	SPE	F1	ACC	SEN	SPE	F1
ResNet34	/	$\mathcal{L}_{class}$	84.91	74.07	88.61	71.43	82.93	76.00	87.30	77.47
Swin-UNTER	/	$\mathcal{L}_{class}$	83.02	78.95	83.91	62.50	83.55	85.45	82.77	75.20
ViT	MAE	$\mathcal{L}_{class}$	83.96	70.00	89.47	71.19	79.79	82.56	77.57	78.48
nnFormer	Alice	$\mathcal{L}_{class}$	86.79	94.12	85.39	69.57	85.10	83.49	85.90	78.81
SwinT	SwinMM	$\mathcal{L}_{class}$	81.13	90.91	80.00	67.80	79.79	91.18	73.60	76.07
ResNet34	VOCO	$\mathcal{L}_{class}$	79.25	78.12	83.13	67.69	83.24	79.20	85.41	76.74
ResNet34	SwAV	$\mathcal{L}_{class}$	86.79	80.00	88.89	74.07	83.40	79.56	85.44	76.91
ResNet34	TP-SwAV	$\mathcal{L}_{class}$	86.82	80.10	89.18	74.27	84.63	82.34	85.80	78.30
ResNet34	TP-SwAV	$\mathcal{L}_{unified}$	88.68	84.00	90.12	77.78	81.87	88.46	77.39	79.77
ResNet34+MoE	TP-SwAV	$\mathcal{L}_{unified}$	88.68	81.48	91.14	78.57	84.97	90.24	81.08	83.62

three recently proposed medical imaging-oriented methods (SwinMM [26], Alice [15], and VOCO [27]). Due to architectural constraints of certain self-supervised frameworks, the backbones were not fully standardized across methods. Specifically, nnFormer [30] was used in Alice, Vision Transformer (ViT) [21] in MAE, and Swin Transformer (SwinT) [17] in SwinMM.

All models were trained under standardized hyperparameters to ensure comparability. For self-supervised pre-training, configurations included: batch size=32, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=0.1), and linear warmup learning rate scheduling (initial lr=1e-4, warmup epochs=100) over 300 total epochs. During fine-tuning, we maintained identical batch-size and optimizer parameters while adjusting learning dynamics: reduced initial learning rate (lr=1e-5), extended warmup duration (150 epochs), and increased the maximum training cycles to 400 epochs. Classification performance was quantified using four metrics: accuracy (ACC), sensitivity (SEN), specificity (SPE), and F1-score (F1). In this study, the evaluation metrics ACC, SEN, SPE, and F1, are expressed as percentages (%). Patient-level diagnostic results were derived from vertebra-level predictions. Specifically, for a patient with N vertebrae, the patient was classified as osteoporosis if any one of the vertebrae was diagnosed or predicted as osteoporosis. This principle guided the computation of patient-level outcomes.

3.2 Main Results

The experimental results presented in Table 2 and Table 3 lead to three principal findings: (1) Compared to SwAV, the proposed TP-SwAV substitutes conventional cropping operations with pre-computed feature pool extraction. This

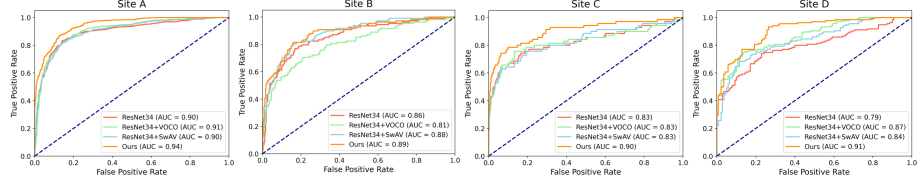


Fig. 3. Receiver operating characteristic (ROC) curves.

Table 4. Ablation study of loss function. Table 5. Ablation study of backbone.

Model	Loss	Site A	Site B	Site C	SiteD
ResNet34	$\mathcal{L}_{class}$	80.49	59.80	65.04	70.54
ResNet34	$\mathcal{L}_{class} + \mathcal{L}_{BMD}$	80.87	60.64	68.47	70.74
ResNet34	$\mathcal{L}_{class} + \mathcal{L}_{position}$	80.54	61.49	65.62	71.52
ResNet34	$\mathcal{L}_{unified}$	81.18	61.70	70.97	72.18

Model	Pre-train	Loss	Site A	Site B	Site C	SiteD
SwinT	/	$\mathcal{L}_{class}$	79.14	60.69	56.46	70.19
SwinT	SwinMM	$\mathcal{L}_{class}$	80.04	63.51	58.00	76.11
SwinT	TP-SwAV	$\mathcal{L}_{class}$	80.00	61.05	58.33	82.44
SwinT+MoE	TP-SwAV	$\mathcal{L}_{unified}$	80.98	64.00	60.18	80.31

modification significantly enhances diagnostic performance across all four clinical centers, substantiating the critical role of texture information in osteoporosis assessment. (2) Deep supervision through concurrent bone mineral density regression and anatomical position prediction demonstrates substantial improvements in diagnostic accuracy, confirming the synergistic benefits of multi-task learning frameworks. (3) The Mixture-of-Experts (MoE) architecture demonstrates superior adaptability to multi-center/multi-device scenarios. When integrated with multi-task learning and TP-SwAV, the ResNet34-based architecture achieves state-of-the-art F1-scores in both vertebra-level and patient-level osteoporosis diagnoses. Notably, patient-level validation yields F1-scores of 86.33% (Site A), 82.81% (Site B), 78.57% (Site C) for internal cohorts, and 83.62% for external testing (Site D). (4) We performed paired t-tests to compare our method with other approaches. The results presented in Table 2 demonstrate statistically significant differences compared to these methods across all four sites. The significance levels are denoted as follows: ‘*’ indicates $p < 0.05$, ‘**’ represents $p < 0.01$, and ‘***’ corresponds to $p < 0.001$. In addition, we constructed receiver operating characteristic (ROC) curves for the ResNet34-based methods to provide additional validation of our approach’s efficacy as shown in Figure 3.

3.3 Ablation Study

To validate the efficacy of multi-task supervision, we conducted experiments using ResNet34 (Table 4), demonstrating that both BMD regression ($\mathcal{L}_{BMD}$) and position prediction ($\mathcal{L}_{position}$) significantly improved performance, as indicated by increased F1-scores over the single-task baseline. Furthermore, to assess generalizability, we implemented our framework with Swin Transformer as the backbone. Results in Table 5 show consistent performance gains (F1-score) across four clinical datasets when incorporating TP-SwAV, MoE structure, and the proposed unified loss ($\mathcal{L}_{unified}$), confirming the adaptability of our approach to both CNN and Transformer architectures.

4 Conclusion

This study presents a unified deep learning framework addressing three critical challenges in opportunistic osteoporosis diagnosis using CT imaging. First, the texture-preserving TP-SwAV algorithm effectively utilizes unlabeled vertebral data through extracted radiomics features, overcoming limitations of random cropping operations in conventional SSL methods. Second, the MoE architecture with adaptive gating mechanisms mitigates device-specific variability in DXA measurements, enhancing cross-device generalizability. Third, multi-task learning integrating BMD regression, position prediction, and diagnostic classification leverages complementary clinical priors to improve diagnostic accuracy. Validated across multiple centers, our framework demonstrates superior adaptability and robustness, providing a scalable solution for osteoporosis screening in resource-constrained regions and reducing reliance on DXA equipment.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (grant 92354307), the National Key R&D Program of China (grant 2024YFF0729202), the Strategic Priority Research Program of the CAS (XDA0460305), and the Fundamental Research Funds for the Central Universities (grant E3E45201X2). Support was also provided by Alibaba Group via the Alibaba Research Intern Program.

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

References

1. Burge, R., Dawson-Hughes, B., Solomon, D.H., Wong, J.B., King, A., Tosteson, A.: Incidence and economic burden of osteoporosis-related fractures in the united states, 2005–2025. *Journal of Bone and Mineral Research* **22**(3), 465–475 (2007)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* **33**, 9912–9924 (2020)
3. Chen, Y.C., Li, Y.T., Kuo, P.C., Cheng, S.J., Chung, Y.H., Kuo, D.P., Chen, C.Y.: Automatic segmentation and radiomic texture analysis for osteoporosis screening using chest low-dose computed tomography. *European Radiology* **33**(7), 5097–5106 (2023)
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems* **26** (2013)
5. El Maghraoui, A., Roux, C.: Dxa scanning in clinical practice. *QJM: An International Journal of Medicine* **101**(8), 605–617 (2008)
6. Fan, B., Lu, Y., Genant, H., Fuerst, T., Shepherd, J.: Does standardized bmd still remove differences between hologic and ge-lunar state-of-the-art dxa systems? *Osteoporosis International* **21**, 1227–1236 (2010)
7. Genant, H.K., Grampp, S., Glüer, C.C., Faulkner, K.G., Jergas, M., Engelke, K., Hagiwara, S., van Kuijk, C.: Universal standardization for dual x-ray absorptiometry: Patient and phantom cross-calibration results. *Journal of Bone and Mineral Research* **9**(10), 1503–1514 (1994)

8. Guo, H., Zhang, J., Yan, K., Lu, L., Xu, M.: Med-query: Steerable parsing of 9-dof medical anatomies with query embedding. *IEEE Journal of Biomedical and Health Informatics* (2024)
9. Haseltine, K.N., Chukir, T., Smith, P.J., Jacob, J.T., Bilezikian, J.P., Farooki, A.: Bone mineral density: clinical relevance and quantitative assessment. *Journal of Nuclear Medicine* **62**(4), 446–454 (2021)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI brainlesion workshop*. pp. 272–284. Springer (2021)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16000–16009 (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
13. Huang, J., Zhou, Y., Liu, G., Guo, H., Yang, G.: Representing topological self-similarity using fractal feature maps for accurate segmentation of tubular structures. In: *European Conference on Computer Vision*. pp. 143–160. Springer (2024)
14. Jang, S., Graffy, P.M., Ziemlewicz, T.J., Lee, S.J., Summers, R.M., Pickhardt, P.J.: Opportunistic osteoporosis screening at routine abdominal and thoracic ct: normative l1 trabecular attenuation values in more than 20 000 adults. *Radiology* **291**(2), 360–367 (2019)
15. Jiang, Y., Sun, M., Guo, H., Bai, X., Yan, K., Lu, L., Xu, M.: Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15859–15869 (2023)
16. Krishnaraj, A., Barrett, S., Bregman-Amitai, O., Cohen-Sfady, M., Bar, A., Chetrit, D., Orlovsky, M., Elnekave, E.: Simulating dual-energy x-ray absorptiometry in ct using deep-learning segmentation cascade. *Journal of the American College of Radiology* **16**(10), 1473–1479 (2019)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10012–10022 (2021)
18. on Osteoporosis, N.C.D.P., Prevention, D., et al.: Osteoporosis prevention, diagnosis, and therapy. *JAMA* **285**(6), 785–795 (2001)
19. Paderno, A., Ataide Gomes, E.J., Gilberg, L., Maerkisch, L., Teodorescu, B., Koç, A.M., Meyer, M.: Artificial intelligence-enhanced opportunistic screening of osteoporosis in ct scan: a scoping review. *Osteoporosis International* **35**(10), 1681–1692 (2024)
20. Pickhardt, P.J., Pooler, B.D., Lauder, T., del Rio, A.M., Bruce, R.J., Binkley, N.: Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Annals of Internal Medicine* **158**(8), 588–595 (2013)
21. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12179–12188 (2021)
22. Siris, E.S., Miller, P.D., Barrett-Connor, E., Wehren, L.E., Abbott, T.A., Berger, M.L., Santora, A.C., Sherwood, L.M.: Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal women: results from the national osteoporosis risk assessment. *JAMA* **286**(22), 2815–2822 (2001)

23. Slosman, D., Rizzoli, R., Donath, A., Bonjour, J.P.: Vertebral bone mineral density measured laterally by dual-energy x-ray absorptiometry. *Osteoporosis International* **1**, 23–29 (1990)
24. Tariq, A., Patel, B.N., Sensakovic, W.F., Fahrenholtz, S.J., Banerjee, I.: Opportunistic screening for low bone density using abdominopelvic computed tomography scans. *Medical Physics* **50**(7), 4296–4307 (2023)
25. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. *Cancer Research* **77**(21), e104–e107 (2017)
26. Wang, Y., Li, Z., Mei, J., Wei, Z., Liu, L., Wang, C., Sang, S., Yuille, A.L., Xie, C., Zhou, Y.: Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 486–496. Springer (2023)
27. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22873–22882 (2024)
28. Yang, J., Liao, M., Wang, Y., Chen, L., He, L., Ji, Y., Xiao, Y., Lu, Y., Fan, W., Nie, Z., et al.: Opportunistic osteoporosis screening using chest ct with artificial intelligence. *Osteoporosis International* **33**(12), 2547–2561 (2022)
29. Yuksel, S.E., Wilson, J.N., Gader, P.D.: Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems* **23**(8), 1177–1193 (2012)
30. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing* **32**, 4036–4045 (2023)