공학석사 학위논문

의료 도메인에서 귀납적 전이 학습을 위한 효과적인 표현 학습 방법

Effective Representation Learning Methods for Inductive Transfer Learning in Medical Domain

울 산 대 학 교 대 학 원
의 과 학 과
경 성 구

귀납적 전이 학습을 사용하는 의료 도메인에 대한 효과적인 표현 학습 방법

지 도 교 수 김 남 국

이 논문을 공학석사 학위 논문으로 제출함

2022 년 08 월

울 산 대 학 교 대 학 원
의 과 학 과
경 성 구

경성구의 공학석사학위 논문을 인준함

심사위원 홍 길 선 (인)

심사위원 김 남 국 (인)

심사위원 이 준 구 (인)

울 산 대 학 교 대 학 원

2022 년 08 월

감사의 글

의료인공지능에 대한 흥미를 갖고 MI2RL 에서 석사과정을 시작한지 어느덧 2년 남짓이 되었습니다. 저에게 있어 희로애락(喜怒哀樂)이 공존했던 연구실 생활의 지식과 경험을 잊지 않고 더 나아가 끊임없이 갈망하고 발전하는 의공학도가 되겠습니다. 무엇보다 부족한 저에게 날카로운 비평과 진심 어린 조언 및 응원으로 많은 가르침을 주신 김남국 교수님, 좋은 연구를 위해 애써 주신 홍길선 교수님, 연구실에 처음 들어왔을 때 사수로서 길잡이가 되어 주신 신기원 선생님께, 항상 긍정적인 마음가짐으로 연구실의 감초(甘草) 같은 역할을 해주시는 류승민 선생님께 진심으로 고개 숙여 감사의 인사를 드립니다.

부족한 저를 믿고 같이 연구에 참여하여 노력해주신 박주영, 조경진, 장미소, 박승주, 김기덕, 정현수 선생님, 사수로서 부족한 점이 많지만 잘 따라와준 원종준, 최창용 선생님, 동기로서 석사 과정을 같이 밟으며 힘이 되어 주신 박현정, 오홍민 선생님께 감사의 인사를 전합니다.

제가 좋은 결과를 얻을 수 있었던 것은 좋은 선배님들과 후배님들의 도움이 컸습니다. 특히, 장령우, 김성철, 김인환, 조성만 선배님들의 선도(善導)를 통해 저를 다잡을 수 있었고, 임지섭, 박강길, 제갈성규, 김지영, 이소영 후배님들의 열정(熱情)을 보며 초심(初心)을 되찾을 수 있었습니다. 연구실 생활동안 서로에게 긍정적인 영향을 주고받으며 보다 알찬 석사 과정을 보낼 수 있게 해준 문소진, 김민지, 김민경 선생님께도 감사드립니다. 바쁘신 와중에도 졸업 심사위원 맡아 주신 이준구 교수님께 진심으로 감사의 인사를 드립니다.

마지막으로 항상 저를 믿고 무한한 응원과 지지를 해 주신 아버지(경규수), 어머니(김태숙), 사랑하는 원구 형, 그리고 6년이라는 오랜 시간을 넘어 평생 서로의 옆을 든든하게 지켜줄 나의 동반자 지수 에게 감사의 인사와 사랑한다는 말을 전합니다.

Abstract

The deep learning technique has been used in a wide range of fields, with impressive results. However, lack of training data, performance degradation due to modality or domain differences, higher-definition images like radiographs and computed tomography (CT) scans, and the robustness of other medical centers, etc., there are still difficulties in applying deep learning in the medical domain. To address these issues, inductive transfer learning of representation learning, a study that skillfully utilizes the features derived from the network, has been intensively researched: *sequential transfer learning* and *multi-task learning*. In this study, three experiments have been performed to confirm how representation learning using inductive transfer learning affects medical domains: 'Application of deep representation on pediatric diagnosis', 'Application of deep representation on brain hemorrhage diagnosis', and 'Application of deep representation on low-dose CT denoising task'. In the first study, sequential transfer learning was applied for performance improvement. We constructed class-balanced pediatric radiographs datasets, PedXnets using labels based on radiographic views, and developed their supervised representations. We validated the effects of the representation learning through pediatric downstream tasks including fracture classification and bone age assessment. As a result, the transfer learning from Model-PedXnets showed improved quantitative performances compared to those of the Model-Baseline. Model-PedXnets had equivalent and in some cases even improved performance than Model-ImageNet. In particular, Model-PedXnets focused on the most meaningful regions. In the second study, multi-task learning was applied for robustness. We proposed a supervised multi-task aiding representation transfer learning network (SMART-Net) for the diagnosis of intracranial hemorrhage (ICH). The proposed framework consists of upstream and downstream components. In the upstream, a weight-shared encoder of the model is trained as a robust feature extractor that captures global features by performing slice-level multi-pretext tasks. In the downstream, the transfer learning was conducted with a pre-trained encoder and 3D operator for volume-level tasks. Experimental results based on four test sets indicate that SMART-Net has better robustness and performance in terms of volume-level ICH classification and segmentation over previous methods. In the third study, multi-task learning was applied for the stabilization of discriminator learning. We propose a multi-task

discriminator based generative adversarial network (MTD-GAN) simultaneously conducting three vision tasks (classification, segmentation, and reconstruction) in a discriminator. To stabilize GAN training, we introduce two novel loss functions termed non-difference suppression (NDS) loss and reconstruction consistency (RC) loss. Furthermore, we take a fast Fourier transform with convolution block (FFT-Conv Block) in the generator to make use of both high- and low-frequency features. Our model has been evaluated by pixel-space and feature-space based metrics in the head and neck LDCT denoising task, and results show outperformance quantitatively and qualitatively than the state-of-the-art denoising methods. All three studies confirmed that representation learning, which includes sequential transfer learning and multi-task learning, could enhance performance, extract semantic information, and make models robust to external data in medical domains. Instead of simply evaluating the performance of models by training scratch models, representation learning should be included in the future application of artificial intelligence to medical domains.

Contents

Abstract ····· i
Contents ····· iii
Contents of Tables ····· iv
Contents of Figures ····· v
Introduction ····· 1
A. Application of deep representation learning to pediatric diagnosis ····· 3
1. Background and Objective ····· 3
2. Materials and Methods ····· 4
3. Experiments and Results ····· 8
4. Discussion ····· 13
5. Summary ····· 14
B. Application of deep representation learning to brain hemorrhage diagnosis ·· 15
1. Background and Objective ····· 15
2. Materials and Methods ····· 17
3. Experiments and Results ····· 21
4. Discussion ····· 29
5. Summary ····· 30
C. Application of deep representation learning to low-dose CT denoising task ·· 31
1. Background and Objective ····· 31
2. Materials and Methods ····· 32
3. Experiments and Results ····· 36
4. Discussion ····· 39
5. Summary ····· 40
Conclusion ····· 40
References ····· 41
Abstract (with Korean) ····· 45

Contents of Tables

Table 1. The performance comparisons of radiographic views recognition task as an upstream task 10
Table 2. The performance comparisons of the fracture classification task 11
Table 3. The performance comparisons of bone age assessment 12
Table 4. Patient demographic information and medical properties in four ICH datasets..... 18
Table 5. Comparisons of up and downstream performance according to consistency loss in the internal test set 25
Table 6. Quantitative results of volume-level target classification and segmentation tasks for a comparative analysis with previous methods on four test sets..... 27
Table 7. Quantitative results of volume-level target classification and segmentation tasks for a comparative analysis with ablation studies on four test sets..... 29
Table 8. Patient demographic information and medical characteristics in the LDCT denoising dataset 32
Table 9. Quantitative results for a comparative analysis with previous methods 38
Table 10. Quantitative results for a comparative analysis in ablation study 38

Contents of Figures

Figure 1. Flow chart for classifying suggested radiographic views in a real-world medical radiography dataset 5
Figure 2. Overview of the labeling procedure for radiographic images for PedXnets..... 7
Figure 3. Illustration of the Model-PedXnets method..... 8
Figure 4. Plots of Model-PedXnet model's Grad-CAM activation maps of radiographic views recognition task 10
Figure 5. Plots of Model-PedXnets' t-SNE maps of radiographic views recognition task 11
Figure 6. Comparisons of activation maps in the fracture downstream task 12
Figure 7. Comparisons of activation maps in the bone age assessment..... 13
Figure 8. Histograms of four datasets are shown according to the quantity of ICH 16
Figure 9. Overview properties of ICH patients in four independent ICH datasets 17
Figure 10. Schematic overview of the SMART-Net framework..... 19
Figure 11. Comparisons of the pre-trained encoder activation map according to multi-pretext task combinations and the previous representation learning approaches..... 24
Figure 12. Illustrations of the output mismatches caused by the target-specific multi-head structure and of the effects of the consistency loss 25
Figure 13. Comparison of volume-level segmentation results for severe and mild ICH cases, and two normal cases 28
Figure 14. Comparisons of false-positive reduction performance on the normal cases in volume-level segmentation using box plots on four test sets. 28
Figure 15. Schematic overview of the MTD-GAN framework 33
Figure 16. A concept of our NDS loss to LSGAN loss in segmentation task 35
Figure 17. The denoising results of previous methods..... 37
Figure 18. The denoising results in the ablation study..... 39

Introduction

Background

In artificial intelligence (AI) technologies, deep learning has achieved an unprecedented performance in a variety of computer vision tasks such as image classification, object detection, semantic segmentation, image reconstruction, visual question answering, etc, showing remarkable performances. Especially, representation learning is an important aspect of deep learning which from raw data automatically discovers useful feature patterns such as those that are interpretable, incorporate latent representations, or can be used for transfer learning [1]. In general, the good representation used for transfer learning has some advantages [2]:

- Reduces the variance of the test results: According to Erhan *et al.* [3], the variance of the predictions of the pre-trained model was significantly low. This affected the test results variance, which means improved reproducibility.
- Relieve the issue of insufficient training data: Transfer learning requires less training data for new deep learning models and saves development time of networks when using pre-trained representation as the majority of the model has already been learned.
- Improved performances and robust models: The entire training process is made more efficient by leveraging previously acquired knowledge. To solve a particular target task by fine-tuning the representation, researchers also can employ a variant approach considering multiple models. The sharing of knowledge between different algorithms can result in a more accurate and generalized model without overfitting.

These promising merits drew a lot of attention from the medical domain, which has a lot of issues such as a lack of training data, the high costs of processing higher-definition images like radiographs and CT scans, and performance degradation when data distribution shifts caused by image modalities, medical centers, and data collection period. However, any machine learning model's performance is highly dependent on the learned representations. According to Ma *et al.* [4], it is important to understand when and why the pre-trained representation works in a particular target task because the pre-trained representation is sometimes helpful but often harmful. To transfer better representation, inductive transfer learning that skillfully handles the features extracted from the raw data has been actively studied: *sequential transfer learning* and *multi-task learning.*

Sequential transfer learning

The objective of sequential transfer learning (STL) is to enhance learning in the target task by transferring pre-trained knowledge from the source task. The initial phase is a pre-training phase in which general representations are learned for a large scale of source tasks. The second phase involves the application of the acquired knowledge to the intended task. On a variety of natural language processing (NLP) tasks, approaches based on these techniques have achieved state-of-the-art performance. Pre-trained language representations such as word2vec, GloVe, GPT, and BERT trained on a large unlabeled text corpus are still utilized as powerful representations and are frequently employed to increase performance in NLP applications [5]. There are two types of self-supervised learning (SSL) in the vision domain [6]: Pretext learning creates representations using pseudo-labels, or labels that are automatically generated depending on the properties of the dataset. It comprises predicting the degree of rotation, filling in a missing portion of an image, coloring a grayscale image, and predicting the relative position of a patch, among other operations. Contrastive learning, including MOCO and SimCLR, learns representations by discriminating between augmented versions of images using pseudo labels and positive or negative image pairings. More recently, BYOL, which does not employ negative image pairs, has been developed.

Multi-task learning

Multi-task learning (MTL) aims to learn multiple distinct tasks simultaneously by maximizing the generalization performance of all tasks using comprehensive information. This strategy has resulted in an average performance improvement and is advantageous for activities that have similar features [7]. Nevertheless, if multi-tasks are unrelated and one group of related tasks dominates the training process, individual tasks may suffer negative transfer in which the multitask model's predictions are inferior to those of the single-task model [8]. To solve the negative transfer, the studies [9, 10] that regulated the equilibrium between multitasking were conducted beforehand. In contrast, previous studies [11-13] focused on task affinity, or which tasks should be performed together.

Objectives

In this study, three experiments were conducted to confirm how representation learning, especially inductive transfer learning including sequential transfer learning and multi-task learning, affects medical domains:

A. Application of deep representation on pediatric diagnosis
B. Application of deep representation on brain hemorrhage diagnosis
C. Application of deep representation on low-dose CT denoising task

A. Application of deep representation learning to pediatric diagnosis

1. Background and Objective

Deep learning technology has often been applied to pediatric medical domains for solving various tasks and has shown excellent research results, such as in disease classification, segmentation, and bone age assessment. However, there are still various obstacles to the real-world clinical application of deep learning models for pediatric tasks despite the impressive achievements of previous studies. One of them is that most pediatric deep learning studies heavily rely on ImageNet pre-trained representations. In general, the ImageNet pre-trained models were often used to relieve the issue of insufficient datasets in pediatric tasks, but recently it has emerged that the pre-trained representation of ImageNet might not be suitable for the medical domain due to a negative transfer by domain difference. To be specific, the ImageNet is a large-scale natural image dataset, so it has three RGB channel spaces. In contrast, medical images such as radiographs, computed tomography (CT), and magnetic resonance images (MRI) have one gray channel space. In addition, medical images typically have a higher resolution than nature images. Thus, these different characteristics can cause a large domain shift while being transferred [14]. According to Ke *et al.* [15], for a large dataset of chest radiographs, ImageNet pre-trained representations showed positive transfer on disease classification task in most CNN models, however, a few models, including the InceptionV3 [16], have either had a negative transfer. In addition, ImageNet pre-trained models often tend to focus sensitively to edge or minor local variations in texture rather than the region of interest (ROI). Thus, it is difficult to interpret how the pre-trained models by ImageNet improved the performances in the medical domain. These results indicate that we need to verify the

effectiveness of ImageNet pre-trained representation on various medical tasks, and representation learning studies suitable for the medical domain have been developed as the transfer learning with pre-trained representation by ImageNet may be limited [17, 18]. In this study, we first constructed class-balanced pediatric radiographs datasets, PedXnets using labels based on radiographic views. Secondly, we conducted the radiographic views recognition task as a pretext task for the development of their supervised representations. Third, the pre-trained representation was used to improve the target task's performance. The Model-PedXnets (i.e., models using the PedXnet pre-trained representation) consist of upstream and downstream tasks like sequential transfer learning. We validated the transferability and positive transfer of our method through pediatric tasks including fracture classification and bone age assessment. The representations of radiographic views using the PedXnets were compared with Baseline-Model (i.e., a trained model with random initialization without pre-trained representations) and Model-ImageNet (i.e., a model using the ImageNet pre-trained representation). In addition, an ablation study was also performed to compare the effects of radiographic view representation in small-scale datasets.

2. Materials and Methods

2.1 Upstream dataset

We retrospectively collected a total of 2,598,404 pediatric radiographs aged from 0 to 18 years in the Asan Medical Center (AMC) between Jan 1997 and Nov 2018. For reflection on the actual frequency of occurrence in the medical center, we divided the original dataset into training and validation sets based on the reference date, Jul 2018. The validation set in the raw original dataset consists of a total of 81,131 radiographs collected over the period from Jul 2018 to Dec 2018. The training dataset in the raw original dataset is composed of the remaining 2,499,598 radiographs. The raw original dataset had a severe imbalance distribution by prescription code. Therefore, when applying our proposed radiographic views labeling which are three types (PedXnet-7C, PedXnet-30C, and PedXnet-68C) according to the degree of anatomy, the class imbalance became highly severe. To address this issue, we under-sampled the data according to the least frequent class and matched the total number of training images equally for a fair comparison between three types of PedXnet datasets (see **Figure 1**).

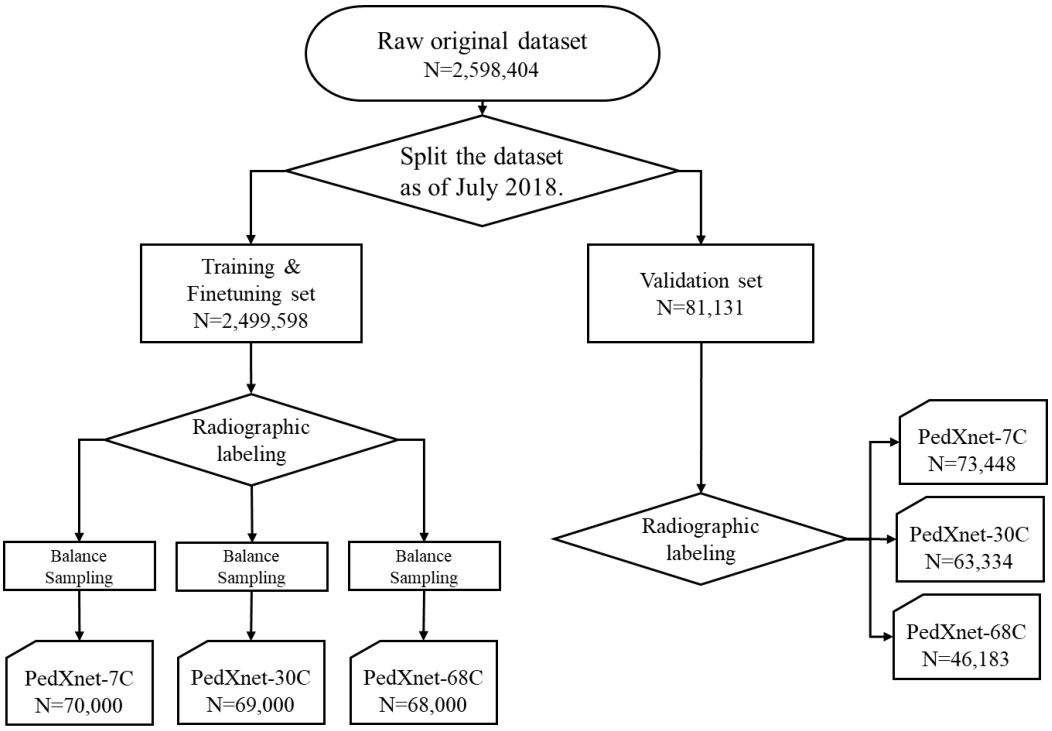

Figure 1. Flow chart for classifying suggested radiographic views in a real-world medical radiography dataset using an upstream dataset. To create balanced datasets for each sort of radiographic view labeling, sampling was carried out individually. The fixed dataset in the upstream validation set had the same radiographic views labeling added to it after a reference date (July 2018), which changed the class-wise mean and variance for each labeling type. The total number of data is N.

2.2 Downstream dataset

2.2.1 Fracture downstream task dataset

To build a fracture task dataset for the downstream, we retrospectively collected 1,772 pediatric radiographs over the period from Jul 2018 to Dec 2018 at Asan Medical Center. The number of fracture cases is 1,010 radiographs and the normal cases are 762 radiographs. The fracture and normal cases were confirmed based on the consensus of two radiologists. We randomly divided the dataset into training, fine-tuning, and validation sets with an 8:1:1 ratio. The fracture task dataset consists of a variety of views of pediatric radiographs including Ankle, Lower leg, Knee, Femur, Shoulder, Humerus, Elbow, Forearm, and Hand.

2.2.2 Bone age assessment (BAA) downstream task dataset

The BAA dataset was released in RSNA Pediatric Bone Age Challenge (2017). The organizers

had provided lists of training, fine-tuning, and validation sets. The dataset has gender information, and the number of training radiographs is 12,611, that of fine-tuning radiographs is 1,425, and that of validation radiographs is 200. According to Halabi *et al.* [19], the training and fine-tuning sets had similar age distributions with an average of 127.321 and 127.156 months, and the validation set had an age distribution with an average of 132.096 months. Radiographs for the training and fine-tuning sets were obtained from Children's Hospital Colorado (Aurora, Colo) and Lucile Packard Children's Hospital at Stanford. The pediatric radiographs for the validation set were collected from Lucile Packard Children's Hospital. The Greulich and Pyle standard method (G-P method) [20] was used by reviewers to determine the ground truth bone age.

2.3 Sequential transfer learning using PedXnet

2.3.1 Radiographic views recognition as upstream tasks

For radiographic views labeling for PedXnets, we benchmarked ImageNet which has a hierarchical structure and is a class-balanced dataset [21]. As shown in **Figure 1**, a hierarchical structure was constructed using anatomical information and the radiographic views information in a large-scale original pediatric dataset. In detail, we first divided the pediatric radiograph dataset into seven major anatomic classes of the human body including the head, chest, upper extremity, abdomen, pelvis, spine, and lower extremity with all pediatric radiographs for construction of the PedXnet-7C dataset. Furthermore, we subdivided the seven classes into 30 classes based on the detailed anatomic areas of radiographs for the composition of the PedXnet-30C dataset. Similarly, the 68 classes were also subdivided using radiographic protocols for the configuration of the PedXnet-68C dataset. We performed classification tasks for radiographic views as upstream tasks using our PedXnet-7C, PedXnet-30C, and PedXnet-68C in order to allow the model to capture representations of radiographic views (see **Figure 2-(a)**). For radiographic views classification tasks, InceptionV3 which is a widely used CNN architecture in studying medical problems (e.g., detecting fractures and bone age assessment tasks) has been trained to classify pediatric radiographs into each corresponding radiographic views class. The model has 11 convolution layers of 1×1, 1×3, and 1×5 kernels, and convolution blocks are applied along with the max-pooling layer for down

sampling. All convolutional layers include batch normalization techniques and rectified linear unit (ReLU) layers. In the upstream tasks, predictions are conducted for target classes with a fully connected layer and a softmax function. We redesigned the last fully connected layer's output channels in accordance with the number of classes of each upstream task.

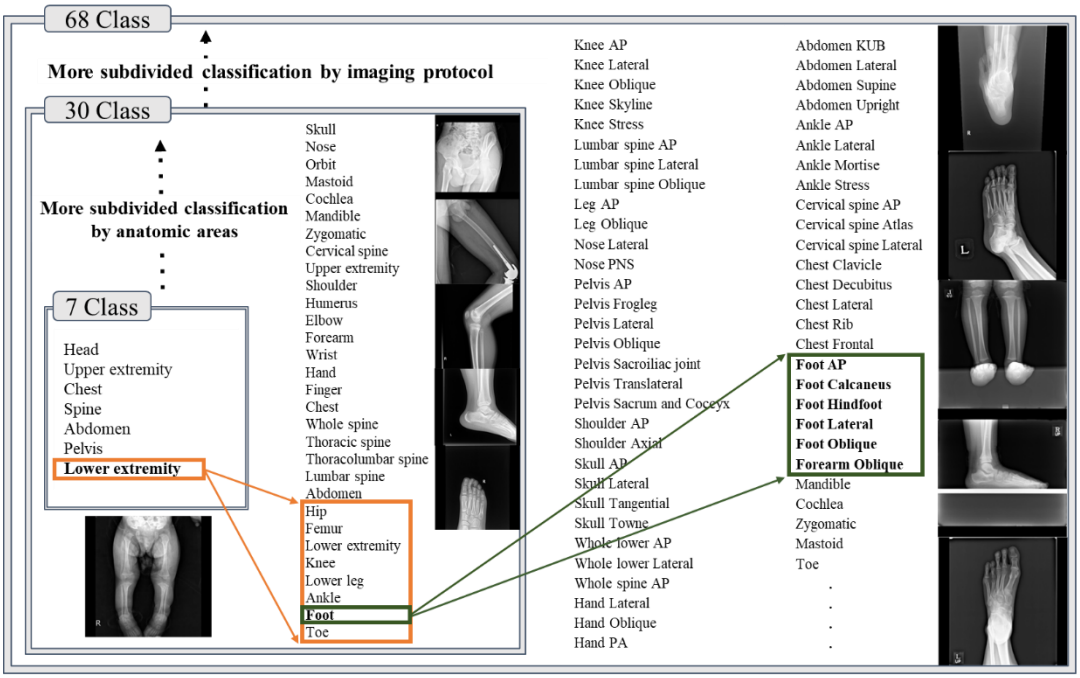

Figure 2. Overview of the labeling procedure for radiographic images for PedXnets. The lower extremities class in the 7 class, for example, can be further subdivided into the hip, femur, knee, lower leg, ankle, foot, and toe classes in the 30 class. Based on the radiograph protocol code, they may be separated into the following subclasses: Foot AP, Foot Calcaneus, Foot Hindfoot, Foot Lateral, and Foot Oblique in the 68 class. Note: KUB, kidney, ureter, and bladder; PNS, paranasal sinus; SI, sacroiliac. PA, posteroanterior; AP, anteroposterior.

2.3.2 Applying pre-trained representation to two pediatric tasks as downstream tasks

To assess whether our pre-trained radiographic view representations by PedXnets benefit applications for medical problems, we conduct two pediatric downstream tasks (see **Figure 2-(b)**). It is important to diagnose pediatric fractures accurately because they will affect the child in his or her growth, and in severe cases, they cause disabilities [22, 23]. A fracture occurs mainly in the upper and lower limbs of the body but anatomically anywhere and frequently occurs during childhood. Therefore, the model should be able to catch fracture features

robustly in multi-view of radiographs. In the fracture classification task, the models should extract general features of fracture in the upper and lower extremities of radiographs. As bone age is an effective indicator for diagnosing various diseases and determining the timing of treatment, the accuracy of bone age assessment is very important [19]. The aim of bone age assessment is to evaluate growth and maturity and to diagnose and manage pediatric disorders. In the bone age assessment task, hand radiographs are mainly used, and the model should extract detailed features from the bones of the wrist, hand, and finger. In this study, we solve these two issues by simply using the pre-trained representations.

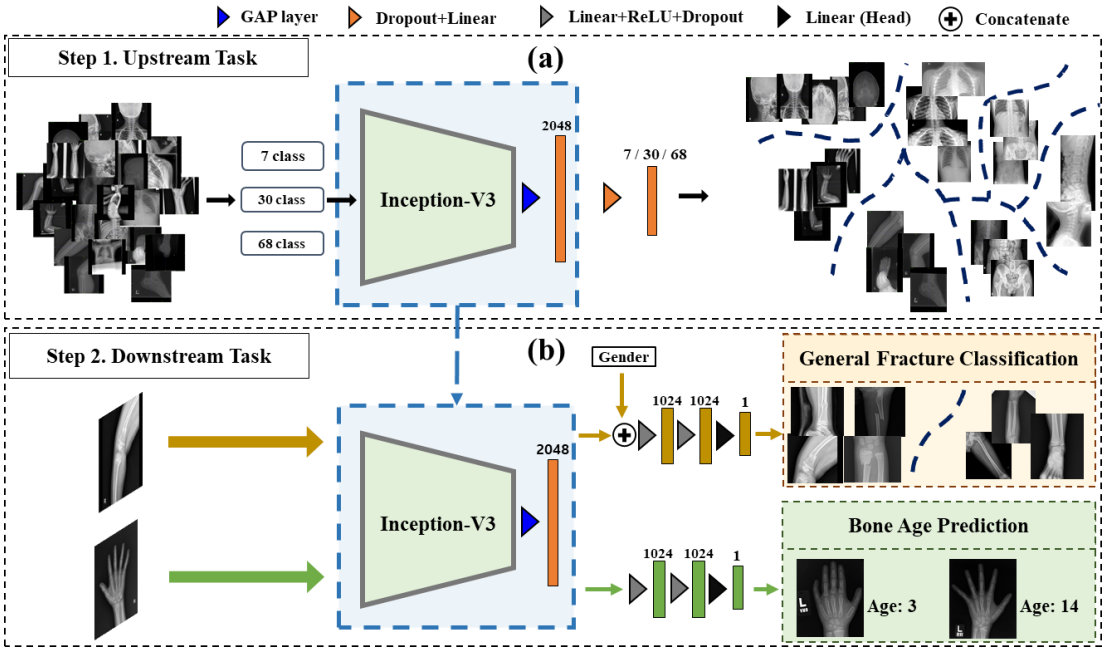

Figure 3. Illustration of the Model-PedXnets method. Upstream and downstream tasks make up the framework. In upstream tasks, radiographic views recognition of pediatric radiographs was conducted to create pre-trained models. Transfer learning using the weights to address two medical issues, such as the categorization of general fractures and the prediction of bone age.

3. Experiments and Results

Metrics. To validate the Model-PedXnets representation in the up and downstream tasks, we employed receiver operating characteristic (ROC), the area under the ROC curve (AUC), F1-score (F1), accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive values (PPV), and negative predictive values (NPV), mean absolute error (MAE), and mean square error (MSE) for quantitative evaluations [24] of multi-class classification in upstream tasks.

We visualized the Model-PedXnets representations with gradient-weighted class activation mapping (Grad-CAM) [25] and channel-wise mean activation where the last convolution layer's features of InceptionV3 were averaged channel-wise, normalized with sigmoid activation, and subsequently interpolated to match the input resolution like Zhou *et al.* [26]. Also, t-distributed stochastic neighbor embedding (t-SNE) [27] is used according to the anatomical hierarchy of radiographic views. The comparisons of Model-Baseline, Model-ImageNet, and Model-PedXnets in downstream tasks are performed using DeLong's ROC comparison [28] and the paired t-test, respectively. The statistical significance level is set at the p-value of 0.05.

Implementation details. For each of the downstream tasks, the same training settings have been applied to Model-Baseline, Model-ImageNet, Model-PedXnet-7C, Model-PedXnet-30C, and Model-PedXnet-68C for fair comparisons.

1) Preprocessing: For each image, min-max normalization with 0.5% clipping of upper and lower bounds was performed to suppress the effect of the L/R mark in radiographs and remove the outlier pixel values. Due to the limitation of GPU resources, all images were resized down into 512×512 by bi-cubic interpolation with the aspect ratio of each original image. In downstream tasks, contrast limited adaptive histogram equalization (CLAHE) [29] is applied to emphasize the bone contrast additionally.

2) Augmentation: There are various radiographic views protocols depending on the age and body size. Thus, we used strong image augmentations to alleviate the heterogeneity of pediatric radiographs and made the model robust for pediatric radiograph protocols applicable to various anatomic locations. We adopted eight augmentation methods from Albumentation [30] as follows: *ShiftScaleRotate*, *HorizontalFlip*, *RandomBrightness*, *RandomContrast*, *RandomGamma*, *GaussNoise*, *Sharpen*, and *RandomBlur*.

3) Setting: The batch size of upstream tasks was 60 and that of downstream tasks was 20. Each model is initialized by a uniform Xavier and trained until a total of 500 epochs with an Adam optimizer, using a learning rate of 1e-4 with a warm-up of 5 epochs, a weight decay of 5e-4, and betas of (0.9, 0.999). The learning rate was reduced during the training following the polynomial learning rate schedule.

3.1 Upstream results of supervised radiographic view representation task

As shown in **Table 1**, Model-PedXnet-7C, Model-PedXnet-30C, and Model-PedXnet-68C, all have sufficiently satisfactory performances in quantitatively (F1>0.78, Accuracy>0.90, Precision>0.84, Recall>0.79). **Figure 4** indicates Model-PedXnet-7C's activation maps that were visualized using Grad-CAM. The Model-PedXnet-7C was activated in the ROI and the activation maps demonstrate that Model-PedXnet-7C can capture clinically meaningful features. In addition, as shown in **Figure 5**, it can be seen that the intermediate features of Model-PedXnets are well clustered in the t-SNE map using the embedded features. These upstream results indicate that Model-PedXnets learned representation without overfitting. Therefore, the model weights of upstream tasks can be applied to downstream tasks.

Table 1. The performance comparisons of radiographic views recognition task as an upstream task.

Upstream validation set	F1 score	Accuracy	Precision	Recall
Model-PedXnet-7C (N=73,448)	0.892	0.911	0.915	0.874
Model-PedXnet-30C (N=63,334)	0.823	0.952	0.933	0.797
Model-PedXnet-68C (N=46,183)	0.785	0.904	0.847	0.798

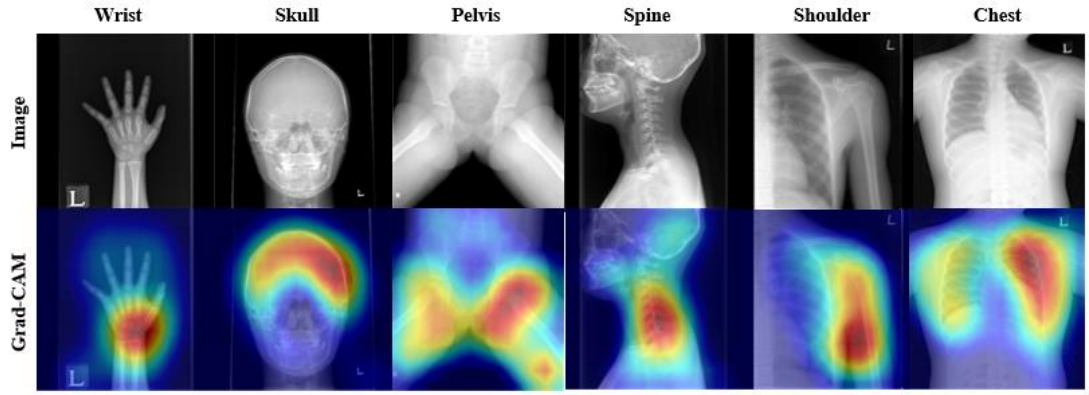

Figure 4. Plots of Model-PedXnet model's Grad-CAM activation maps of radiographic views recognition task as an upstream task with independent validation sets according to each labeling method.

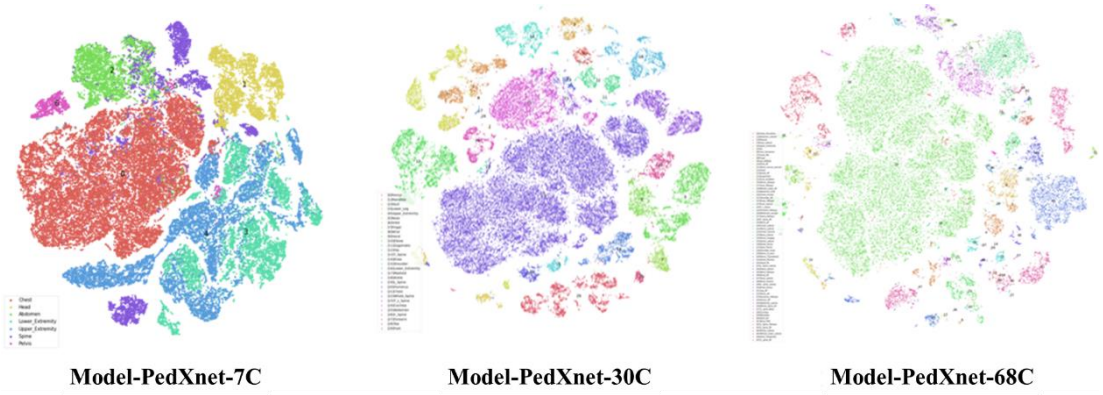

Figure 5. Plots of Model-PedXnets' t-SNE maps of radiographic views recognition task as an upstream task with independent validation sets according to each labeling method. Refer to Supplemental Table 1 for each class detail for upstream validation sets.

3.2 Downstream task results for fracture classification task

When comparing Model-PedXnets and Model-Baseline, the AUC scores of Model-PedXnets are remarkably improved as can be seen in **Table 2**. Although Model-ImageNet achieved the highest AUC 0.880 and SEN 0.853 performances, Model-PedXnet-7C achieved the highest performance with 0.843 in F1, 0.827 in ACC, and 0.893 in PPV. We visualized the features of the last convolution layer of Model-Baseline, Model-PedXnet-7C, and Model-ImageNet using the Grad-CAM to confirm the representations (see **Figure 6**). Model-PedXnet-7C focused exactly where the broken areas were in the two radiographs.

Table 2. The performance comparisons of the fracture classification task.

Network	AUC	F1	ACC	SEN	SPE	PPV	NPV
Model-Baseline	0.795	0.758	0.732	0.721	0.747	0.798	0.659
Model-PedXnet-7C	0.877**	**0.843**	**0.827**	0.798	0.867	**0.893**	0.756
Model-PedXnet-30C	0.861*	0.824	0.799	0.808	0.787	0.840	0.747
Model-PedXnet-68C	0.865*	0.798	0.782	0.740	**0.880**	0.865	0.700
Model-ImageNet	**0.880****	0.790	0.810	**0.853**	0.778	0.735	**0.880**

Note: *, $p<0.05$, **, $p<0.005$. DeLong's test method was adopted for pairwise ROC comparison between the baseline and each model.

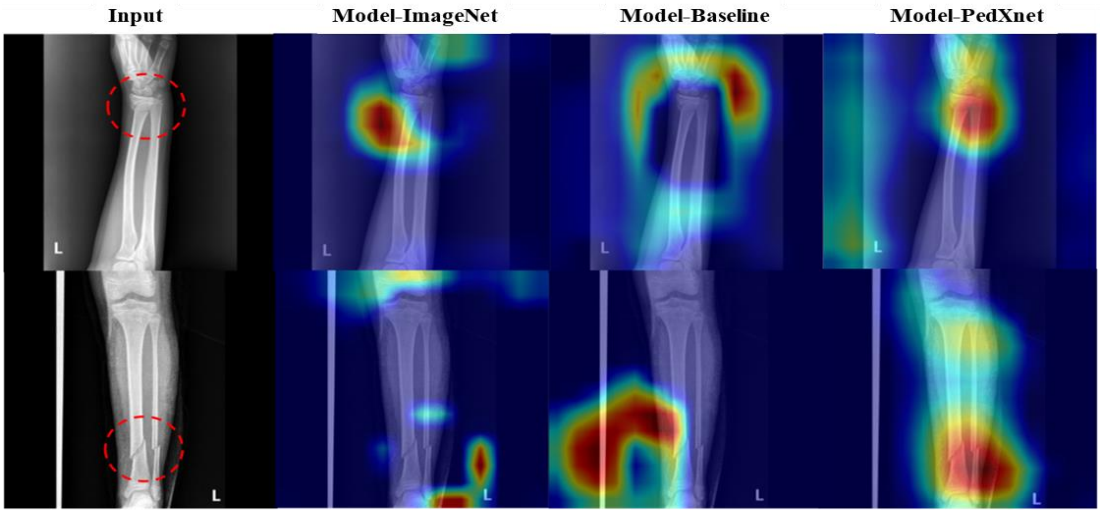

Figure 6. Comparisons of activation maps in the intermediate layer of Model-PedXnet, Model-Baseline, and Model-ImageNet models using Grad-CAM in the fracture downstream task.

3.3 Downstream task results for bone age assessment task

As shown in **Table 3**, Model-PedXnet-7C achieved the best performance of 5.245 in MSE, 42.857 in MAE, and 0.974 in R^2. Model-PedXnet-7C and Model-PedXnet-30C show performance improvements in MAE compared to the baseline model. **Figure 7** indicates that Model-PedXnet-7C captured the most important regions to predict bone age such as carpus and metacarpophalangeal joints, most intensively.

Table 3. The performance comparisons of bone age assessment.

Network	MAE (Month)	MSE (Month)	R^2 Score
Model-Baseline	5.645	52.694	0.968
Model-PedXnet-7C	**5.245**	**42.857**	**0.974**
Model-PedXnet-30C	5.567	49.347	0.971
Model-PedXnet-68C	5.851	55.082	0.970
Model-ImageNet	5.308	47.630	0.971

Note: *, p<0.05, **, p<0.005, MAE, mean absolute error, MSE, mean square error. paired t-test method was adopted for MAE comparison between the baseline and each model, but there is no significance.

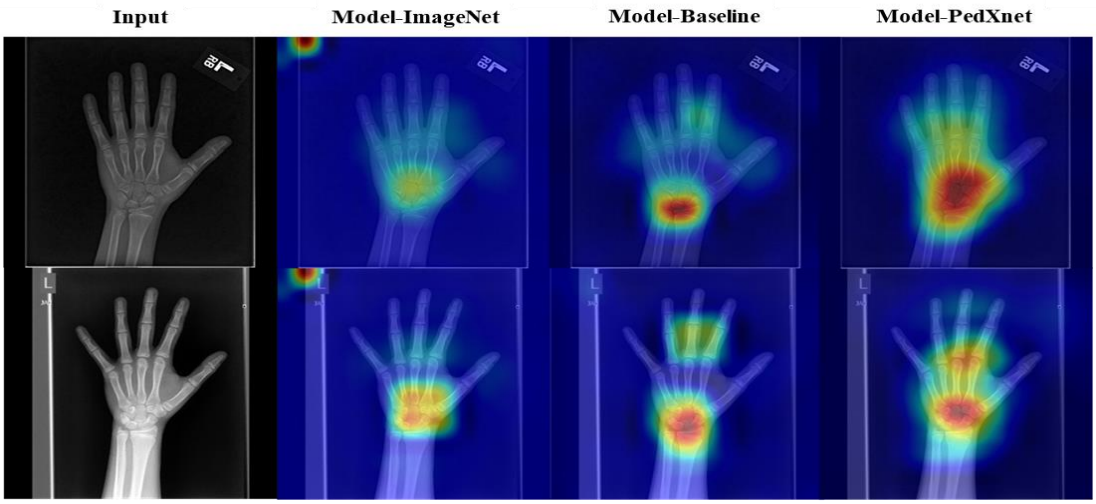

Figure 7. Comparisons of activation maps in the intermediate layer of Model-PedXnet, Model-Baseline, and Model-ImageNet using channel-wise mean activation map in the BAA. The sample in the first row is 152 months old, and the sample in the second row is 167 months old.

4. Discussion

Most of the previous medical tasks mainly use transfer learning due to the scarcity of cases. Especially most pediatric studies rely on ImageNet representation. However, it is still debated whether the ImageNet representation is suitable for the medical domain [36]. In this study, we constructed a class-balanced pediatric dataset, PedXnets, and proposed our Model-PedXnets framework to reap the benefits of transfer learning in medical domains. In **Tables 2** and **3**, the Model-PedXnets showed superior performance improvements by a large margin as compared to Model-Baseline in downstream pediatric tasks including fracture classification and bone age assessment. Even though PedXnets has smaller-scale datasets than ImageNet, Model-PedXnet showed equal and in some cases even superior performance than Model-ImageNet. As shown in **Figures 6** and **7**, performance gaps between Model-ImageNet and Model-PedXnet were noticeably different in activation map visualization. The Model-ImageNet focused on some minor local context, while the Model-PedXnet-7C focused more on medically meaningful ROI. In addition, comparing the ablation studies' results amongst Model-PedXnets, it was found that for radiographic view representation made with fewer classes of datasets, Model-PedXnet-7C, was more effective than other Model-PedXnets. It can be interpreted that pediatric radiographs include views of various sizes according to age and excessive divided classes based on the radiographs' protocol can lead to label noise because

there were a lot of overlapping regions. For example, the chest AP view images were similar to abdomen AP views in newborns and infants. Therefore, the label noises made the network miss meaningful features and resulted in a negative transfer issue, which decreased the transfer learning effects. Despite the improved performance, our method has some limitations. First, as our proposed methodology can rely on the backbone network and pre-processing, it can lead to sufficiently different results depending on different backbone networks and pre-processing. Second, as shown in **Figure 2**, we performed excessive random under-sampling in the raw original dataset to build class-balanced datasets according to the anatomical hierarchy of radiographic views. This has reduced the total number of training data and there might be a possibility that the total number of data was insufficient compared to ImageNet, so it did not show an appropriate effect. Third, the labeling of radiographic views may vary depending on the radiologist, which could change the results since our proposed method is of a supervised manner based on subjective labeling according to an anatomical or radiographic perspective.

5. Summary

In this study, we introduced a supervised manner of medical representation learning for pediatric tasks with radiographic view labels. In the upstream task, we designed the class-balanced pediatric radiograph datasets (PedXnets) by radiographic views labels and conducted representation learning for pediatric problems through a radiographic-views classification task on the PedXnets in a supervised manner. According to two downstream evaluation results, the representation by seven major anatomical view labels was the most effective and the transfer effect of Model-PedXnet-7C was positive in both pediatric downstream tasks including fracture classification and bone age assessment tasks. Model-PedXnets showed superior results by a large margin compared to Model-Baseline. Model-PedXnets even showed that were results equivalent and in some cases better than Model-ImageNet, even though PedXnets were smaller datasets than ImageNet. In addition, the proposed representation learning allowed networks to capture more semantic features in the ROI of radiographs. Our study could be helpful for medical domains, particularly for pediatric radiographs research for which the data is difficult to obtain.

B. Application of deep representation learning to brain hemorrhage diagnosis [31]

1. Background and Objective

With the recent development of deep learning, the classification and segmentation tasks using computer-aided diagnosis (CAD) in non-contrast head computed tomography (NCCT) for intracranial hemorrhage (ICH) have become popular in emergency medical care. However, a few challenges remain:

ICH's heterogeneity makes training difficult. ICH can arise anywhere in the skull and have numerous subtypes. ICH can be clear to dim and thick to diffused and may lessen over time [32]. Real emergency medical centers witness a lot of tiny ICHs (see **Figure 8**). Even competent radiologists can struggle with little ICHs, therefore they consider all of their properties. To overcome these variations, CAD models must be trained with many ICH features in mind, as radiologists do.

High sensitivity and specificity are required. Most ICH models are sensitive but lack specificity. A deep learning-based diagnostic model for triage must have good sensitivity and specificity. Even a slight traumatic brain hemorrhage overlooked (false negative) can worsen disease severity and induce neurologic impairment. Due to limited resources in the emergency department, false positives can delay diagnostic workflows for more critical patients [33].

Volume predictions are costly. 3D volumetric CT scans including NCCT demand a big GPU memory. Previous studies avoided this issue by using 3D patch- or 2D slice-based approaches. 3D patch-based captures more spatial information than 2D slice-based. However, the outcomes can vary greatly depending on patch size. In particular, excessively small patch sizes cause the network to ignore spatial information in the entire area [34]. The 2D slice-based method uses fewer GPU resources than the 3D patch-based method, allowing the complete image to be utilized and readily expanded to 3D by stacking slice-level features or processing the stacked features by a 3D operator. However, the results can seem strange because it renders 3D through simply stacking.

There is an external data vulnerability. If the training data doesn't adequately reflect the task domain, the trained model may perform poorly on external data. Voter *et al.* [35] highlighted that the type and amount of ICH in their external test dataset could affect performance. This is

a common problem, especially in medical deep learning, and recent research has focused on generalizability and robustness and should adopt the learning approach suitable not only for internal but also external data for generalizability. For real-world clinical applications, the network must be robust to disease-specific data distributions, and external validation tests are needed to ensure network robustness. In this study, we proposed a supervised multi-task aiding representation transfer learning network (SMART-Net) for ICH to overcome these challenges. The proposed framework consists of the up- stream and downstream components. In the upstream, a weight-shared encoder of the model is trained as a robust feature extractor that captures global features by performing slice-level multi-pretext tasks (classification, segmentation, and reconstruction). We added a consistency loss to regularize discrepancies between classification and segmentation heads, which improved representation and transferability. In the downstream, the transfer learning was conducted with a pre-trained encoder and 3D operator (classifier or segmenter) for volume-level tasks. Excessive ablation studies were conducted, and the SMART-Net was developed with optimal multi-pretext task combinations and a 3D operator. Using four test sets (one internal and two external test sets that reflect a natural incidence of ICH, and one public test set with a relatively small volume of ICH cases), we compared our SMART-Net at the volume level with the previous state-of-the-art ICH classification methods, ICH segmentation methods, and representation learning methods.

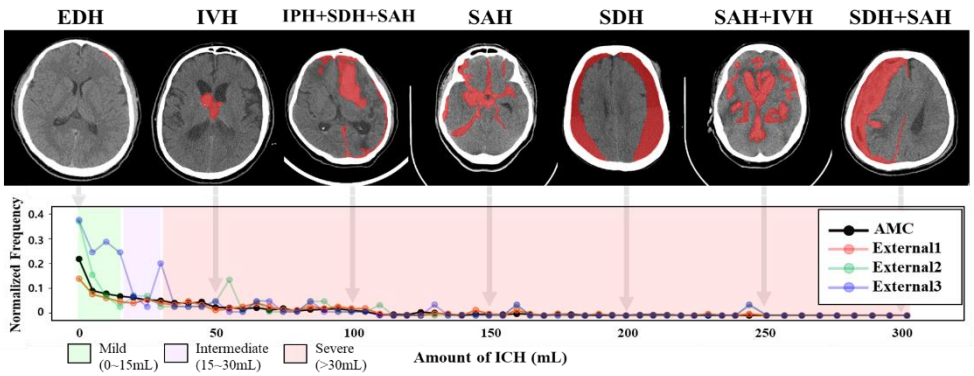

Figure 8. Histograms of four datasets are shown according to the quantity of ICH. Above the graphs, samples corresponding to each severity and subtype class are presented. (Note: AMC, Asan Medical Center; CPH, cerebral parenchymal hemorrhage; IVH, intraventricular hemorrhage; EDH, epidural hemorrhage; SDH, subdural hemorrhage; SAH, subarachnoid hemorrhage)

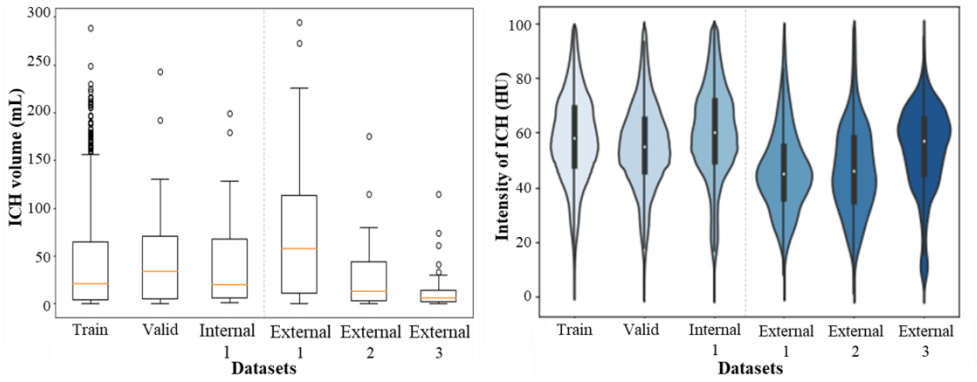

Figure 9. Overview properties of ICH patients in four independent ICH datasets. (Left) box plots of the bleeding volume for ICH patients in each dataset, (right) violin plots of the hemorrhage intensity for ICH patients in each dataset.

2. Materials and Methods

2.1 Dataset

The AMC training dataset CT images were gathered by searching the AMC database for individuals who had successive NCCT exams from September 2009 to April 2017. In the AMC training dataset, CT scans are 512×512 pixels with depths between [28, 50]. The AMC training dataset included 782 ICH patients and 760 healthy controls based on CT scan clinical radiology reports. We used the hold-out method to segregate 5% of the AMC training dataset. **Table 4** summarizes ICH datasets. There was a class imbalance in the five ICH subtypes such as cerebral parenchymal hemorrhage, intraventricular hemorrhage, epidural hemorrhage, subdural hemorrhage, and subarachnoid hemorrhage. We combined all ICH subtype classes and performed binary classification and segmentation at the volume level to reduce class imbalance and focus on brain bleeding in emergencies. Three neuroradiologists independently conducted labeling of ICH segmentation masks. Two veteran emergency radiologists with 14 and 23 years of experience reviewed the data. Senior reviewers adjusted segmentation masks after reaching a consensus. The masks were pixel-annotated with 1 for ICH and 0 for non-ICH. We employed one internal and two external test datasets to verify the models in a clinical setting. From March to June 2017, AMC collected the internal dataset. Nowon Eulji Medical Center in Korea gathered external1 from July to October 2018. Pohang Stroke and Spine Hospital collected external2 from March to June 2019. We used a public dataset from

PhysioNet (1.3.1) with 75 patients to verify the models in the ICH case setting. The PhysioNet dataset has a mean ICH volume of 15.10 mL, which is low (see **Figure 9**). We trained and tuned using only the AMC training dataset, and other dataset properties were only referenced.

Table 4. Patient demographic information and medical properties in four ICH datasets.

Dataset		Slice (Patient)		Sex		Age	Spacing	Manu.		
		H	N	M	F	Mean±Std	Mean	S	G	P
Train (AMC)		10,805 (747)	39,958 (722)	872	597	55.2±17.6	x,y : 0.41 z : 4.80	1,378	97	0
Validation (AMC)		541 (39)	1,907 (38)	40	37	53.7±18.9	x,y : 0.41 z : 4.80	71	6	0
TEST	Internal1	451 (29)	3,961 (100)	69	60	55.6±17.7	x,y : 0.41 z : 4.80	125	4	0
	External1	1405 (86)	5,950 (162)	Unknown			x,y : 0.42 z : 5.00	116	132	0
	External2	282 (22)	7,507 (203)	96	130	59.6±18.8	x,y : 0.43 z : 5.00	0	0	225
	External3	315 (34)	2,421 (39)	39	34	32.1±17.0	x,y : 0.43 z : 5.00	73	0	0

Note: H, hemorrhage; N, normal; M, male; F, female; Manu., manufacturer of device; S, Siemens; G, General Electronics; P, Philips.

2.2 Sequential transfer learning using multi-task pretraining

We describe SMART-Net (see **Figure 10**). Upstream and downstream tasks comprise our framework. Upstream tasks include classification (*CLS*), segmentation (*SEG*), and reconstruction (*REC*) (see **Figure 10-(a)**). In upstream tasks, a shared encoder may extract comprehensive features across all tasks, resulting in a robust feature extractor. In addition, we introduced consistency loss to better representation for the regulation of misaligned between classification and segmentation heads. In downstream tasks, there are two types of volume-level target tasks; classification and segmentation of ICH, respectively (see **Figure 10-(b, c)**). We employed a pre-trained encoder as a feature extractor and connected it with optimal 3D operators to address volume-level downstream tasks.

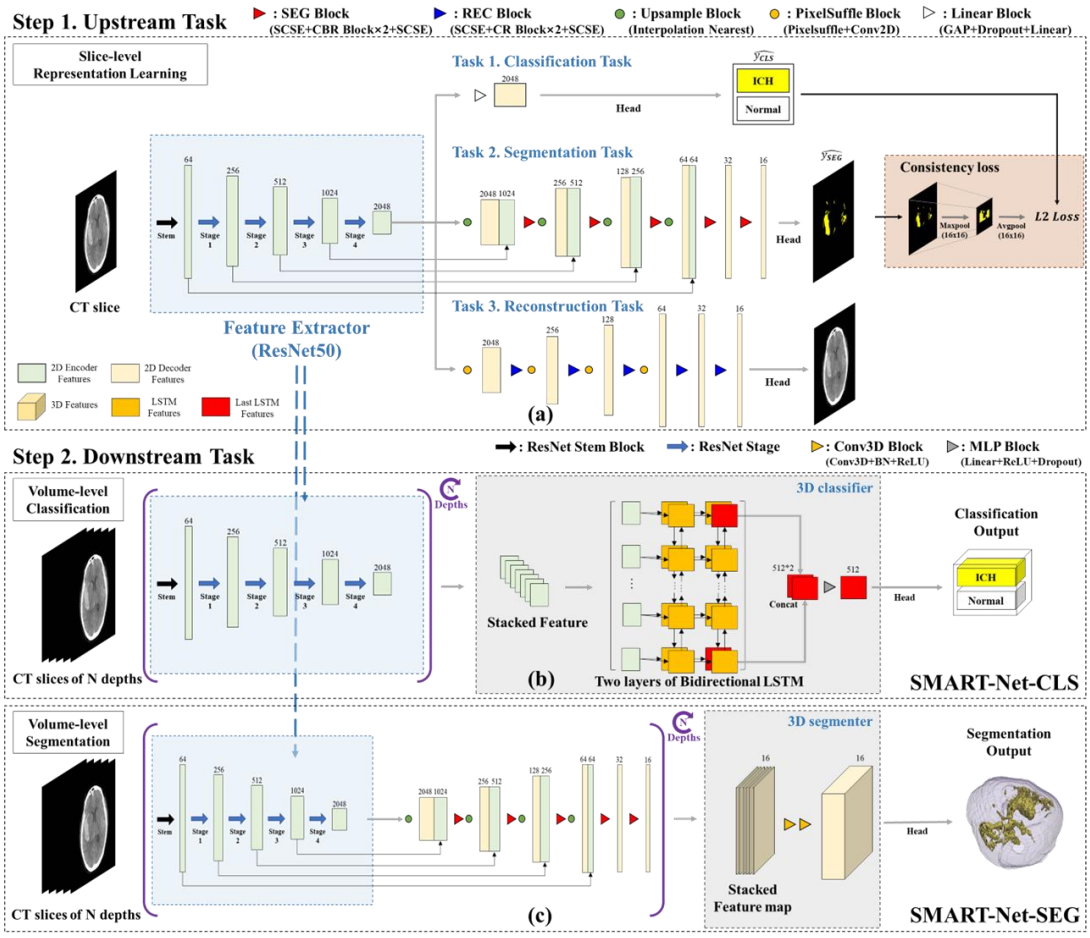

Figure 10. Schematic overview of the SMART-Net framework. Our framework comprises upstream and downstream tasks: (a) An upstream task is a slice-level representation learning process using three multi-pretext tasks (classification, segmentation, reconstruction) with consistency loss. A downstream task is a process of extending to volume-level ICH tasks by transfer learning using the pre-trained encoder in the upstream. The pre-trained encoder, as a feature extractor, extracts and accumulates features in a slice-by-slice way, as much as the depth of the CT scan. (b) Accumulated features are transmitted to an LSTM-based classifier to process continuous information and perform volume-level classification. (c) For volume-level segmentation, the stacked feature maps are extracted with the pre-trained encoder and an initialized slice-level decoder. The feature maps are processed by the Conv3D-based segmenter to complement volumetric spatial details.

2.2.1 Multi-pretext tasks learning for representation as upstream task

To create a robust feature extractor that captures worldwide ICH features through pretext jobs, we used a hard parameter shared architecture [7] with a shared encoder and three task-specific layers. As an encoder, we chose ResNet-50. The encoder uses one stem block and four stages of ResNet to extract features from a CT slice image (see **Figure 10-(a)**). Each target-specific layer receives the encoder's last layer's features. Each ResNet stage's features are sent as skip

connections to depth-level segmentation decoders. The first task-specific layer for *CLS* has a simple linear classifier with a global average pool, dropout, and bias. Our *CLS* upstream framework is the same as ResNet-50. This *CLS* task determines ICH in each CT slice, allowing the encoder to discover its crux characteristics. The *CLS* loss is defined as the binary cross-entropy loss (BCE loss). The second task-specific layer for *SEG* has U-Net decoder blocks with skip connections. Our *SEG* upstream architecture has skip connections like U-Net. The depth of the *SEG* decoder is selected considering the down sampling of ResNet stages, and we added spatial and channel squeeze & excitation (SCSE) [36] blocks at the beginning and end of each decoder block to reinforce the representation of the feature. This *SEG* task localizes the ICH lesion and segments the brain hemorrhage in each CT slice, allowing the encoder to catch local features of the lesion [37]. The *SEG* loss was defined as the overlap-based dice-coefficient loss (DICE loss). The third task-specific layer for *REC* contains PixelShuffle [38] decoder blocks, which super-resolve low-resolution data into high-resolution data with low computational cost. Our *REC* upstream framework is similar to an autoencoder without skip connections to avoid trivial solutions and generate more important features [39]. This *REC* task reconstructs the input CT slice images via encoding-decoding, which is widely employed in unsupervised learning to help the encoder learn the disentangled feature representation by refining image features [40]. The *REC* loss was defined as the mean absolute error loss. In multi-head structures, the preference of characteristics in the heads of various target-specific layers might affect training [41-43]. This discrepancy also emerged during our training (see **Figure 12**). We incorporated consistency loss to regularize *CLS* and *SEG* head differences and improved the concurrence rate. Since the *SEG* head and *CLS* head don't match the resolution, we down sampled the *SEG* head to the *CLS* head. Two outputs were compared by mean square error loss using the nonparametric down sampling techniques Maxpool and Avgpool of the 16×16 kernel. The loss allows encoder to focus on typical *CLS* and *SEG* elements.

2.2.2 Volume-level classification and segmentation tasks as downstream task

To properly leverage the general representation, we employed transfer learning with a pre-trained encoder as a feature extractor. To reduce the degradation of individual performance [44], we kept the shared encoder and eliminated target-specific layers. To achieve volume-

level classification or segmentation, we processed the 3D features by stacking 2D features predicted by the feature extractor iteratively by CT depths. The 3D features are transmitted to the 3D operators to enhance volumetric information. In volume-level target classification, we used a 3D classifier with a variable-length LSTM layer to capture sequential information in successive slices. The volume-level classification has been performed by determining the existence of ICH using BCE loss. In volume-level target segmentation, we chose a 3D segmenter based on 3D convolution layers capable of capturing 3D spatial features. The slice-level decoder was randomly initialized and retrained to avoid negative transfer bias and focus only on the target task. Localizing the ICH lesion at the volume level utilizing BCE and DICE losses was used for target segmentation. To avoid catastrophic forgetting [45], gradual unfreezing was used instead of direct fine-tuning [46]. We trained the 3D operator with the encoder frozen, then slowly unfroze the encoder.

3. Experiments and Results

Metrics. We used receiver operating characteristic (ROC) analysis, including AUCs, sensitivity (SEN), specificity (SPE), and F1-scores (F1) for the quantitative evaluation of ICH classification [24], and dice similarity coefficient (DSC) for ICH segmentation [47]. The DSC is an overlap-based indicator used only for positive cases. A metric was needed to evaluate segmentation performance in negative instances. Although pixel-wise specificity might be utilized, it was not useful as a criterion for performance comparisons in the ICH segmentation job since brain hemorrhage accounted for such a small part of CT scans. To show the difference in negative case performance, we established the false-positive volume (FPV), which is clinically meaningful despite being a rough estimate. FPV was determined by multiplying segmentation results with patient pixel spacing as follows:

$$FPV = \sum_{normal}(\hat{y}_{seg_i} * s_i), \tag{1}$$

where $\hat{y}_{seg_i}$ is the output in volume-level *SEG*, and s_i is meta information of the pixel spacing volume in CT scans for converting to the volume unit (μL). False-positive reduction improves when FPV decreases. We also propose the concurrency rate to measure consistency loss as follows:

$$Concurrence\ rate = \frac{1}{N}\sum_{i=1}^{N}[round(\hat{y}_{cls_i}) == \max(round(\hat{y}_{seg_i}))], \tag{2}$$

where $\hat{y}_{cls_i}$ and $\hat{y}_{seg_i}$ are the output of the *CLS* head and *SEG* head at slice-level, respectively, and N is the positive or negative slice. […] is the Iverson bracket which equals 1 when the expression within it is true and equals 0 otherwise. The max function was applied to the output of the SEG head for matching the *CLS* head output's dimension. The round function is the round-to-even method. We used an activation map to analyze pretext task combinations, consistency loss, and previous studies to determine how the pretext task affects the feature extractor (see **Figure 11**). To test consistency loss's influence on representation learning, we evaluated classification and segmentation upstream/downstream performance (see **Table 5**). We examined the activation maps of the convolution layer before each target-specific head to visualize consistency loss and misalignment difficulties in multi-head structure (see **Figure 12**). We tested our SMART-Net with existing state-of-the-art models in volume-level ICH classification and segmentation using four test sets to assess the models' robustness and performance (see **Table 6**). ImageNet (pre-trained models by ImageNet), Model Genesis [48], and Autoencoder [49] were chosen as comparison groups for SMART-Net representation learning methods. For volume-level classification tasks, we compared SMART-Net-CLS to [50-52]. Studies [53, 54] compared SMART-Net-SEG to volume-level segmentation. We conducted excessive ablation investigations on all multi-pretext task combinations of *CLS*, *SEG*, and *REC* with consistency loss to compare transferability and prove SMART-Net as the best combination (see **Table 7**). To compare ablation studies fairly, all ablation study models' encoders were set to ResNet-50, and training settings were the same. DeLong *et al.* [28] compared ROC curves for AUC values of correlated data in the classification task and paired t-tests were used to compare DSC and FPV values in the segmentation task.

Implementation details. For each of the up and downstream tasks, the same training settings have been applied for fair comparisons. We clipped [0, 80] Hounsfield units (HU) and scaled them to [0, 1] for network input. CLAHE was used to boost CT scan contrast. Due to GPU restrictions, the image was linearly interpolated to 256×256. We did not use cropping nor did we make patch-volume images, which could lead to unstable results and false positives [34]. To reduce ICH heterogeneity, seven transformations were used from Albumentation [30]: *ShiftScaleRotate*, *RandShiftIntensity*, *HorizontalFlip*, *RandBrightness*, *RandContrast*,

GaussNoise, and *Blur*. The batch size was set to the GPU's maximum memory. The network was initialized by a uniform Xavier, and we used an Adam optimizer with a learning rate of 1e-4 in the upstream and 5e-6 in the downstream, a 5-epoch warmup, and 5e-4 weight decay and betas (0.9, 0.999). The learning rate was reduced by the polynomial learning rate schedule. Upstream training had 1000 epochs and downstream training had 500. We manually balanced multi-task weights between three pretext tasks. We fixed multi-task weights identically to compare MTL's role as a pretext task rather than comparing multiple combinations.

3.1 Upstream results

3.1.1 Comparison of pretext tasks' effects on the upstream task

Figure 11 shows that despite the identical training settings, the encoder might focus differently depending on the upstream pretext tasks. ImageNet's pre-trained feature extractor recognizes the whole brain as an object and focuses on cerebral textures and edges. The Model Genesis feature extractor identifies ROI by focusing on the cortical region inside the brain and the image edge outside the brain, similar to the *REC* pretext task. According to ablation research, *CLS*, *SEG*, and *REC* each have a distinct personality. *CLS* allows the feature extractor to focus on the most prominent points of ICH, while *SEG* and *REC* collect the overall picture characteristics. The *REC* pretext feature extractor is equally activated in the brain, but the *SEG* feature extractor is sporadically focused on the entire image. Dual pretext tasks have harmonic qualities. In *CLS*+*SEG* and *CLS*+*REC* activation maps, the intensely activated *CLS* pretext task and the extensively activated *SEG* or *REC* pretext tasks work together, resulting in more accurate ROI activation. *SEG*+*REC* pretext task evened out intermittently active areas. Triple pretext tasks combine the representations of three pretext tasks to create an ROI-friendly representation. In summary, the type and quantity of pretext tasks affect feature extraction, and when all three were run simultaneously, the most relevant activation map appeared.

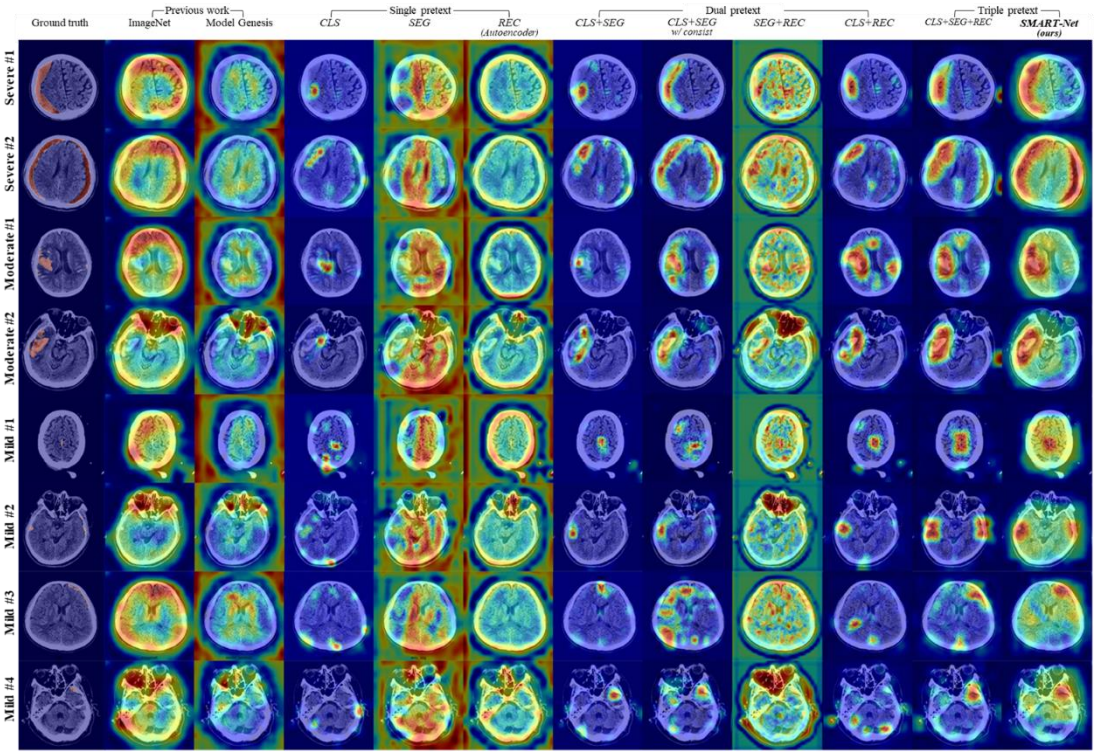

Figure 11. Comparisons of the pre-trained encoder activation map in the upstream task according to multi-pretext task combinations and the previous representation learning approaches. To perform a comparison based on the severity of ICH cases, two severe cases (≥ 30 mL), two moderate cases (15-30 mL), and four mild cases (≤ 15 mL) of ICH are displayed.

3.1.2 Effect of consistency loss

The consistency loss between *CLS* and *SEG* heads affected the encoder's representation. In **Figure 11**, applying consistency loss to dual and triple pretext tasks makes the ROI activation area more significant. **Table 5** shows that the concurrence rate rose by 7% in hemorrhage cases and 9.5% in healthy controls. Even while upstream slice-level performance may have declined, downstream volume-level performance improved, meaning consistency loss increased transferability. With the addition of the *REC* pretext task, the *CLS*+*SEG* model's downstream AUC performance increased ($p<0.05$), and when consistency loss was included, the performance was much better ($p<0.01$). In downstream segmentation, adding *REC* pretext to *CLS*+*SEG*+w/cosists. enhanced DSC ($p<0.05$). Adding consistency loss to *CLS*+*SEG*+*REC* enhanced DSC and FPV ($p<0.05$). In **Figure 12**, the left three columns show three types of discrepancies in multi-head structures: the *SEG* head detected lesions, but the *CLS* head missed them (SEG-TP/CLS-FN), the *SEG* head missed lesions, but the *CLS* head detected them (SEG-FN/CLS-TP), and both heads detected lesions but focused on different locations (SEG-

TP/CLS-TP). With consistency loss, all three types of activation zones were adjusted to meaningful ROIs, and two target-specific heads focused on the same sites. Based on the experimental upstream findings, we used the encoder trained by *CLS*, *SEG*, and *REC* pretext job with consistency loss in the SMART-Net framework.

Table 5. Comparisons of up and downstream performance according to consistency loss in the internal test set.

	Upstream slice-level comparison								Downstream volume-level comparison					
	Concurrence		Classification				Segmentation		Classification				Segmentation	
Pretext Task	Pos.	Neg.	AUC	F1	SEN	SPE	DSC	FPV↓	AUC	F1	SEN	SPE	DSC	FPV↓
CLS+SEG	0.828	0.841	0.954	0.638	**0.871**	0.903	0.504	4.9	0.956	0.800	0.897	0.900	0.604	23.8
+ w/ consist.	**0.871**	**0.914**	**0.958**	**0.731**	0.815	**0.953**	**0.515**	**4.3**	**0.959**	**0.831**	**0.931**	**0.910**	**0.612**	**7.7**
CLS+SEG+REC	0.831	0.824	**0.968**	**0.691**	**0.902**	**0.920**	**0.530**	3.2	0.984	0.871	0.931	0.940	0.622	12.7
+ w/ consist.	**0.901**	**0.919**	0.963	0.685	0.897	0.919	0.506	**2.4**	**0.988**	**0.933**	**0.966**	**0.970**	**0.642**	**4.9**

Note: ↓, the lower the value, the better performance; consist., consistency loss; Pos., hemorrhage cases; Neg., healthy controls.

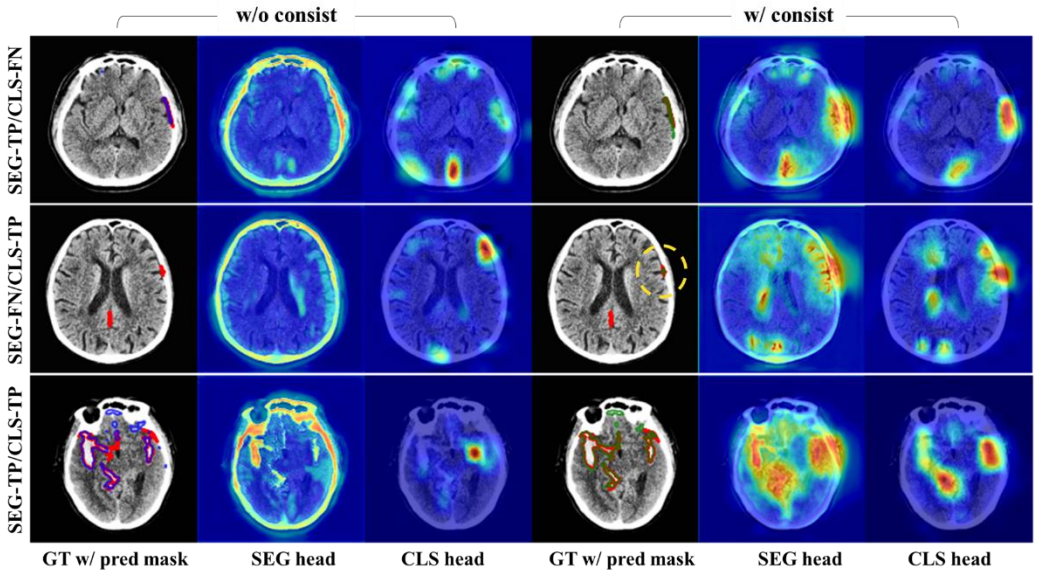

Figure 12. Illustrations of the output mismatches caused by the target-specific multi-head structure and of the effects of the consistency loss in the upstream task. The first and fourth columns are the CT image overlapped with the ground truth mask (red) and prediction masks (blue, w/o the consistency loss; green, w/ the consistency loss). The second and fifth columns are the activation maps of the segmentation head according to the consistency loss. The activation map of the classification head according to the consistency loss is shown in the third and sixth columns. Note: TP, true positive; FN, false negative.

3.2 Comparisons of target tasks in downstream analysis

3.2.1 Comparison of performance of previous work

Table 6 displays the quantitative findings of volume-level target tasks, where the top of the table corresponds to a comparison with the prior classification studies and the bottom refers to a comparison with the previous segmentation research. For the classification task, the proposed SMART-Net-CLS model outperforms other models in all metrics, with AUC, F1, SEN, and SPE values of 0.988, 0.933, 0.966, and 0.970 in the internal test set, respectively, and with the same tendency in external test sets by a large margin, with the exception of the full-version model derived from the study by Patel *et al.* [51]. The SMART-Net-CLS is steady with respect to external datasets, whereas other methods display unstable performance. There is a statistically significant distinction between the SMART-Net-CLS and the simple version of Patel *et al.*, Singh *et al.* [52], Model Genesis ($p<0.005$), and ImageNet ($p<0.001$), which is more pronounced in external datasets. For the segmentation task, the suggested SMART-Net-SEG model achieved the best DSC and FPV in the internal test set, with 0.642 and 4.9μL, respectively. By comparing the findings from Isensee *et al.* [54] and SMART-Net-SEG, we can see that although the 3D version of Isensee *et al.* has the best DSC performance in several test sets, it has an excessive number of false positives exceeding 660μL in the control groups of all test sets. In contrast, SMART-Net-SEG produced the most balanced findings in both the ICH group and the control group and is more robust in external datasets than Isensee *et al*.'s 3D version. Consequently, the SMART-Net-SEG achieved the best DSC and FPV at 0.57 and 4.5μL, respectively, based on the average result of the four test sets. Statistically, SMART-Net-SEG differs from the techniques of Patel *et al.* [53] ($p<0.05$), Model Genesis, and ImageNet ($p<0.001$), and the difference is more obvious in external datasets. **Figure 13** illustrates the slice-level segmentation results for two ICH patients (Severe#1: approximately 199.45 mL and Mild#1: approximately 11.99 mL) and two healthy controls (Normal#1: beam hardening artifact and Normal#2: no lesion) in the internal test set. Our SMART-Net-SEG shows more consistent findings considering the ground truth in both severe and mild ICH situations, and our model has a reduced false-positive rate in normal cases, including the artifact case, compared to previous approaches. **Figure 14** demonstrates that our SMART-Net-SEG surpasses all other models in terms of false-positive reduction performance.

Table 6. Quantitative results of volume-level target classification and segmentation tasks for a comparative analysis with previous methods on four test sets.

Classification results

	Internal 1				External 1				External 2				External 3			
Methods	AUC	F1	SEN	SPE	AUC	F1	SEN	SPE	AUC	F1	SEN	SPE	AUC	F1	SEN	SPE
Patel *et al.* (2019b) (Full)	0.940	0.852	0.897	0.940	0.928*	0.828	0.837	0.901	0.888	0.650	0.591	0.975	0.918	0.816	0.912	0.718
Patel *et al.* (2019b) (Simple)	0.875**	0.632	0.621	0.900	0.747***	0.560	0.488	0.864	0.596***	0.345	0.227	0.990	0.733***	0.800	0.882	0.205
Singh *et al.*	0.864**	0.607	0.586	0.900	0.868***	0.718	0.709	0.858	0.715***	0.270	0.545	0.729	0.776**	0.603	0.559	0.774
Scratch (end-to-end)	0.914***	0.724	0.724	0.920	0.836***	0.696	0.733	0.802	0.805***	0.384	0.636	0.818	0.808**	0.699	0.853	0.487
ImageNet (=Nguyen *et al.*)	0.903*	0.780	0.793	0.930	0.834***	0.629	0.523	0.926	0.661***	0.414	0.273	0.995	0.768**	0.730	0.794	0.667
Autoencoder (Hinton *et al.*)	0.714***	0.462	0.724	0.590	0.627***	0.522	0.628	0.586	0.576***	0.190	0.909	0.167	0.676***	0.633	0.912	0.154
Model Genesis (Zhou *et al.*)	0.862**	0.505	0.793	0.610	0.804***	0.626	0.779	0.623	0.745**	0.191	0.955	0.128	0.654***	0.633	0.912	0.154
SMART-Net-CLS (Ours)	0.988	0.933	0.966	0.970	**0.966**	0.866	0.942	0.877	0.947	0.706	0.864	0.946	0.953	0.842	0.971	0.744

Segmentation results

	Internal 1		External 1		External 2		External 3		Average	
Methods	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)
Isensee *et al.* (2D version)	0.627	40.3	0.599	14.6***	0.526	60.2***	0.505	25.0*	0.564	35.0
Isensee *et al.* (3D version)	0.622	408.0***	0.665***	511.5***	0.541	254.2***	0.462	1469.5***	0.573	660.8
Patel *et al.* (2019a)	0.491***	21.1	0.504***	2.8	0.450**	223.0***	0.394***	74.1	0.460	80.3
Scratch (end-to-end)	0.534**	44.4***	0.419***	17.6***	0.351***	6.6	0.456**	36.7	0.440	26.3
ImageNet	0.603*	37.1*	0.591	29.6***	0.529	21.3**	0.494	249.7	0.554	84.4
Autoencoder (Hinton *et al.*)	0.612	77.9**	0.601	104.3***	0.554	106.3***	0.441**	357.7***	0.552	156.5
Model Genesis (Zhou *et al.*)	0.600*	72.5***	0.565***	34.4**	0.480**	29.3**	0.486	83.5*	0.533	54.9
SMART-Net-SEG (Ours)	0.642	4.9	0.602	2.2	0.537	7.9	0.519	3.1	0.575	4.5

Note: ↓, the lower the value, the better performance; p-values were calculated between SMART-Net vs. others: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. The red color represents the best performance, and the blue color represents the second-highest performance.

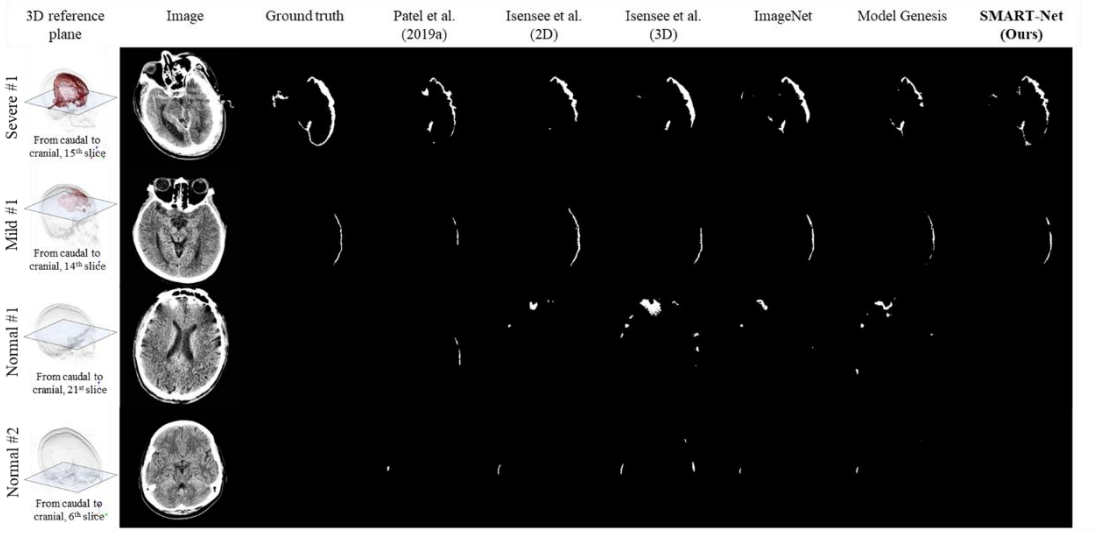

Figure 13. Comparison of volume-level segmentation results for severe (#1: 199.45 mL) and mild (#1: 11.99 mL) ICH cases, and two normal (#1: beam hardening artifact, #2: no lesion) cases. Owing to the nature of sporadic cerebral hemorrhage, the slice-level comparison is more appropriate than the 3D rendering visualization comparison, thus, we present the segmentation result of volume-level at the slice level.

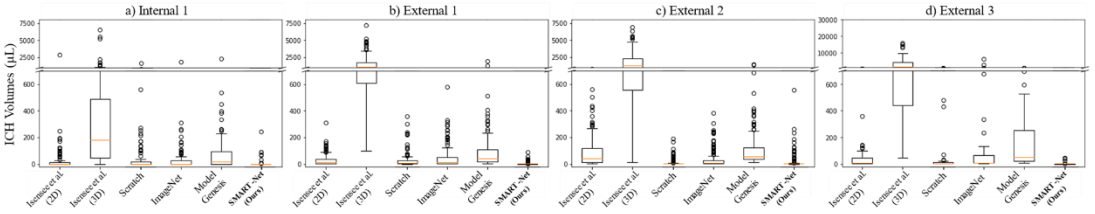

Figure 14. Comparisons of false-positive reduction performance on the normal cases in volume-level segmentation using box plots on four test sets.

3.2.2 Performance of ablation studies on pretext task

The quantitative outcomes of pretext task ablation studies in volume-level classification and segmentation tasks are presented in **Table 7**. Triple pretext tasks with consistency loss yielded the best results for both the classification and segmentation of targets. As demonstrated in **Table 7**, models trained using a single pretext task in the upstream task and the end-to-end scratch technique are more unstable on external test sets than models trained using multiple pretext tasks. In addition, the model employing consistency loss in the upstream job has a false-positive rate that is lower than 10μL and lower than other models. In classification, significant differences between ablation experiments and SMART-Net-CLS/SEG were more prominent than in segmentation.

Table 7. Quantitative results of volume-level target classification and segmentation tasks for a comparative analysis with ablation studies on four test sets.

Classification results

	Internal 1				External 1				External 2				External 3			
Pretext task	AUC	F1	SEN	SPE	AUC	F1	SEN	SPE	AUC	F1	SEN	SPE	AUC	F1	SEN	SPE
Only CLS	0.947*	0.776	0.897	0.880	0.920***	0.759	0.860	0.784	0.943	0.326	0.955	0.576	0.902*	0.790	0.941	0.615
Only SEG	0.456***	0.367	1.000	0.000	0.644***	0.515	1.000	0.000	0.638***	0.178	1.000	0.000	0.381***	0.636	1.000	0.000
Only REC	0.714***	0.462	0.724	0.590	0.627***	0.522	0.628	0.586	0.576***	0.190	0.909	0.167	0.676***	0.633	0.912	0.154
CLS+SEG	0.956**	0.800	0.897	0.900	0.943*	0.826	0.884	0.864	0.911	0.642	0.712	0.916	0.786***	0.734	0.853	0.590
CLS+SEG w/ consist	0.959	0.831	0.931	0.910	0.957	0.841	0.860	0.901	0.936	0.653	0.727	0.946	0.936	0.800	0.941	0.641
CLS+REC	0.975	0.862	0.966	0.920	0.935**	0.831	0.860	0.889	0.910	0.607	0.773	0.916	0.897	0.762	0.941	0.538
SEG+REC	0.909*	0.716	0.828	0.860	0.880***	0.621	0.953	0.407	0.803**	0.263	0.818	0.522	0.829**	0.694	1.000	0.231
CLS+SEG+REC	0.984	0.871	0.931	0.940	0.958	0.845	0.919	0.864	0.934	0.654	0.773	0.936	0.945	0.825	0.941	0.667
SMART-Net-CLS	0.988	0.933	0.966	0.970	**0.966**	0.866	0.942	0.877	0.947	0.706	0.818	0.946	0.953	0.842	0.971	0.744

Segmentation results

	Internal 1		External 1		External 2		External 3		Average	
Pretext task	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)	DSC (ICH)	FPV↓ (Normal)
Only CLS	0.594*	40.4	0.573***	83.3**	0.489*	15.1	0.510	21.1*	0.542	40.0
Only SEG	0.618	18.1	0.607	13.8*	0.514*	16.3	0.475*	40.3	0.554	22.1
Only REC	0.612	77.9**	0.601	104.3***	0.554	106.3***	0.441**	357.7***	0.552	156.5
CLS+SEG	0.604*	23.8*	0.593	73.7***	0.512	21.6*	0.503	117.9	0.553	59.3
CLS+SEG w/ consist	0.612*	7.7	0.587*	5.1	0.516	9.3	0.511	5.2	0.557	6.8
CLS+REC	0.604	141.1**	0.561*	138.3***	0.515	139.0***	0.377***	464.4***	0.514	220.7
SEG+REC	0.597	30.5	0.591	10.6**	0.502*	11.6	0.507*	9.8	0.549	15.6
CLS+SEG+REC	0.622	12.7	0.595	9.3**	0.520	12.9	0.517	7.0	0.564	10.5
SMART-Net-SEG	0.642	4.9	0.602	2.2	0.537	7.9	0.519	3.1	0.575	4.5

Note: ↓, the lower the value, the better performance; consist., consistency loss; p-values were calculated between SMART-Net vs. others: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. The red text represents the best performance, and the blue text represents the second-highest performance.

4. Discussion

There are difficulties because ICH's heterogeneity makes training difficult, high sensitivity and specificity are required, volume predictions are costly, and there is an external data vulnerability. In this study, we devised and evaluated SMART-Net to overcome these difficulties. **Figure 11** shows that feature extractor representations differ by multi-pretext task combination. Our three multi-pretext tasks with consistency loss improved encoder representation. Our training technique extracts global pretext features, allowing encoders to handle features in multiple ways. More upstream multitasking led to more significant

activations. Adding consistency loss raised the two heads' concurrence rate and transferability (see **Table 5**). In **Figure 12**, consistency loss can help refine the feature extractor's representation because it stabilizes learning by regulating discrepancies between the classification and segmentation heads by forcing them to find the same meaning from a binary task perspective. **Table 5** shows that while upstream performance is poor, downstream performance can be improved, which is similar to Salman *et al.* [55]. Multi-pretext tasks and consistency loss in the upstream task improved transfer learning. **Table 6** shows that transfer learning with 3D operators and the multi-pretext feature extractor enhances volume-level segmentation and classification. SMART-Net-CLS and SMART-Net-SEG had the best classification and segmentation results in four test sets. Due to negative transfer, Model Genesis and Autoencoder performed worse than the scratch model in volume-level ICH classification. In segmentation, Isensee *et al.*'s 3D version had a significant false-positive rate despite good sensitivity. Other models were external data-dependent and unstable. As seen in **Table 7**, the encoder trained with more upstream pretext tasks improved downstream performance and became more robust to external input. These results might be used to add SMART-Net to an emergency triage system to prioritize severe cases and the CAD system to exclude real normal cases from radiologists' reading lists, reflecting a high negative predicted value. Our technique has drawbacks despite enhanced performance. MTL requires task balancing [9, 10]. Even in our research, pretext imbalances could impair upstream analyses. Second, because our research mainly relies on feature extractors, different backbone models can produce different outcomes. Third, segmentation masks differ by radiologist's experience and CT vendor.

5. Summary

We presented a multi-task representation learning network for ICH volume-level classification and segmentation in NCCT. Initially, we constructed a robust feature extractor using multi-pretext tasks and incorporated consistency loss for the concurrence of classification (*CLS*) and segmentation (*SEG*) heads. In order to tackle the volume-level ICH challenges, we paired a pre-trained feature extractor with the best 3D operator using ablation research. By investigating the interaction between many pretext tasks, we discovered that the combination

of pretext tasks impacts the performance of the feature extractor. The proposed framework was derived from our exploratory studies and subsequently evaluated using one internal and three external test sets that reflected the actual incidence of ICH in emergency care settings. In comparison to existing approaches, our framework obtained state-of-the-art volume-level outcomes for both ICH classification and segmentation performance, as indicated by the evaluation results. Consequently, SMART-Net may be practically useful in actual clinical settings.

C. Application of deep representation learning to low-dose CT denoising task

1. Background and Objective

Computed tomography (CT) is one of the most important diagnostic modalities utilized in modern medical facilities. X-rays have the potential to produce genetic damage and cancer proportional to the radiation dose [56, 57]. ALARA (As Low As Reasonably Achievable) principles are commonly utilized in CT imaging to avoid unintended side effects [58]. Reducing the radiation dose increases noise and artifacts in the generated images, which may affect the diagnostic confidence and accuracy of radiologists. Consequently, substantial effort has been devoted to creating enhanced image reconstruction or image processing approaches for low-dose CT (LDCT). Even though many deep learning algorithms have been applied to LDCT denoising in recent years, radiologists still face challenges such as over smoothness and visual discomfort. In this paper, we propose a multi-task discriminator-based generative adversarial network (MTD-GAN) capable of performing simultaneously three visual tasks (classification, segmentation, and re-construction) in a discriminator. To stabilize GAN training, we present two novel loss functions referred to as non-difference suppression (NDS) loss and reconstruction consistency (RC) loss. In addition, we employ a fast Fourier transform with convolution block (FFT-Conv Block) in the generator to utilize both high- and low-frequency characteristics. Our model has been tested by pixel-space and feature-space based metrics in the head and neck LDCT denoising task, and the findings demonstrate that it outperforms the state-of-the-art denoising algorithms statistically and qualitatively.

2. Materials and Methods

2.1 Dataset

The head and neck CT denoising dataset was gathered by scanning the database of the Asan Medical Center in the Republic of Korea for patients who underwent consecutive CT scans between July 2020 and August 2020. **Table 8** summarizes LDCT denoising datasets. In CT data sets for 130 patients, 6,054 pairs of pictures (100 patients) were randomly selected as the training set, 845 pairs of images (15 patients) as the validation set, and 859 pairs of images (15 patients) as the test set. A reconstruction program [59] inserts Poisson random noise into quarter-dose LDCT pictures and normal-dose CT (NDCT) images to train the model. All CT images are 3mm thick and B30 kernel.

Table 8. Patient demographic information and medical characteristics in the LDCT denoising dataset.

Dataset		Slice (Patient)		Gender		Age	Spacing
		LDCT	NDCT	Male	Female	Mean±Std	Mean
Head & Neck CT	Train	6,054 (100)	6,054 (100)	38	62	58.9±19.8	x, y : 0.40 z : 3.00
	Valid	845 (15)	845 (15)	5	10	48.5±14.9	x, y : 0.40 z : 3.00
	Test	859 (15)	859 (15)	7	8	55.1±16.4	x, y : 0.40 z : 3.00

2.2 Multi-task discriminator GAN (MTD-GAN)

Here, we present information about MTD-GAN (see **Figure 15**). Our design incorporates both a discriminator and a generator. To differentiate between the denoiser's output and the NDCT image target, the discriminator performs three vision multitasks: reconstruction, segmentation, and classification. In addition, we added two losses for better representation and sensitivity regulation. The generator is constructed as a denoiser and can be utilized independently during testing. Considering the virtues of the Fourier domain, we incorporated the FFT-Conv Block into our generator to enhance performance. Our MTD-GAN is optimized by means of an adversarial technique.

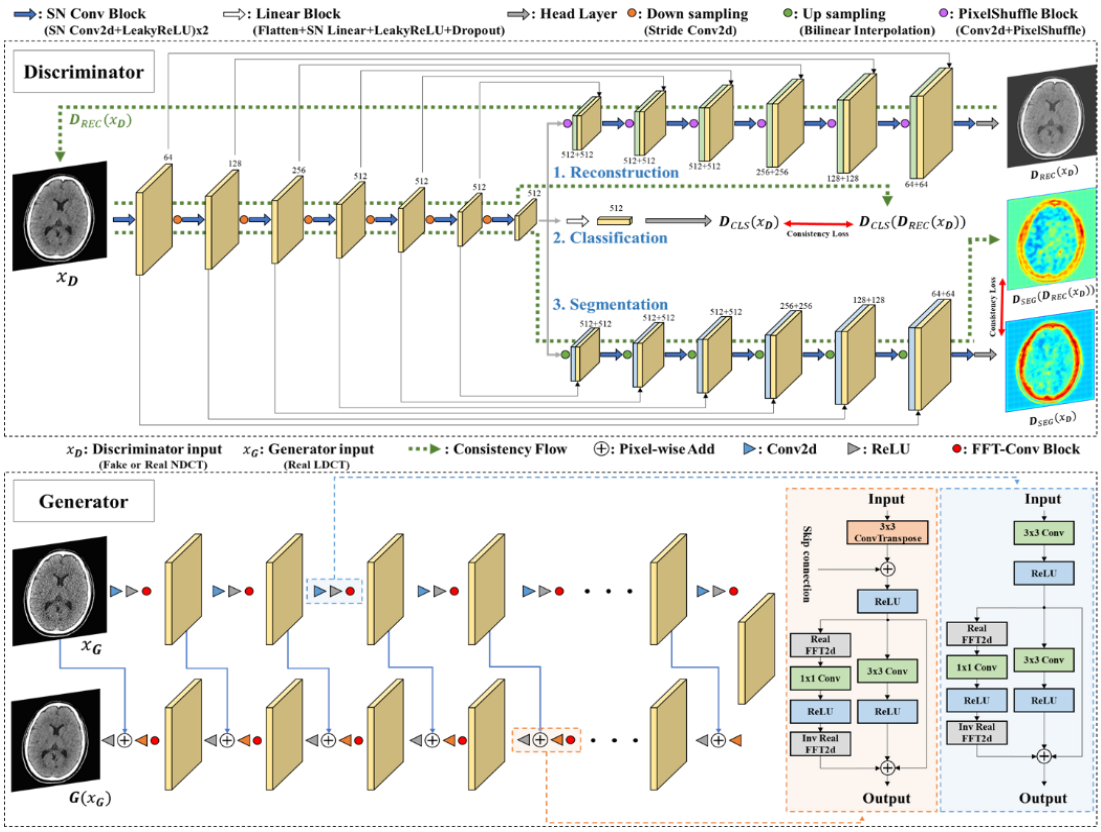

Figure 15. Schematic overview of the MTD-GAN framework. Our framework comprises a discriminator and generator: A discriminator is trained for distinguishing the real normal-dose and denoised images by three multi-tasks (classification, segmentation, reconstruction) with consistency loss using reconstruction output. In the generator, we adopted the RED-CNN as a base denoiser and added the FFT-Conv block at every layer for better denoising performance. Note: SN; spectral normalization [60].

2.2.1 Multi-task discriminator

A conventional discriminator is prone to forgetting prior samples since the distribution of synthetic samples moves as the generator constantly varies throughout training, resulting in an inability to retain a robust representation for identifying global and local image differences [61]. To address the issues, we present multi-task discriminators that perform classification, segmentation, and reconstruction. We implemented a hard parameter sharing architecture with a shared encoder and three types of task-specific layers for specified vision tasks in order to conduct successful multi-task learning for obtaining semantically denoised and normal-dose image features through diverse vision tasks. Our multi-task discriminator acquires a robust representation capable of characterizing both global and local changes between normal-dose

and denoised images (see **Figure 15**. discriminator part).

Reconstruction (REC). The first target-specific layer for REC leads the discriminator to do a reconstruction task, hence enhancing the discrimination and generalization skills of the classifier module. This REC task is done unsupervised to reconstruct the fake or actual NDCT image via the encoding-decoding process, allowing the network to better learn the semantic contextual representation by capturing image properties [31]. The REC loss was defined as the mean absolute error loss.

Classification (CLS). The second task-specific layer for CLS determines whether the image is fake or real, similar to conventional discriminators with a scalar value. This enables the network to learn the most discriminate difference by focusing on the global structure between fake and real images, which regularizes the generator accordingly. The CLS loss is defined as the LSGAN [62].

Segmentation (SEG). The third task-specific layer for SEG is used to determine if a picture is fake or real using a per-pixel confidence map, which allows the network to identify the difference in local features between normal-dose and denoised images. The SEG loss is also defined as the LSGAN.

2.2.2 Non-difference suppression loss and consistency loss

Non-difference suppression (NDS) loss. Due to the nature of the medical image, the regions of background and bone with the same LDCT and NDCT occupy a significant amount of the CT image, resulting in significant uncertainty when the discriminator makes a determination. In other words, according to the prior procedure [63], the background portions of NDCT and LDCT are identical, but the label values are calculated differently, resulting in the transmission of inaccurate information to the generator. To improve the stability of GAN training, we employ the NDS loss, which excludes regions without difference from the loss calculation in the difference mask between LDCT and NDCT images (see **Figure 16**). The NDS-SEG deficit is defined as follows:

$$L_{NDS-SEG} = L_{SEG} \times boolean(|\boldsymbol{I}_{LDCT} - \boldsymbol{I}_{NDCT}|), \quad (3)$$

where L_{SEG} is the loss of segmentation and $\boldsymbol{I}_{LDCT}$ and $\boldsymbol{I}_{NDCT}$ are the image of low-dose CT scans and normal-dose CT scans.

Reconstruction consistency (RC) loss. As illustrated in **Figure 15**, green flows, we presented a novel consistency regularization for stabilizing GAN training utilizing reconstruction output. Consistency regularization penalizes the discriminator's sensitivity to enhance the performance of GANs [64]. In this article, the discrepancy between the input and reconstructed images is examined in detail. The RC loss is defined as follows:

$$L_{Consist.} = \mathbb{E}_{x_D}[\|\boldsymbol{D}_{CLS}(x_D) - \boldsymbol{D}_{CLS}(\boldsymbol{D}_{REC}(x_D))\| + \|\boldsymbol{D}_{SEG}(x_D) - \boldsymbol{D}_{SEG}(\boldsymbol{D}_{REC}(x_D))\|], \quad (4)$$

where x_D is the input of the discriminator and $\boldsymbol{D}_{CLS}$, $\boldsymbol{D}_{SEG}$ and $\boldsymbol{D}_{REC}$ are the CLS, SEG, and REC layers of multi-task discriminator scans and normal-dose CT scans.

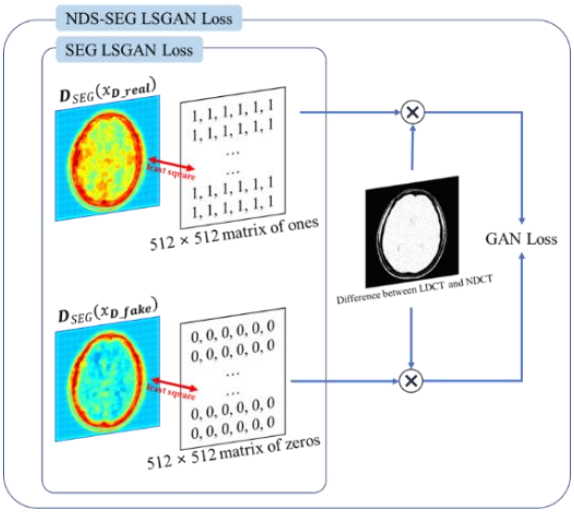

Figure 16. A concept of our NDS loss to LSGAN loss in segmentation task.

2.2.3 FFT-Generator

We chose the RED-CNN [65] base denoiser with 10 (de)convolutional layers at both the encoder and decoder to demonstrate the influence of the multitask-based discriminators. A convolutional operator excels at retrieving high-frequency details but may be incapable of examining low-frequency data. According to the spectral convolution theorem [66] in Fourier theory, updating a single value in the spectrum domain has a global effect on the image, giving it the benefit of a large receptive field [67]. Thus, we incorporated the FFT-Conv Block to gain the benefits of modeling both high- and low-frequency differences between hazy and clear features, as well as long- and short-term interactions. In addition to a standard spatial residual flow, as shown in **Figure 15**. generator section, the FFT-Conv Block translates initial spatial

characteristics into a spectrum domain, performs fast updates on spectral data, and then converts data back into the spatial domain. The original adversarial LSGAN loss for the generator is calculated as follows:

$$L_{adv} = \mathbb{E}_{x_G}[\boldsymbol{D}_{CLS}(x_{\boldsymbol{D\_fake}}) - 1]^2 + \mathbb{E}_{x_G}[\boldsymbol{D}_{SEG}(x_{\boldsymbol{D\_fake}}) - 1]^2, \quad (5)$$

where x_G and x_{D_fake} are the input of the generator and the fake input of the discriminator. We also applied the NDS loss to adversarial loss in the generator as follows:

$$L_{NDS-adv} = \mathbb{E}_{x_G}[\boldsymbol{D}_{CLS}(x_{\boldsymbol{D\_fake}}) - 1]^2 + \mathbb{E}_{x_G}[\boldsymbol{D}_{SEG}(x_{\boldsymbol{D\_fake}}) - 1]^2 \times boolean(|\boldsymbol{I}_{LDCT} - \boldsymbol{I}_{NDCT}|), \quad (6)$$

We used the Charbonnier loss [68] to achieve better performance on the denoising task. To further improve the fidelity of high-frequency details, we used the additional edge loss [69] to control the high-frequency components between the NDCT image and the denoised image.

3. Experiments and Results

Metrics. We employed peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and root mean square error (RMSE) for the quantitative evaluation of denoising image quality (RMSE). We further adopted perceptual loss (PL) [70], texture matching loss (TML) [71], and Fréchet inception distance (FID) [72] to assess feature space similarity because these metrics are susceptible to over smoothed images [63, 73]. RED-CNN [65], EDCNN [73], Restormer [74], CTformer [75], WGAN-VGG [70], MAP-NN [69], and DU-GAN [63] are state-of-the-art previous methods that are compared with our model.

Implementation details. For a fair comparison, the training settings, including optimizer, patch size, epochs, and learning rate scheduler, were identical.

1) Preprocessing: For network input, we employed a brain window that clips [0, 80] Hounsfield units (HU) and scales to [0, 1]. We clipped the foreground of the CT picture and then randomly extracted eight 64×64 patches from each 512×512 original image in each epoch.

2) Augmentation: Data augmentation is used to extend the dataset by randomly rotating (90 degrees, 180 degrees, or 270 degrees) and flipping a duplicate of the original image (up, down, left, or right).

3) Training configuration: To ensure a fair comparison under the constraints of restricted resources, the batch size of each trial was set to the maximum for the single GPU's memory. The network was initialized by a uniform Xavier, and an AdamW optimizer with a learning

rate of 1e-4, a warmup of 10 epochs, weight decay of 5e-4, and betas of 0.1 was utilized (0.9, 0.99). During the training of the polynomial learning rate schedule, the learning rate was lowered. The maximum number of epochs was 500. We empirically set the multi-task weights to be similar to compare the function of MTL in depth. We employed the projecting conflicting gradients (PCGrad) technique [76], which projects the gradient of one job onto the normal plane of the gradient of any other activity with a conflicting gradient, to eliminate gradient interference between several tasks.

3.1 Comparison of previous works

According to **Table 9**, our model has the best FID, PL, TML, and SSIM performance compared to other models, as well as competitive RMSE and PSNR performance. As shown in **Figure 17**, our MTD-GAN has the best qualitative performance in decreasing noise/artifacts and preserving clinically significant anatomical structures, resulting in pictures that are radiologist-friendly.

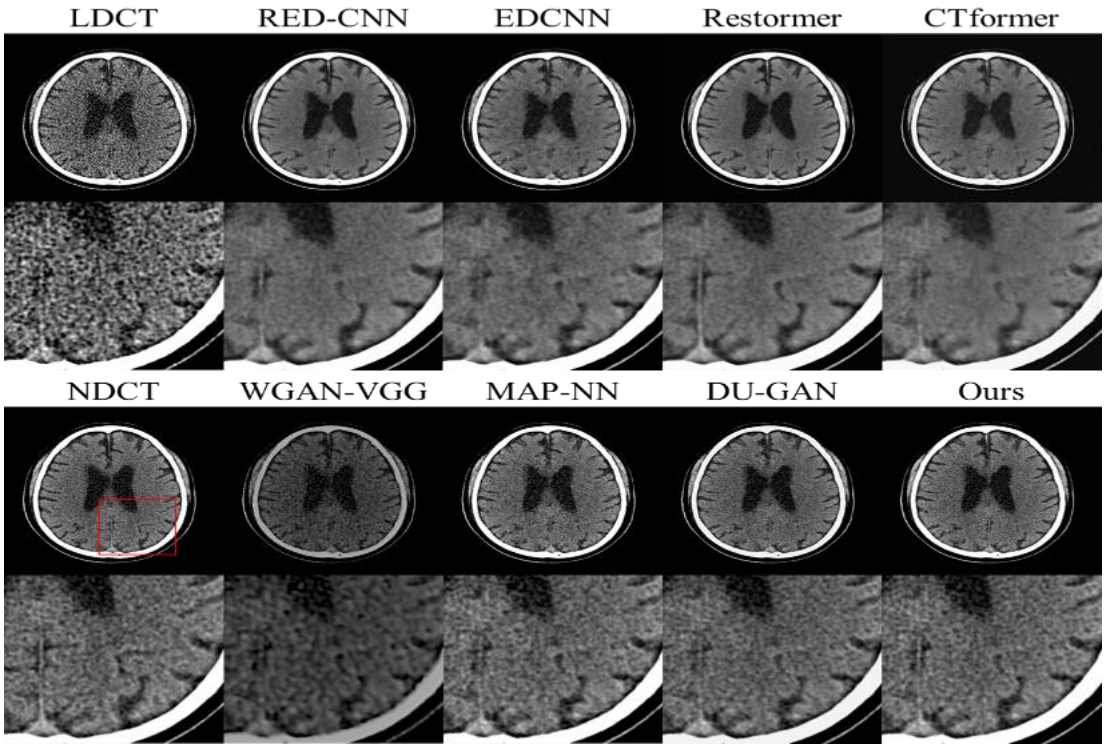

Figure 17. The denoising results of previous methods on the test set. The display window is [0, 80] HU. We zoomed ROI marked by the red square where have clinically meaningful area.

Table 9. Quantitative results for a comparative analysis with previous methods on the test set.

Previous works	FID↓	PL↓	TML↓	RMSE↓	PSNR↑	SSIM↑
RED-CNN [65]	34.7098	0.1493	15.5072	0.0330	33.3131	0.9012
EDCNN [73]	38.0299	0.1487	14.9475	0.0343	32.9481	0.8988
Restormer [74]	32.7898	0.1469	15.2586	**0.0323**	**33.5237**	0.9026
CTformer [75]	33.5319	0.1496	14.2172	0.0344	32.1764	0.8979
WGAN-VGG [70]	53.9002	0.2629	29.7508	0.0589	26.3789	0.8454
MAP-NN [69]	19.5778	0.1373	11.4723	0.0383	32.0321	0.8995
DU-GAN [63]	18.8156	0.1285	10.5245	0.0364	32.5533	0.9035
MTD-GAN (Ours)	**17.6212**	**0.1276**	**10.3559**	0.0369	31.2932	**0.9037**
GT (NDCT)	9.8371	0.0	0.0	0.0	100.0	1.0
Input (LDCT)	39.7624	0.1821	20.4434	0.0575	28.9550	0.8743

3.2 Comparison of ablation studies

To study the effect of the various MTD-GAN components, we conducted excessive ablation tests. To ensure a fair comparison, the training settings were fixed equally for a fair comparison and the number of epochs in the ablation study was up to 200 and the batch size was 20. **Table 10** and **Figure 18** demonstrate that both quantitative and qualitative performance improved as MTD-GAN elements were incrementally introduced.

Table 10. Quantitative results for a comparative analysis in ablation study on the test set.

Ablation Study	FID↓	PL↓	TML↓	RMSE↓	PSNR↑	SSIM↑
(a) G: RED-CNN / D: CLS-Discriminator	20.0615	0.1332	11.1556	**0.0366**	**32.5082**	0.9032
(b) + D: CLS&SEG-Discriminator	19.1339	0.1329	10.9765	0.0375	32.3491	0.9027
(c) + D: CLS&SEG&REC-Discriminator	18.3820	0.1324	10.9761	0.0372	32.4050	0.9025
(d) + D: L_{NDS}	18.2724	0.1329	11.0543	0.0369	32.4347	0.9028
(e) + D: $L_{\mathrm{Consist.}}$	18.2577	0.1329	11.0562	0.0370	**32.4502**	0.9026
(f) + G: FFT-Generator	**17.9308**	**0.1275**	**10.2946**	0.0371	31.2766	**0.9035**
(g) + PCGrad (Ours)	**17.6212**	**0.1276**	**10.3559**	**0.0369**	31.2932	**0.9037**
High	9.8355	0.0	0.0	0.0	100.0	1.0
Low	39.7582	0.1821	20.4448	0.0575	28.9550	0.8743

Note: The red color represents the best, and the blue color represents the second-highest performance.

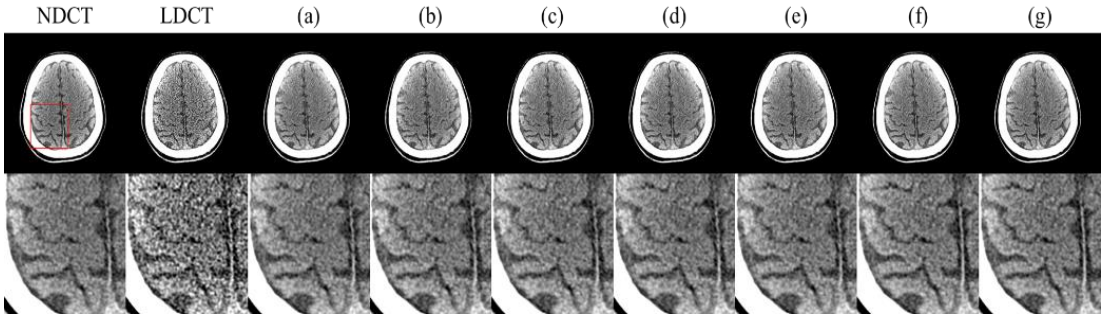

Figure 18. The denoising results in the ablation study on the test set. The display window is [0, 80] HU. (a)-(g) are from the same as the label shown in **Table 10**.

4. Discussion

For radiologists, excessive smoothness and ocular discomfort remain. To solve these difficulties, we proposed MTD-GAN. We discovered that feature-space measurements offered the most reasonable credibility to radiologists, but pixel-based metrics are sensitive to hazy images. In actuality, the radiologist mistook the outputs of the model with the highest FID score performance for actual LDCT images. **Table 9** shows that our model beats others in FID, PL, TML, and SSIM and is competitive in RMSE and PSNR. MTD-GAN maintains clinically significant anatomical features, giving radiologist-friendly pictures (see **Figure 17**). In CNNs, particularly RED-CNN and EDCNN, the output has a high PSNR but is over-smoothed. In transformers, the Restormer outputs have the highest PSNR and RMSE, but they're too smooth for radiologists. CTformer contains boundary artifacts in the stitched image because of setting input patch sizes. The WGAN-VGG has the worst image quality because of unstable training, while the MAP-NN and DU-GAN create high-fidelity denoised images. In the ablation studies, **Table 10** shows that multitasking the discriminator improves the denoiser's FID. A multitask-based discriminator can both provide per-pixel feedback to the denoiser and focus on the global structure at a semantic level. LDCT and NDCT image contexts can be distinguished. The FID performance improved when confusion was removed from the segmentation LSGAN loss and strong perturbation training was utilized to recover the discriminator. Through the FFT-Conv Block, blending low-and high-frequency data helps generate realistic images, demonstrating that performance can be improved by optimizing the generator. These MTD-GAN results could be applied in a real medical context to reduce patient risk. Despite the better

performance, our technique has flaws. Task balancing is critical in MTL. Even in our research, we used PCGrad. However, it's still insufficient and unstable, thus multitasking experiments are done. Inspired by task balancing studies, we may explore fitted weights among multitasks in future trials. Second, feature-space metrics like FID, PL, and TML can lead to inadequate task evaluations. These indicators used a model pre-trained by nature photos, which is improper for medical usage [17, 18]. Many studies are currently underway, and feature-space metrics that use pre-trained medical images will emerge and be enhanced in the future.

5. Summary

In this paper, a unique GAN method employing a discriminator based on multiple tasks is created for clinical applications. This is, as far as we are aware, the first study to apply three vision multitasks to the discriminator in the LDCT denoising task. Primarily, we make three contributions: (1) An architecture based on several tasks strengthens the discriminator, directing the generator to synthesize images with global and local realism. (2) For stable GAN training, the NDS loss makes the discriminator robust by removing confusing areas in the segmentation task, while the RC loss utilizing the generated reconstruction enables the network to acquire more contextual knowledge. (3) The addition of the FFT-Conv Block to the LDCT denoising operation permits the generator to use both high- and low-frequency components and to increase the receptive field, resulting in images with a higher level of detail. Experimentally, the suggested MTD-GAN achieves superior denoising performance compared to previous approaches and has the potential for clinical use.

Conclusion

In this study, three experiments were performed to evaluate representation learning, particularly inductive transfer learning and multitask learning, in the medical domain. In the first study, sequential transfer learning was utilized to predict pediatric diagnoses, resulting in enhanced performance and distinct ROI activation. In the second study, multi-task learning was applied to develop a robust feature extractor for the brain hemorrhage identification task, resulting in improved performance even when using external data. In the third study, multi-task learning was applied to improve the discriminator for low-dose CT denoise tasks, thereby

stabilizing GAN training. As evidenced by these findings, it is preferable to apply inductive transfer learning of representation learning rather than learn a model from scratch in medical domain tasks.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35, 1798-1828 (2013)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
3. Erhan, D., Courville, A., Bengio, Y., Vincent, P.: Why does unsupervised pre-training help deep learning? In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 201-208. JMLR Workshop and Conference Proceedings, (Year)
4. Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V.: Deep neural nets as a method for quantitative structure–activity relationships. Journal of chemical information and modeling 55, 263-274 (2015)
5. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. Science China Technological Sciences 63, 1872-1897 (2020)
6. Shurrab, S., Duwairi, R.: Self-supervised learning methods and applications in medical imaging analysis: A survey. arXiv preprint arXiv:2109.08685 (2021)
7. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE transactions on pattern analysis and machine intelligence (2021)
8. Wu, S., Zhang, H.R., Ré, C.: Understanding and improving information transfer in multi-task learning. arXiv preprint arXiv:2005.00944 (2020)
9. Liu, S., Liang, Y., Gitter, A.: Loss-balanced task weighting to reduce negative transfer in multi-task learning. In: Proceedings of the AAAI conference on artificial intelligence, pp. 9977-9978. (Year)
10. Chen, Z., Badrinarayanan, V., Lee, C.-Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International conference on machine learning, pp. 794-803. PMLR, (Year)
11. Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? In: International Conference on Machine Learning, pp. 9120-9132. PMLR, (Year)
12. Vandenhende, S., Georgoulis, S., De Brabandere, B., Van Gool, L.: Branched multi-task networks: deciding what layers to share. arXiv preprint arXiv:1904.02920 (2019)
13. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4106-4115. (Year)
14. Mustafa, B., Loh, A., Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S.M., Sieniek, M., Winkens, J., Liu, Y., Bui, P.: Supervised transfer learning at scale for medical imaging. arXiv preprint arXiv:2101.05913 (2021)
15. Ke, A., Ellsworth, W., Banerjee, O., Ng, A.Y., Rajpurkar, P.: CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In: Proceedings of the Conference on Health, Inference, and Learning, pp. 116-124. (Year)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. (Year)
17. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems 32, (2019)
18. Alzubaidi, L., Fadhel, M.A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., R. Oleiwi, S.: Towards a better understanding of transfer learning for medical imaging: a case study. Applied Sciences 10, 4523 (2020)
19. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan,

I., Pereira, L.A., Sousa, R.T., Abdala, N.: The RSNA pediatric bone age machine learning challenge. Radiology 290, 498 (2019)
20. Iannaccone, G.: WW Greulich and SI Pyle: Radiographic atlas of skeletal development of the hand and wrist. I volume-atlante di 256 pagine. Stanford University Press, Stanford, California, 1959. Acta geneticae medicae et gemellologiae: twin research 8, 513-513 (1959)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, (2012)
22. Naranje, S.M., Erali, R.A., Warner, W.C., Sawyer, J.R., Kelly, D.M.: Epidemiology of pediatric fractures presenting to emergency departments in the United States. Journal of Pediatric Orthopaedics 36, e45-e48 (2016)
23. Ravichandiran, N., Schuh, S., Bejuk, M., Al-Harthy, N., Shouldice, M., Au, H., Boutis, K.: Delayed identification of pediatric abuse-related fractures. Pediatrics 125, 60-66 (2010)
24. Lones, M.A.: How to avoid machine learning pitfalls: a guide for academic researchers. arXiv preprint arXiv:2108.02497 (2021)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618-626. (Year)
26. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921-2929. (Year)
27. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research 9, (2008)
28. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 837-845 (1988)
29. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing 39, 355-368 (1987)
30. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information 11, 125 (2020)
31. Kyung, S., Shin, K., Jeong, H., Kim, K.D., Park, J., Cho, K., Lee, J.H., Hong, G.-S., Kim, N.: Improved performance and robustness of multi-task representation learning with consistency loss between pretexts for intracranial hemorrhage identification in head CT. Medical Image Analysis 102489 (2022)
32. Parizel, P., Makkat, S., Van Miert, E., Van Goethem, J., Van den Hauwe, L., De Schepper, A.: Intracranial hemorrhage: principles of CT and MRI interpretation. European radiology 11, 1770-1783 (2001)
33. Jha, S.: Value of triage by artificial intelligence. Academic radiology 27, 153-155 (2020)
34. Hu, S., Coupé, P., Pruessner, J., Collins, L.: Validation of appearance-model based segmentation with patch-based refinement on medial temporal lobe structures. In: MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion, pp. 28-37. (Year)
35. Voter, A.F., Meram, E., Garrett, J.W., John-Paul, J.Y.: Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. Journal of the American College of Radiology 18, 1143-1152 (2021)
36. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International conference on medical image computing and computer-assisted intervention, pp. 421-429. Springer, (Year)
37. Chen, S., Ma, K., Zheng, Y.: Transfer learning for 3d medical image analysis. arXiv preprint arXiv (1904)
38. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1874-1883. (Year)
39. Feng, R., Zhou, Z., Gotway, M.B., Liang, J.: Parts2Whole: Self-supervised contrastive learning via

reconstruction. Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, pp. 85-95. Springer (2020)
40. Amyar, A., Modzelewski, R., Li, H., Ruan, S.: Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. Computers in Biology and Medicine 126, 104037 (2020)
41. Song, G., Liu, Y., Wang, X.: Revisiting the sibling head in object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11563-11572. (Year)
42. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., Fu, Y.: Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10186-10195. (Year)
43. Qin, R., Liu, Q., Gao, G., Huang, D., Wang, Y.: Mrdet: A multi-head network for accurate oriented object detection in aerial images. arXiv preprint arXiv:2012.13135 (2020)
44. Liu, P., Qiu, X., Huang, X.: Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742 (2017)
45. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences 114, 3521-3526 (2017)
46. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
47. Fenster, A., Chiu, B.: Evaluation of segmentation algorithms for medical imaging. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 7186-7189. IEEE, (Year)
48. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis 67, 101840 (2021)
49. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science 313, 504-507 (2006)
50. Nguyen, N.T., Tran, D.Q., Nguyen, N.T., Nguyen, H.Q.: A CNN-LSTM architecture for detection of intracranial hemorrhage on CT scans. medRxiv (2020)
51. Patel, A., Van De Leemput, S.C., Prokop, M., Van Ginneken, B., Manniesing, R.: Image level training and prediction: intracranial hemorrhage identification in 3D non-contrast CT. Ieee Access 7, 92355-92364 (2019)
52. Singh, S.P., Wang, L., Gupta, S., Gulyas, B., Padmanabhan, P.: Shallow 3D CNN for detecting acute brain hemorrhage from medical imaging sensors. IEEE Sensors Journal 21, 14290-14299 (2020)
53. Patel, A., Schreuder, F.H., Klijn, C.J., Prokop, M., Ginneken, B.v., Marquering, H.A., Roos, Y.B., Baharoglu, M., Meijer, F.J., Manniesing, R.: Intracerebral haemorrhage segmentation in non-contrast CT. Scientific reports 9, 1-11 (2019)
54. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203-211 (2021)
55. Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., Madry, A.: Do adversarially robust imagenet models transfer better? Advances in Neural Information Processing Systems 33, 3533-3545 (2020)
56. Brenner, D.J., Hall, E.J.: Computed tomography—an increasing source of radiation exposure. New England journal of medicine 357, 2277-2284 (2007)
57. de Gonzalez, A.B., Darby, S.: Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. The lancet 363, 345-351 (2004)
58. Valentin, J., Boice Jr, J., Clarke, R., Cousins, C., Gonzalez, A., Lee, J., Lindell, B., Meinhold, C., Mettler Jr, F., Pan, Z.: Published on behalf of the International Commission on Radiological Protection. (2007)
59. Kramer, M., Ellmann, S., Allmendinger, T., Eller, A., Kammerer, F., May, M.S., Baigger, J.F., Uder, M., Lell, M.M.: Computed tomography angiography of carotid arteries and vertebrobasilar system: a simulation study for radiation dose reduction. Medicine 94, (2015)
60. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
61. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8207-8216. (Year)

62. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2794-2802. (Year)
63. Huang, Z., Zhang, J., Zhang, Y., Shan, H.: DU-GAN: Generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising. IEEE Transactions on Instrumentation and Measurement 71, 1-12 (2021)
64. Zhang, H., Zhang, Z., Odena, A., Lee, H.: Consistency regularization for generative adversarial networks. arXiv preprint arXiv:1910.12027 (2019)
65. Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J., Wang, G.: Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE transactions on medical imaging 36, 2524-2535 (2017)
66. Katznelson, Y.: An introduction to harmonic analysis. Cambridge University Press (2004)
67. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems 33, 4479-4488 (2020)
68. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14821-14831. (Year)
69. Shan, H., Padole, A., Homayounieh, F., Kruger, U., Khera, R.D., Nitiwarangkul, C., Kalra, M.K., Wang, G.: Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. Nature Machine Intelligence 1, 269-276 (2019)
70. Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G.: Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE transactions on medical imaging 37, 1348-1357 (2018)
71. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE international conference on computer vision, pp. 4491-4500. (Year)
72. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30, (2017)
73. Liang, T., Jin, Y., Li, Y., Wang, T.: Edcnn: Edge enhancement-based densely connected network with compound loss for low-dose ct denoising. In: 2020 15th IEEE International Conference on Signal Processing (ICSP), pp. 193-198. IEEE, (Year)
74. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728-5739. (Year)
75. Wang, D., Fan, F., Wu, Z., Liu, R., Wang, F., Yu, H.: CTformer: Convolution-free Token2Token Dilated Vision Transformer for Low-dose CT Denoising. arXiv preprint arXiv:2202.13517 (2022)
76. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in Neural Information Processing Systems 33, 5824-5836 (2020)

Abstract (with Korean)

딥 러닝은 다양한 분야에서 사용되었고, 인상적인 결과를 가져왔다. 그러나 훈련 데이터 부족, 촬영장비 또는 도메인 차이로 인한 성능 저하, 방사선 사진 및 CT 스캔과 같은 고화질 영상 및 다른 의료 센터에 대한 견고성 등 의료 영역에서 딥 러닝을 적용하는 데 여전히 어려움이 있다. 이러한 문제를 해결하기 위해 네트워크에서 파생된 기능을 능숙하게 처리하는 연구인 표현 학습의 귀납적 전이 학습 (inductive transfer learning), 즉 순차 전이 학습 (sequential transfer learning)과 다중 작업 학습 (multi-task learning)이 활발히 연구되었다. 본 연구에서는 표현 학습, 특히 순차 전이 학습과 멀티태스킹 학습을 포함한 귀납적 전이 학습이 의료 영역에 어떤 영향을 미치는지 확인하기 위해 세가지 실험을 수행했다: '소아 진단에 대한 심층 표현에 대한 연구', '뇌출혈 진단에 대한 심층 표현 연구', 그리고 '저선량 CT 노이즈 제거 작업에 대한 심층 표현 연구'. 첫 번째 연구에서는 성능 향상을 위해 순차적 전이 학습을 적용했다. 우리는 방사선 사진 뷰를 기반으로 레이블을 사용하여 클래스 균형 소아 방사선 사진 데이터 세트, PedXnet 을 구성하고 감독된 표현 (supervised representation)을 개발했다. 골절 분류 및 골 연령 평가를 포함한 소아 다운 스트림 작업을 통해 표현 학습의 효과를 검증했다. 그 결과, Model-PedXnets 의 전이 학습은 Model-Baseline 의 것에 비해 향상된 정량적 성능을 보여주었다. Model-PedXnets 는 Model-ImageNet 과 동등하고 경우에 따라서는 성능이 향상되었습니다. 특히 Model-PedXnets 는 가장 의미 있는 ROI 에 초점을 맞췄다. 두 번째 연구에서는 견고성을 위해 다중 작업 학습을 적용했다. 우리는 두개내 출혈 (ICH)의 진단을 위해 감독된 다중 작업 지원 표현 전달 학습 네트워크 (SMART-Net)를 제안했다. 제안된 프레임워크는 업 스트림 및 다운 스트림 구성 요소로 구성된다. 업 스트림에서 모델의 가중치 공유 인코더는 슬라이스 레벨 다중 사전 정의 작업 (pretext task)을 수행하여 글로벌 기능을 캡처하는 강력한 기능 추출기로 훈련된다. 다운 스트림에서, 전송 학습은 환자 단위 작업을 위해 사전 훈련된 인코더와 3D 연산자를 사용하여 수행되었다. 네 가지 테스트 세트를 기반으로 한 실험 결과는 SMART-Net 이 이전 방법에 비해 볼륨 레벨 ICH 분류 및 세분화 측면에서 견고성과 성능이 우수함을 보여준다. 세 번째 연구에서는 판별기 학습의 안정을 위해 멀티태스킹 학습을 적용했다. 우리는 저 선량 컴퓨터 단층 촬영 (LDCT) 노이즈 제거 모델을 더 잘 정규화 하기 위해 다중 작업 판별

기 GAN (MTD-GAN)을 제안한다. 이 모델은 GAN 프레임워크에서 판별기에 대한 세 가지 다중 작업을 활용하여 노이즈 제거 이미지와 정상 선량 이미지 사이의 전역 및 로컬 차이를 모두 학습한다. 또한 이미지와 푸리에 도메인을 모두 사용하여 미세한 구조적 세부 사항을 학습할 수 있는 FFT-Generator 를 제안하여 CT 노이즈 제거 작업을 개선하였다. 결과적으로, MTD-GAN 은 정량적 결과와 정성적 결과에서 이전 방법보다 방사선 전문가 친화적인 성능을 달성한다. 세 가지 연구 모두 의료 영역에서 순차적 전이 학습과 다중 작업 학습을 포함한 표현 학습이 성능을 향상시키고 의미론적 특징을 추출하고 외부 데이터에 대해 모델을 견고하게 만들 수 있음을 확인했다. 의료 영역에 인공지능을 적용하는 미래에는 단순히 스크래치 모델을 훈련시켜 성능을 평가하기보다는 표현 학습을 고려해야 한다.