IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 January 2023, accepted 22 February 2023, date of publication 3 March 2023, date of current version 8 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3252361

RESEARCH ARTICLE

Multi-View Computed Tomography Network for Osteoporosis Classification

DONG HWAN HWANG[1], SO HYEON BAK[2] TAE-JUN HA[3], YOON KIM[1,4], WOO JIN KIM[5], AND HYUN-SOO CHOI[1,6]

[1]Ziovision Inc., Chuncheon 24341, South Korea
[2]Department of Radiology, University of Ulsan College of Medicine, Seoul 44610, South Korea
[3]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Seoul 05505, South Korea
[4]Department of Computer Science and Engineering, Kangwon National University, Chuncheon 24341, South Korea
[5]Department of Internal Medicine, Kangwon National University, Chuncheon 24341, South Korea
[6]Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Hyun-Soo Choi (choi.hyunsoo@seoultech.ac.kr)

This work was supported by the Promotion of Innovative Businesses for Regulation-Free Special Zones funded by the Ministry of Small and Medium-sized Enterprises (SMEs) and Startups (MSS, South Korea) (P0020626).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Kangwon National University Hospital under Approval No. KNUH-A-2021-03-020-002.

ABSTRACT Osteoporosis is a skeletal disease that is difficult to identify in advance of symptoms. Existing skeletal disease screening methods, such as dual-energy X-ray absorptiometry, are only used for specific purpose due to cost and safety reasons once symptoms develop. Early detection of osteopenia and osteoporosis using other modalities for relatively frequent examinations is helpful in terms of early treatment and cost. Recently, many studies have proposed deep learning-based osteoporosis diagnosis methods for various modalities and achieved outstanding results. However, these studies have limitations in clinical use because they require tedious processes, such as manually cropping a region of interest or diagnosing osteoporosis rather than osteopenia. In this study, we present a classification task for diagnosing osteopenia and osteoporosis using computed tomography (CT). Additionally, we propose a multi-view CT network (MVCTNet) that automatically classifies osteopenia and osteoporosis using two images from the original CT image. Unlike previous methods that use a single CT image as input, the MVCTNet captures various features from the images generated by our multi-view settings. The MVCTNet comprises two feature extractors and three task layers. Two feature extractors use the images as separate inputs and learn different features through dissimilarity loss. The target layers learn the target task through the features of the two feature extractors and then aggregate them. For the experiments, we use a dataset containing 2,883 patients' CT images labeled as normal, osteopenia, and osteoporosis. Additionally, we observe that the proposed method improves the performance of all experiments based on the quantitative and qualitative evaluations.

INDEX TERMS Computed tomography, osteopenia, osteoporosis, deep learning, diagnosis.

I. INTRODUCTION

Osteoporosis is a skeletal disease that increases the risk of fractures because of a decrease in calcium mass and weakening of bone strength [5]. Osteoporosis is a major cause of fractures and is difficult to recognize in advance of symptoms. For these reasons, over half of fracture patients have never received a timely osteoporotic diagnosis [6], [7]. As the population ages, many experts expect the number of osteoporotic patients to rise, increasing economic costs, injuries, and death rates [8]. Early detection and treatment of osteoporosis are adequate to prepare for these risks, and prevention through early diagnosis and treatment is essential [9].

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian A. Linte.

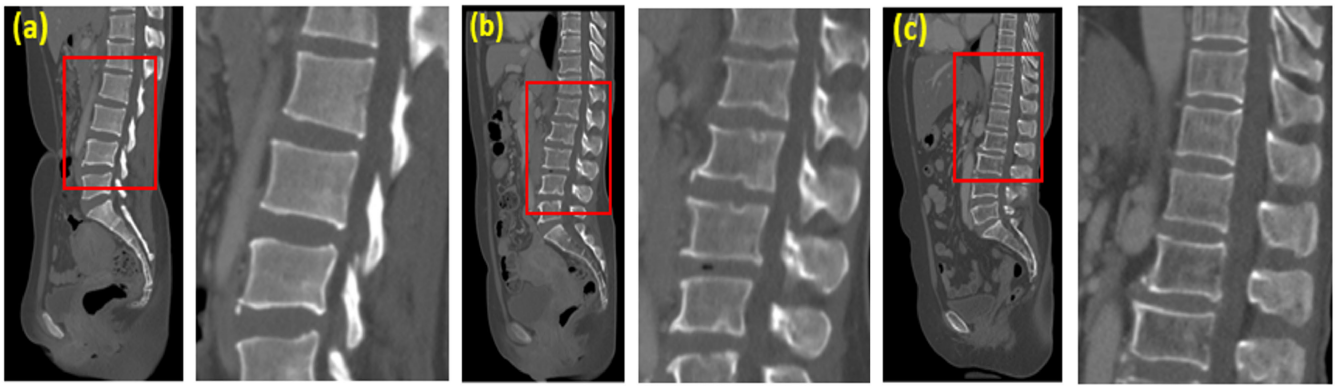

FIGURE 1. An example of sagittal CT scan dataset. (a) Normal, (b) Osteopenia and (c) Osteoporosis. The display window is [−450, 1000] HU.

TABLE 1. List of osteoporosis diagnosis studies.

Method	Modality	Dataset (Train / Test)	ROI	Model	AUC
Nam et al. [1]	CT	158 / 40	w/	Logistic classifier	0.90
Yasaka et al. [2]	CT	1,665 / 95	w/	4-layer CNN	0.97
Rastegar et al. [3]	DXA	147 / cross validation	w/	RF, Ensemble	0.78
Zhang et al. [4]	X-ray	1,616 / 396, 348	w/	5-layer CNN	0.81
MVCTNet	CT	2,283 / 600	w/o	ResNet-18	0.96

Dual-energy X-ray absorptiometry (DXA) is widely used to diagnose osteoporosis risk. However, DXA is not frequently performed until symptoms (e.g., fracture) appear and can cause unnecessary exposure to the other organs located around the test site. For these reasons, it is necessary to utilize other common modalities for early diagnosis. For example, computed tomography (CT) is frequently performed and has relatively less risky. Additionally, CT can be used to assess osteoporosis risks by experts. Fig. 1 shows the CT images of normal, osteopenia, and osteoporosis patients. The regions marked by red rectangles in the images show the differences between the three cases. The normal case has a uniform texture, whereas osteopenia and osteoporosis show texture differences with a decrease in the Hounsfield unit (HU). Because of these characteristics in CT, it is possible to analyze the presence of osteoporosis and osteopenia. This can provide an opportunity for early treatment and reduce the social and economic burden of osteoporotic fractures. Therefore, we intend to use CT images to diagnose osteoporosis. We conduct an osteoporotic classification task on a sagittal CT image which is relatively easy to obtain. For the task, we acquired our dataset from patients who underwent contrast-enhanced abdomen CT scans and DXA tests. Further, we provide details of the dataset in Section III-E.

With recent developments in deep learning, some studies [2], [3], [10], [11] have attempted to diagnose osteoporosis and achieved meaningful success. However, they still face significant challenges because of limited data and manual processes, which involve tedious work by radiology experts. Yasaka et al. [2] required the CT images to be cropped around the vertebrae. Yamamoto et al. [11] manually cropped the hip joint areas on the X-ray images. Moreover, they only used a CT image clipped within a certain HU range. HU has a more extensive pixel range than a photo image, and medical experts use a specific HU range to identify osteoporosis. However, deep neural networks may degrade the diagnostic performance of osteoporosis because of insufficient or unnecessary information.

In photo-domain computer vision, many studies [12], [13], [14], [15] have proposed powerful methods that leverage different views of the same instance in various fields, such as self-supervised and semi-supervised learning. Motivated by these, we propose a simple but powerful method called a multi-view CT network (MVCTNet), which uses two views generated by the CT domain knowledge (different HU clip settings in the same CT image). The MVCTNet consists of two feature extractors that extract different features from two perspectives and three target layers that learn the classification tasks. Further, the main task layer, one of the three task layers, aggregates features from the feature extractors. To classify the risk of osteoporosis, we define a task loss using multi-class and ordinal classification. We also use a dissimilarity loss that leverages the feature extractors' outputs and enables them to capture different characteristics from each view.

For the experiments, we conduct quantitative and qualitative evaluations of our MVCTNet using the dataset.

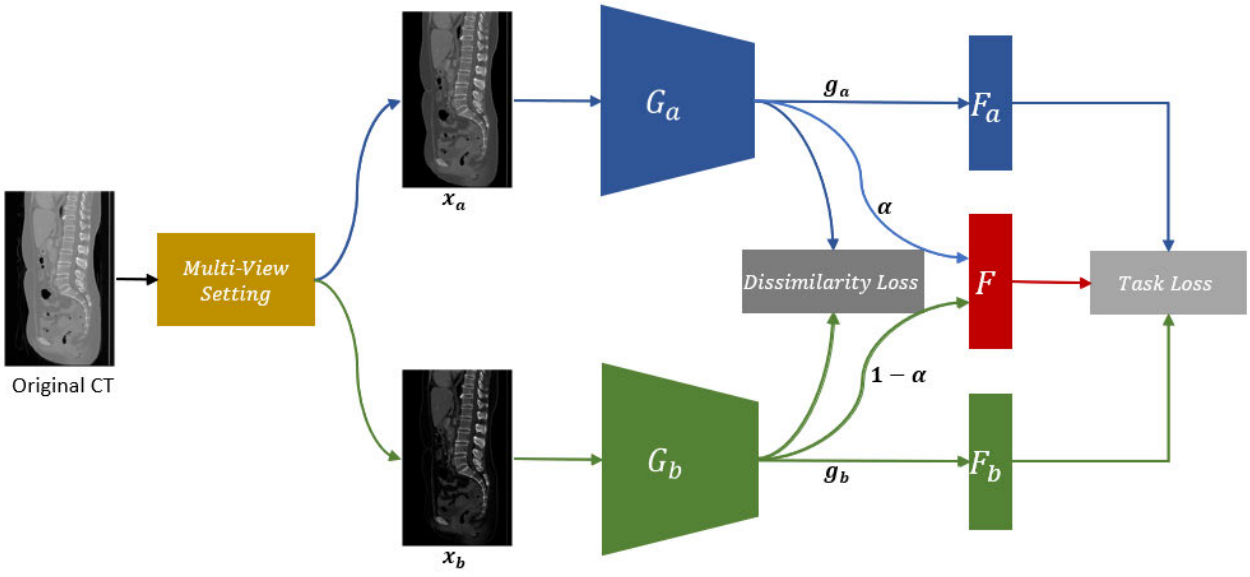

FIGURE 2. Overview of the proposed Multi-View CT Ensemble Network (MVCTNet). MVCTNet consists of feature extractors G_a, G_b and three task layers F_a, F_b, F. The task loss focuses on the target task and the dissimilarity loss minimizes the similarity of representations between the two feature extractors.

The experimental results show that our proposed method improves the automatic osteoporosis classification using intact CT images without manual work. In addition, we provide ablation studies and confirm the effectiveness of the modules in improving the classification performance of the model.

As a summary, our major contributions in this work, which extends from our previous work [16], are stated below:

- We present an osteoporosis classification task based on a sagittal CT image of the abdomen that can aid in early treatment and cost reduction.
- We propose a simple but powerful framework called MVCTNet that leverages different views from CT images and does not require manual processing.
- We demonstrate the effectiveness of the MVCTNet in the osteoporosis classification task by comparing it with strong baseline models, which achieved remarkable progress in image recognition.

II. RELATED WORK

A. DEEP LEARNING-BASED OSTEOPOROSIS DIAGNOSIS

Recently, computer vision applications for medical imaging have been studied and shown significant developments. Among these studies, several studies [1], [2], [3], [4] have conducted osteoporosis diagnosis in modalities other than the DXA test, such as bone mineral density (BMD), X-ray, and CT, other than the DXA test because it provides a helpful solution for many medical professional clinicians.

Table 1 presents the investigations of previous automatic osteoporosis diagnosis studies and our MVCTNet. Nam et al. [1] attempted to predict osteoporosis or non-osteoporosis using a logistic classifier with preoperative lumbar CT ROIs drawn encapsulating only cancellous bone. Yasaka et al. [2] estimated the BMD values on axial CT images of cropped regions around the vertebrae using a four-layer CNN model. Rastegar et al. [3] segmented BMD images and extracted texture features from ROIs, including the lumbar (L1-L4). For classification, they used ensemble machine learning methods such as random forest and K-nearest neighbor. These methods assess the osteoporosis diagnoses using regression or binary classification. Zhang et al. [4] classified normal, osteopenia, and osteoporosis using a five-layer CNN model on cropped X-ray images, including spine bone regions.

All the methods mentioned above have achieved significant advances. However, they have disadvantages in that a person must extract the ROIs manually or require an additional ROI extraction algorithm. If the ROI extraction performance is not sufficiently precise, a subsequent diagnosis cannot be performed. The MVCTNet is an automatic diagnostic classification model that uses CT images through our multi-view settings without any manual procedures, except slice selection, which we discuss in Section V.

B. VIEW-BASED METHODS

In computer vision, many works [12], [13], [14], [15], [17], [18], [19], [20] have proposed view-based methods for solving advanced problems, such as self-supervised, semi-supervised learning, and medical imaging. Some studies [12], [13] introduced a contrastive learning framework for self-supervised representation learning by utilizing the similarities of different views of the same instance. Sohn et al. [15]

transformed an unlabeled image into weakly and strongly transformed images to train deep neural networks using the weakly transformed images' pseudo-labels for semi-supervised learning.

Inspired by advances in computer vision, medical image analysis studies [17], [18], [19], [20] proposed medical domain-specific methods that adjust to their medical domain attributes. References [17] and [18] attempted to integrate contrastive learning and volumetric medical domain-specific knowledge for volumetric medical segmentation. Li et al. [19] proposed a self-learning scheme that leverages multi-style and multi-view to reduce the vendor-style domain gap for mammography detection. Yang et al. [20] designed a hybrid representation learning which use domain-specific knowledge in histopathological images. These methods are based on the idea of [12], [13] that different views from the same instance should train neural networks to have comparable features. Our MVCTNet method uses not only domain-specific knowledge, but also a different approach to leverage views. MVCTNet consists of two feature extractors that learn different features from views using a dissimilarity loss. Unlike contrastive loss, which maximizes the similarity of views from the same instance for representation learning, our dissimilarity loss minimizes it for supervision. MVCTNet's three task layers leverage different features to adjust for osteoporosis classification.

III. PROPOSED METHODOLOGY

A. MULTI-VIEW SETTINGS IN CT

In the medical field, professional clinicians diagnose diseases using CT by monitoring a specific HU range. This allows them to focus more on areas such as organs or bones they want to see. However, this may provide neural networks with insufficient information, which can degrade their performance. To address this issue, we provide more information to the neural network using two specific HU ranges. Given the labeled CT scan $\{ct, y\}$, we clamp the CT scans HU within specific ranges according to the different settings. Therefore, we can obtain two CT images with different texture information from each CT setting. Subsequently, we use the images as input for the MVCTNet, which might be seen as the different views for the same instance. In this work, we utilize two different CT settings to obtain two images $\{x_a, y\}$ and $\{x_b, y\}$ from the CT scan $\{ct, y\}$ using the following equation,

$$x_n(h,w) = \begin{cases} \mathrm{HU}_n^{\min} & \text{if } \mathrm{HU}(h,w) < \mathrm{HU}_n^{\min} \\ \mathrm{HU}_n^{\max} & \text{if } \mathrm{HU}(h,w) > \mathrm{HU}_n^{\max} \\ \mathrm{HU}(h,w), \end{cases} \quad (1)$$

$$x_n(h,w) = \frac{x_n(h,w) - \mathrm{HU}_n^{\min}}{\mathrm{HU}_n^{\max} - \mathrm{HU}_n^{\min}}, \quad (2)$$

where $\mathrm{HU}(h,w)$ denotes (h,w) the HU value of location (h,w), and HU_n^{min} and HU_n^{max} denote the minimum and maximum HU values to clamp the HU scan according to $n \in \{a,b\}$. We empirically fix $\mathrm{HU}_a^{\max}$ and $\mathrm{HU}_b^{\max}$ to 1050. We set $\mathrm{HU}_a^{\min}$ to -128 and $\mathrm{HU}_b^{\min}$ to -450 for our MVCTNet. We also compare our method with a method using the original CT scan, clamped by the entire range of the HU, in Section 4.

B. MULTI-VIEW CT NETWORK

Fig. 2 shows our MVCTNet architecture, which consists of two feature extractors (G_a and G_b) and three task layers (F_a, F_b, and F). The feature extractors G_a and G_b learn different features of CT through x_a and x_b. The task layers learn the osteoporosis classification task by using high-level representations and their combinations.

We train the feature extractors of MVCTNet both in performing the osteoporosis classification task and capture the characteristics of different views. For this, we extract representations $g_a = G_a(x_a)$ and $g_b = G_b(x_b)$, which are the outputs of the global average pooling (GAP) layer of the features extractors, which are the inputs of the task layers and our dissimilarity loss. We employ ResNet-18 [21] and EfficientNet-b0 [22] pre-trained on ImageNet dataset as the feature extractors (G_a, G_b) and the two task layers (F_a, F_b) in our experiments.

For the task layers, which consist of one MLP layer, we train two of them (F_a, F_b) using g_a and g_b, respectively, to fit in the osteoporosis classification task. We also train the other one of them (F) through an output of aggregation between g_a and g_b to leverage their differences. To aggregate the representations, we add them using the hyper-parameter α, which could be a fixed or learnable parameter. The aggregation operation is defined as follows:

$$g_{ab} = \alpha g_a + (1-\alpha) g_b. \quad (3)$$

The main task layer F predicts the class probabilities using the combined features g_{ab} obtained as the input, which can be obtained by using (3). Because we design F_a and F_b to train the feature extractors according to the task, the task layers are used only for training. Then, we discard them in the inference phase using only the main task layer F for prediction. To choose the value of α, we conduct an experiment on α as the fixed value and learnable parameter, as described in Section IV-C.

C. LOSS FUNCTION

Our goal is to simultaneously capture different characteristics in multiple CT views while performing osteoporosis classification. For this purpose, we use task loss and dissimilarity loss, the former optimizes our model to perform the task, and the latter minimizes the similarity between two representations (g_a, g_b) from MVCTNet's feature extractors. We calculate the task loss using the sum of the loss function of the task layers. Because there is a sequential relationship between normality, osteopenia, and osteoporosis, we employ ordinal regression, not only multi-class classification for the osteoporosis classification task. The task loss function is

Algorithm 1 MVCTNet Training Procedure

Input : Feature extractors G_a and G_b, task layers F_a, F_b and F, total epochs E, mini batch B, CT images x, and trade-off hyper-parameter λ.

for $i = 1$ **to** E **do**
  for $j = 1$ **to** B **do**
    obtain x_a and x_b using (1) and (2);
    $g_a = G_a(x_a)$;
    $g_b = G_b(x_b)$;
    obtain g_{ab} using (3);
    $\hat{y}_a = F_a(g_a)$;
    $\hat{y}_b = F_b(g_b)$;
    $\hat{y} = F(g_{ab})$;
    define $\mathcal{L}_{\text{task}}(\hat{y}_a, \hat{y}_b, \hat{y}; G_a, G_b, F_a, F_b, F)$ using (4);
    define $\mathcal{L}_{\text{dis}}(g_a, g_b; G_a, G_b)$ using (6);
    $\mathcal{L} = \mathcal{L}_{\text{task}}(\hat{y}_a, \hat{y}_b, \hat{y}; G_a, G_b, F_a, F_b, F) + \lambda\mathcal{L}_{\text{dis}}(g_a, g_b; G_a, G_b)$;
    optimize $\mathcal{L}$;
  end
end
Output: Feature extractor G_a and G_b, task layer F.

formulated as follows:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{cls}}(F_a(g_a), y) + \mathcal{L}_{\text{cls}}(F_b(g_b), y) + \mathcal{L}_{\text{cls}}(F(g_{ab}), y), \tag{4}$$

where $\mathcal{L}_{\text{cls}}(\cdot, \cdot)$ denotes the standard cross-entropy loss (CE) or ordinal regression loss (OR) [23]. We use the ordinal regression loss which has usually been studied in age estimation [23], [24], [25]. Following [23], we use ordinal regression loss as cross-entropy of $K-1$ binary classifiers:

$$\mathcal{L}_{\text{or}} = -\sum_{k=1}^{K-1} y^k \log \hat{y}^k + (1 - y^k) \log(1 - \hat{y}^k), \tag{5}$$

where $\hat{y}$ denotes the sigmoid function output of the task layers and K is the number of classes. We also experiment with both the standard multi-class and the ordinal regression for the classification task in Section IV-B.

By optimizing the task loss, feature extractors can train osteoporosis diagnostic tasks. Through the task loss, the representations from the two feature extractors can converge to similar properties. As a result, the MVCTNet can not utilize the different characteristics from two perspectives. To address this issue, we use dissimilarity loss, which minimizes the cosine similarity between the representations g_a and g_b,

$$\mathcal{L}_{\text{dis}} = \frac{g_a^T g_b}{\|g_a\| \, \|g_b\|}. \tag{6}$$

By optimizing the dissimilarity loss, we can constraint the feature extractors from having similar characteristics from different perspectives. In summary, by integrating task loss and dissimilarity losses, the overall loss in our framework can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{dis}}, \tag{7}$$

where $\lambda > 0$ denotes the trade-off hyper-parameter.

Algorithm 1 summarizes the overall training procedure of our MVCTNet. Note that while all models, including feature extractors and task layers, are updated using the task loss, the dissimilarity loss only optimizes feature extractors.

TABLE 2. Patient baseline characteristics.

	Train dataset	Test dataset
Men / women (n)	505 / 1,778	87 / 513
Mean age (years)	58.16 (± 21.82)	58.61 (± 17.92)
Mean body weight (kg)	59.50 (± 12.17)	58.26 (± 11.50)
Mean body height (cm)	157 (± 9)	157 (± 8)
Mean body mass index (kg / m^2)	58.16 (± 21.82)	58.61 (± 17.92)

D. IMPLEMENTATION DETAILS

We implement all networks and experimental settings in PyTorch framework. The entire network is trained on a single NVIDIA RTX 3090 GPU, and all network parameters are optimized by Adam optimization algorithm in an end-to-end manner. All inputs of the models are uniformly resized to 448 × 224, and a random horizontal flip is applied with a probability of 0.5 for training. The feature extractors are trained using a learning rate of $1e-4$, which is ten times lower than that of the task layers that consist of a single MLP layer. We split the training dataset into training (80%) and validation (20%) datasets, and all hyper-parameters were tuned in the validation phase. The entire network is trained for 60 epochs with a batch size of 32, and the trade-off hyper-parameter λ is set to 0.1. For the parameter α, is used as a fixed value or learnable parameter. When used as a learnable parameter, it is initialized to 0.5 and then updated. We initialize the feature extractor's weights with pre-trained weights on ImageNet dataset for all experiments and randomly initialize the task layers.

E. DATASET COLLECTION AND ANNOTATION

We enrolled patients who underwent a contrast-enhanced abdominal CT and dual-energy X-ray absorptiometry (DXA) test between January 2015 and October 21, 2015. A total of 2,883 images were collected from 2,883 de-identified patients, 592 males and 2,291 females aged ≥ 20 years, and randomly divided into 2,283 images for training and 600 images for testing. We also divided the training datasets into training and validation datasets in a ratio of 8:2. Images were labeled considering the three stages as normal (T-score ≥ −1.0), osteopenia (−2.5 < T-score < −1.0), and osteoporosis (T-score ≤ −2.5) according to the World Health

Organization criteria. Particularly, the patients in their 20s belonged to the normal group, but did not undergo DXA testing; thus, they underwent additional verification by the radiology professor.

During the collection process, the patients with visible surgical or bone cement and those without CT multi-planar reformation forming the sagittal axis were excluded. For slice selection, radiology experts label each patient's sagittal slice image, including all vertebrae, according to its osteoporosis risk. Then we use all the slices determined by the experts.

This study was conducted and approved in accordance with the relevant guidelines and regulations of Kangwon National University Hospital IRB (Approval No. KNUH-A-2021-03-020-002). Patient consent was not required because the data were de-identified.

IV. EXPERIMENTS

A. EVALUATION METRICS

In this study, we employ five metrics for quantitative evaluation, including the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, F1-score, and mean absolute error (MAE). The first four metrics are used to evaluate the classification performances. The definitions of the sensitivity, specificity, and F1-score are formulated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{9}$$

$$\text{F1} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}, \tag{10}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative respectively.

We also employ MAE metric to evaluate the ordinal regression performance and MAE is formulated as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{11}$$

where N denotes the total number of samples, and $\hat{y}_i$ and y_i denote the predictions and labels of the samples, respectively.

B. COMPARISONS WITH BASELINE METHODS

To evaluate our proposed method, we employ two strong baseline methods: ResNet-18 [21] and EfficientNet-b0 [22], both of which have high performance in image recognition and medical applications. In addition, we compare them with the performances of the MVCTNet, using two baseline methods as the feature extractors, on our dataset. We also investigate the baseline methods and MVCTNet with CE and OR loss functions. Note that we use only x_b, G_b and F_b as the baseline methods in this experiment. Empirically, we found that x_b setting shows better results, and we investigate original CT and x_a settings in Section. IV-C. Because comparisons of AUC evaluation in the ordinal regression task are complicated, we estimate it using the average AUC of osteopenia and osteoporosis. As shown in Table 3, we observe that methods with OR loss show better performance than methods with CE loss, regardless of baseline and metrics. From these observations, we can confirm that ordinal regression is more suitable for osteoporosis diagnosis. We also observe that the MVCTNet outperforms all other baseline methods in all metrics and all task losses. The MVCTNet achieves the best performance on five evaluation metrics with sensitivity of 81.33%, specificity of 90.67%, F1-score of 81.24%, AUC of 0.9640, and MAE of 0.1900 in the ordinal regression task. In particular, in the multi-class classification task, the MVCTNet with ResNet-18 exhibit the most improved performance, achieving a sensitivity improvement of 4.66%. From these results, we can confirm that our proposed method is effective for the osteoporosis classification.

C. ABLATION STUDIES

For a more detailed analysis of the proposed method, we conduct ablation studies on an ordinal regression task.

1) ANALYSIS OF VIEWS

To evaluate the effectiveness of our multi-view settings, we conduct experiments with various view settings, such as original CT, x_a, and x_b as the input of the baseline methods and the MVCTNet. As shown in Fig. 4, we first observe that the baseline methods with original CT (ct) have lower performance than other baseline methods with x_a or x_b. We probably infer that using the whole range of HU may provide neural networks with unnecessary information, such as the HU value, which does not require finding the spine in CT. We also observe that the baseline with x_b outperforms the baseline with x_a. It might be reasonable when considering that x_b is more similar than x_a to the bone setting, which professional clinicians use for diagnosis.

To compare different views and the same views in the MVCTNet, we use x_a (or x_b) as inputs of MVCTNet's two feature extractors. We can observe that the different perspective method outperforms the same perspective methods in all metrics on the ordinal regression task. From this observation, we can confirm that the performance does not increase simply as the number of parameters increases, but it uses different views.

2) EFFECTIVENESS OF THE MVCTNet COMPONENTS

We conduct an ablation study to investigate the effectiveness of components of our MVCTNet using the ResNet-18 as feature extractors. Fig. 3 shows the performance of the vanilla ResNet-18 with x_b, MVCTNet without dissimilarity loss, which uses only the task loss, and MVCTNet with dissimilarity loss on sagittal CT scan osteoporotic ordinal regression tasks. We can observe that the MVCTNet without the dissimilarity loss outperforms the ResNet-18 with a

TABLE 3. Performance comparison of baseline methods and their MVCTNet combinations on a sagittal CT scan osteoporotic classification dataset. If MVCTNet is checked, baseline methods are used as the feature extractors of MVCTNet. If MVCTNet is unchecked, pure baseline networks are used. MVCTNet consistently improves performance when combined with baseline methods, regardless of categorical (CE) and ordinal (OR) loss.

Baseline	Task Loss	MVCTNet	Sensitivity	Specificity	F1-score	AUC	MAE
ResNet-18	CE		0.7517	0.8758	0.7544	0.8971	0.2517
	CE	✓	0.7983	0.8992	0.7959	0.9279	0.2100
	OR		0.7767	0.8883	0.7712	0.9467	0.2300
	OR	✓	**0.8133**	**0.9067**	**0.8124**	**0.9640**	**0.1900**
EfficientNet-b0	CE		0.7550	0.8775	0.7497	0.9045	0.2467
	CE	✓	0.8033	0.9017	0.7811	0.9290	0.2017
	OR		0.7817	0.8908	**0.8027**	0.9505	0.2233
	OR	✓	**0.8050**	**0.9025**	**0.8027**	**0.9620**	**0.1950**

TABLE 4. Ablations on Multi-view setting. Note that input ct indicates the original CT scan. MVCTNet uses ResNet-18 as its features extractors. All experiments are conducted on the ordinal regression task.

Method	Input	Sensitivity	Specificity	F1-score	AUC	MAE
ResNet-18	ct	0.7217	0.8608	0.7247	0.9482	0.2800
	x_a	0.7650	0.8825	0.7634	0.9502	0.2383
	x_b	0.7767	0.8883	0.7712	0.9467	0.2300
MVCTNet	(x_a, x_a)	0.7700	0.8850	0.7620	0.9580	0.2383
	(x_b, x_b)	0.7933	0.8966	0.7884	0.9561	0.2117
	(x_a, x_b)	**0.8133**	**0.9067**	**0.8124**	**0.9640**	**0.1900**

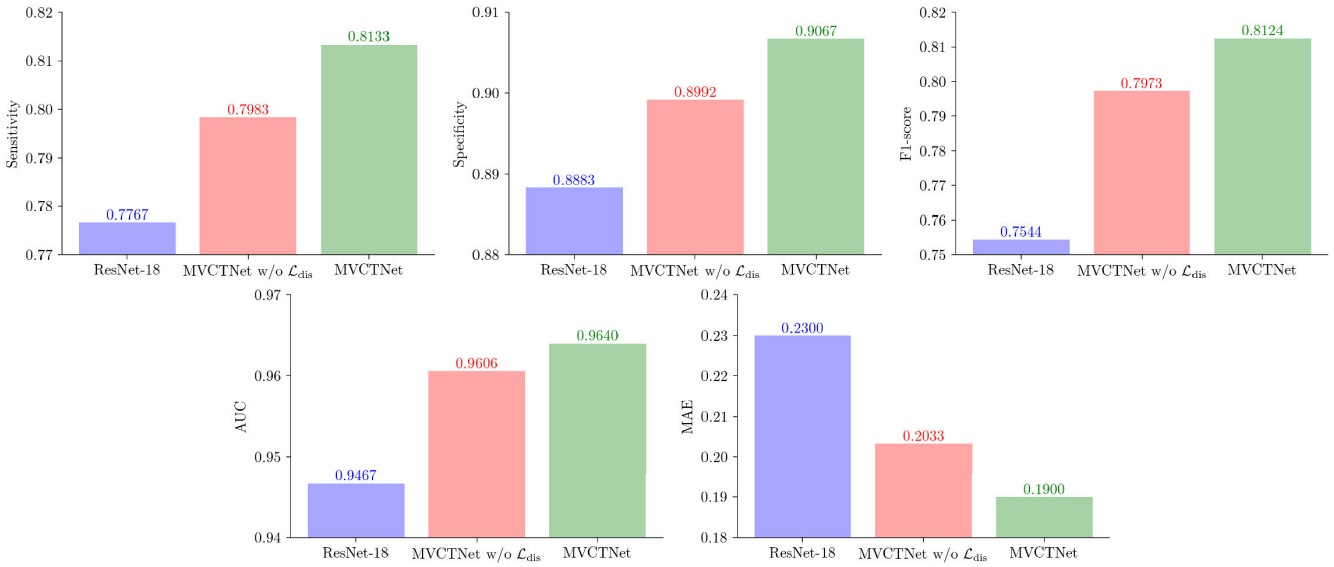

FIGURE 3. Ablations on components of MVCTNet, which uses ResNet-18 as its feature extractors. All experiments are conducted on the ordinal regression task.

performance of 79.73% in F1-score, which improves the baseline by 3.87%. When combined with dissimilarity loss, the MVCTNet exhibit better performance in all metrics. From these observations, we can confirm that all components of MVCTNet are effective in improving performance.

3) ANALYSIS OF α

Table 5 compares the performance of the cases where α is fixed to a specific value and when it is a learnable parameter in operation for the aggregation of the features through the two feature extractors, which do not require a

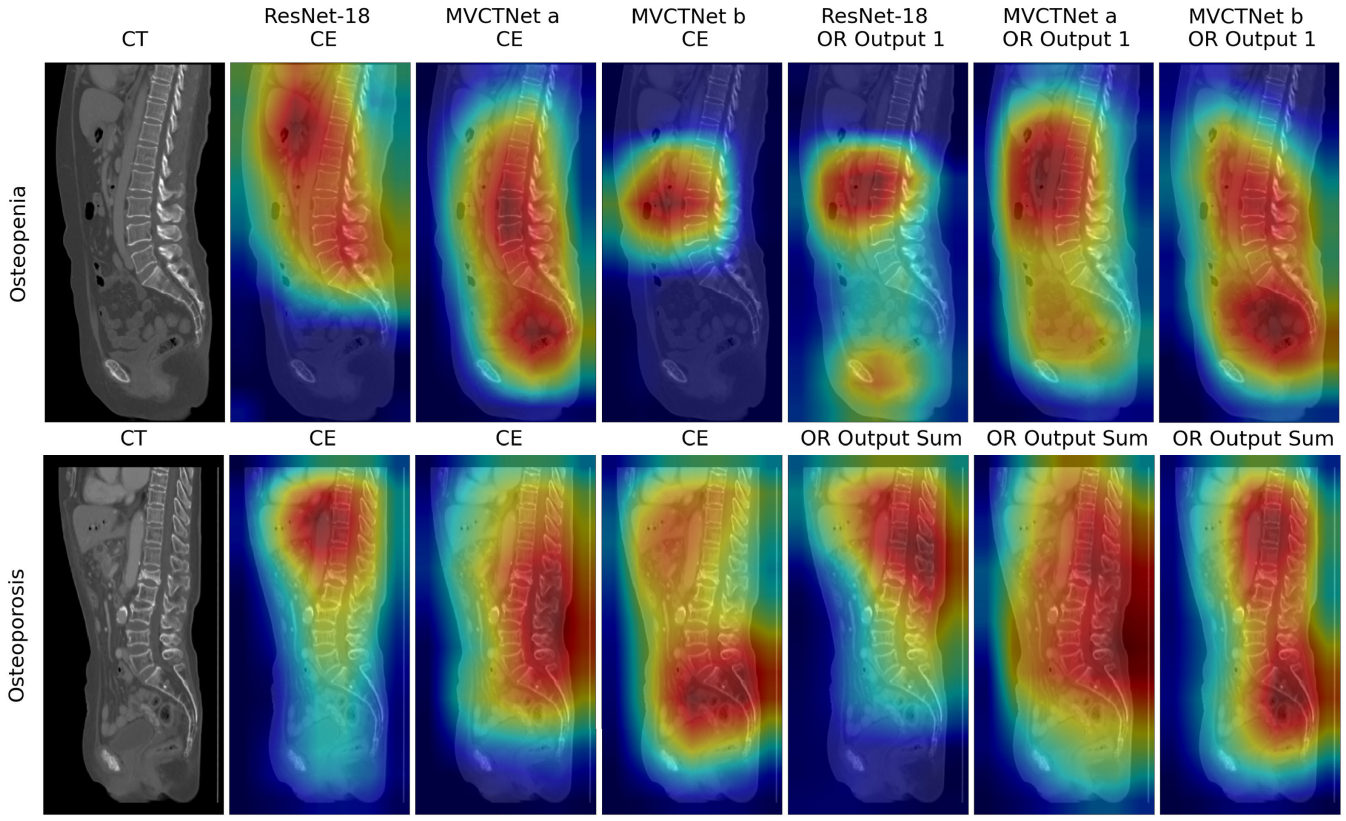

FIGURE 4. Visualization results with Grad-CAM. We compare Grad-CAM results between ResNet-18 and MVCTNet in osteopenia and osteoporosis according to multi-class classification(CE) and ordinal regression (OR). For OR, we visualize the osteopenia class (OR Output 1) on the first node of the classifier. And we operate the sum of gradients of the first and second nodes (OR Output Sum) of the classifier for visualization of osteoporosis.

TABLE 5. Ablations on α. MVCTNet uses ResNet-18 as its features extractors. All experiments are conducted on the ordinal regression task.

α	Sensitivity	Specificity	F1-score	AUC	MAE
0.1	0.7850	0.8925	0.7793	0.9605	0.2266
0.3	0.8017	0.9008	0.7976	0.9635	0.2066
0.5	0.8100	0.9050	0.8078	0.9639	0.1933
0.7	**0.8133**	**0.9067**	**0.8132**	0.9613	**0.1900**
0.9	0.8067	0.9033	0.8071	0.9556	0.1966
learnable	**0.8133**	**0.9067**	0.8124	**0.9640**	**0.1900**

hyper-parameter search to find its optimal value. The method with $\alpha = 0.7$ outperforms other fixed α methods, with the highest F1-score of 81.32%, which is 0.08% better than the learnable parameter. Furthermore, the result of learnable α shows similar or better performance when compared to $\alpha = 0.7$. In particular, it achieves the highest performance in AUC, which is approximately 0.27% better than the that of method with $\alpha = 0.7$. From these results, it is the best choice to set α as 0.7 or a learnable parameter. In all experiments, including comparisons with baseline methods and other ablation studies, we used α as the learnable parameter.

D. VISUALIZATION WITH GRAD-CAM

Fig. 4 shows the spatial importance visualization of networks for osteopenia and osteoporosis samples in the test dataset using Grad-CAM [26]. We investigate the ResNet-18 with x_b and MVCTNet with ResNet-18 feature extractors using CE and OR. Particularly, we visualize the Grad-CAM results of the MVCTNet's networks including G_a, F_a (MVCTNet a) and G_b, F_b (MVCTNet b) obtained by our multi-view setting. The Grad-CAM results of the ResNet-18, MVCTNet a, and MVCTNet b are obtained by the operation of the corresponding class gradients for the multi-class task. For the ordinal regression, we visualize the results of the osteopenia class via the same operation in the multi-class. Particularly, in the case of the osteoporosis class, we add the gradients of all nodes in the classifier (OR Output Sum) of each model.

From the visualization results, we can observe that the Grad-CAM masks of our MVCTNet, except MVCTNet b in osteopenia and CE, cover a more extensive target region, including the spine, which is important for an osteoporosis diagnosis than ResNet-18. Furthermore, the MVCTNet a and MVCTNet b results have complementary visualization masks. From these observations, we confirm that the two feature extractors complement each other's viewpoints, as intended.

V. DISCUSSION AND CONCLUSION

In this work, we have presented an osteoporosis diagnosis task in sagittal CT scans and proposed a simple but powerful architecture, MVCTNet, for this task. We collected a large dataset for the task, involving 2,883 patients who underwent contrast-enhanced abdominal CT. Additionally, we performed the task using the MVCTNet without manual processes, which were previously used for automatic osteoporosis diagnosis, as presented in Table 1. Quantitative and qualitative experiments demonstrated that the MVCTNet consistently outperformed other baseline methods in multi-class classification and ordinal regression tasks. In addition, from the ablation studies, we confirmed that the components of the MVCTNet are effective for the osteoporosis diagnosis. Our MVCTNet has advantages; however, an actual medical scenario may involve additional challenges, such as the selection of a CT slice. In future work, we will offer an extensive version of our model to overcome these challenges such as a 3D medical image model.

ACKNOWLEDGMENT

(Dong Hwan Hwang and So Hyeon Bak are co-first authors.)

REFERENCES

[1] K. H. Nam, I. Seo, D. H. Kim, J. I. Lee, B. K. Choi, and I. H. Han, "Machine learning model to predict osteoporotic spine with Hounsfield units on lumbar computed tomography," *J. Korean Neurosurgical Soc.*, vol. 62, no. 4, pp. 442–449, Jul. 2019.

[2] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe, "Prediction of bone mineral density from computed tomography: Application of deep learning with a convolutional neural network," *Eur. Radiol.*, vol. 30, pp. 3549–3557, Jun. 2020.

[3] S. Rastegar, M. Vaziri, Y. Qasempour, M. R. Akhash, N. Abdalvand, I. Shiri, H. Abdollahi, and H. Zaidi, "Radiomics for classification of bone mineral loss: A machine learning study," *Diagnostic Interventional Imag.*, vol. 101, no. 9, pp. 599–610, Sep. 2020.

[4] B. Zhang, K. Yu, Z. Ning, K. Wang, Y. Dong, X. Liu, S. Liu, J. Wang, C. Zhu, Q. Yu, and Y. Duan, "Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study," *Bone*, vol. 140, Nov. 2020, Art. no. 115561.

[5] I. M. Wani and S. Arora, "Computer-aided diagnosis systems for osteoporosis detection: A comprehensive survey," *Med. Biol. Eng. Comput.*, vol. 58, pp. 1873–1917, Sep. 2020.

[6] A. D. Smith, "Screening of bone density at CT: An overlooked opportunity," *Radiology*, vol. 291, no. 2, pp. 368–369, May 2019.

[7] T. Sözen, L. Özişik, and N. C. C. Başaran, "An overview and management of osteoporosis," *Eur. J. Rheumatology*, vol. 4, no. 1, p. 46, 2017.

[8] J. Wu, Y. Qu, K. Wang, and Y. Chen, "Healthcare resource utilization and direct medical costs for patients with osteoporotic fractures in China," *Value Health Regional Issues*, vol. 18, pp. 106–111, May 2019.

[9] P. J. Mitchell, "Fracture liaison services: The UK experience," *Osteoporosis Int.*, vol. 22, no. 3, pp. 487–494, Aug. 2011.

[10] P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes, "Classification of osteoporotic vertebral fractures using shape and appearance modelling," in *Proc. Int. Workshop Comput. Methods Clin. Appl. Musculoskeletal Imag.* Cham, Switzerland: Springer, 2017, pp. 133–147.

[11] N. Yamamoto, S. Sukegawa, A. Kitamura, R. Goto, T. Noda, K. Nakano, K. Takabatake, H. Kawai, H. Nagatsuka, K. Kawasaki, Y. Furuki, and T. Ozaki, "Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates," *Biomolecules*, vol. 10, no. 11, p. 1534, Nov. 2020.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, Jul. 2020, pp. 1597–1607.

[14] J.-B. Grill, F. Strub, F. Altche, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, and B. Piot, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[15] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.

[16] S.-H. Choi, D.-H. Hwang, D.-H. Kim, S.-H. Bak, and Y. Kim, "Efficient osteoporosis prediction using a pair of ensemble models," *J. Korea Soc. Comput. Inf.*, vol. 26, no. 12, pp. 45–52, 2021.

[17] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

[18] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi, "Positional contrastive learning for volumetric medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 221–230.

[19] Z. Li, Z. Cui, S. Wang, Y. Qi, X. Ouyang, Q. Chen, Y. Yang, Z. Xue, D. Shen, and J.-Z. Cheng, "Domain generalization for mammography detection via multi-style and multi-view contrastive learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 98–108.

[20] P. Yang, Z. Hong, X. Yin, C. Zhu, and R. Jiang, "Self-supervised visual representation learning for histopathological images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 47–57.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[23] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

[24] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020.

[25] H. Zhu, H. Shan, Y. Zhang, L. Che, X. Xu, J. Zhang, J. Shi, and F.-Y. Wang, "Convolutional ordinal regression forest for image ordinal estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 4084–4095, Aug. 2022.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

DONG HWAN HWANG received the B.S. and M.S. degrees in computer science and engineering from Kangwon National University, South Korea, in 2019 and 2021, respectively. He is currently a Researcher with Ziovision. His research interests include deep learning and computer vision.

SO HYEON BAK received the B.S. degree in biological science and the M.S. and Ph.D. degrees in medicine from Konkuk University, in 2004, 2009, and 2017, respectively. From February 2015 to February 2022, she was with Kangwon National University Hospital. Since March 2022, she has been with the Department of Radiology, Asan Medical Center, as a Clinical Assistant Professor.

TAE-JUN HA received the B.S. degree in computer science and the M.S. degree in information systems engineering from Hansung University, in 2017 and 2019, respectively. Since 2021, he has been a Researcher with Kangwon National University Hospital. His research interests include computer vision and parallel programming.

YOON KIM received the B.S., M.S., and Ph.D. degrees in electronic engineering from Korea University, in 1993, 1995, and 2003, respectively. In 2004, he joined the Department of Computer Engineering, Kangwon National University, where he is currently a Professor. Since 2016, he has been with Ziovision, Chuncheon, Gangwon, South Korea. His research interests include deep learning and computer vision.

WOO JIN KIM received the B.S. and M.S. degrees from the Medical College, Seoul National University, in 1994 and 2004, respectively, and the Ph.D. degree in medicine from Hallym University, in 2006. In 2004, he joined the Department of Internal Medicine, Kangwon National University, where he is currently a Professor. His research interest includes biomedical informatics.

HYUN-SOO CHOI received the B.S. degree in computer and communication engineering majored in brain and cognitive science from Korea University, in 2013, and the integrated M.S./Ph.D. degree in electrical and computer engineering from Seoul National University, South Korea, in 2020. From 2020 to 2021, he was a Senior Researcher with Vision AI Laboratories, SK Telecom. From 2021 to 2023, he was with Kangwon National University, South Korea, as an Assistant Professor. Since 2021, he has been a Chief Technical Officer with Ziovision. Since 2023, he has been an Assistant Professor with the Department of Computer Science and Technology, Seoul National University of Science and Technology.