IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 January 2024, accepted 22 January 2024, date of publication 2 February 2024, date of current version 8 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3360272

RESEARCH ARTICLE

A Deep Learning Harmonization of Multi-Vendor MRI for Robust Intervertebral Disc Segmentation

CHAEWOO KIM[1,2], SANG-MIN PARK[3], SANGHOON LEE[3], AND DEUKHEE LEE[1,2,4], (Member, IEEE)

[1]Center for Healthcare Robotics, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea
[2]Division of AI-Robotics, Korea University of Science and Technology, Seoul 02792, Republic of Korea
[3]Spine Center and Department of Orthopaedic Surgery, Seoul National University College of Medicine, Seoul National University Bundang Hospital, Seongnam 13620, Republic of Korea
[4]Yonsei-KIST Convergence Research Institute, Yonsei University, Seoul 03722, Republic of Korea

Corresponding author: Deukhee Lee (dkylee@kist.re.kr)

This work was supported in part by KIST Institutional Programs under Project 2E32983; and in part by the Korea Medical Device Development Fund grant funded by the Korean Government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health and Welfare, and the Ministry of Food and Drug Safety) under Project 1711138281 and Project RS-2020-KD000145.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Seoul National University Bundang Hospital under Application No. B-2304-822-101.

ABSTRACT Magnetic resonance imaging (MRI) provides enhanced soft tissue contrast and high spatial resolution. However, the relationship between intensity values among soft tissues in MRI is inconsistent, even when obtained under the same conditions (e.g., vendors and acquisition protocols). This inconsistency hinders accurate medical image segmentation and disease classification. Therefore, we propose a framework to harmonize multi-vendor MRI using a novel radiomics approach for robust segmentation. The proposed model comprises a cycle-consistent adversarial network (CycleGAN)-based network and a segmentation network. The CycleGAN-based network harmonizes MRI with the support of a radiomics-based method (radiomic feature (RF) loss function newly designed for this study). The segmentation network encourages the CycleGAN-based network to enhance intervertebral disc (IVD) segmentation features using dice loss functions during harmonization. Furthermore, publicly available datasets and diverse MRI scans provided by a collaborating hospital were used to make our model more robust to MRI variability. The proposed model was evaluated for segmentation using the Dice coefficient, intersection-over-union (IoU), F1 score, precision, and recall. It outperformed other segmentation methods (Dice = 0.920, IoU = 0.853, F1 score = 0.920, precision = 0.940, and recall = 0.902), even on diverse test datasets with disease information. The harmonization performance was assessed using the relative error of the RF values between the target (standard) and harmonized data. It achieved the four best scores (≈ 0) among the five features in a relative error of RF compared to other harmonization methods (e.g., conventional histogram-based method and deep learning model).

INDEX TERMS Harmonization, magnetic resonance imaging, radiomics, segmentation.

I. INTRODUCTION

A. RESEARCH BACKGROUND

Magnetic resonance imaging (MRI) is a popular medical imaging technique that detects radio signals from protons in tissues using a magnetic field. Protons in different tissues produce distinct signals due to their varying proton densities. These MR signals thus provide high soft tissue contrast and submillimeter spatial resolution without radiation exposure, in contrast to X-ray and computerized tomography (CT) scans. Owing to this advantage, MRI has been used for a wide range of diagnoses (e.g., tumors in the brain or breast,

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

inflammation in the blood vessels, and disc disease in the spine). However, owing to the sensitivity of the inhomogeneous magnetic field during MRI acquisition, the radiomic features of MRI are highly dependent on inter-vendor variability and acquisition parameters (e.g., spacing between slices, slice thickness, pixel spacing, repetition time, echo time, and magnetic field strength) [1]. Consequently, the intensity values in MR images do not provide consistently reliable physical and biological data, unlike CT scans, where the radiodensity of tissues determines pixel intensity by the Hounsfield Unit (HU). This variability can significantly affect the accuracy of diagnosis, prognosis, disease monitoring, and clinical applications of MRI [2]. Accordingly, the emphasis on harmonizing multi-vendor MRI has increased to address these challenges [3], [4], [5], [6], [7].

For years, conventional statistical methods have been widely used for intensity normalization (e.g., Z-score, WhiteStripe, and Nyúl intensity normalization). These methods can yield reliable outcomes when the inter-pixel relationship is linear. However, they might be insufficient for MRI studies because of the nonlinear relationship between intensity values of soft tissues in MR images [8]. In contrast, deep learning (DL) algorithms can autonomously extract feasible features and learn from them. They can also extract high-level features to overcome inter-pixel nonlinearity in MRI. Consequently, DL approaches have been extensively developed to harmonize multi-vendor MRI in various regions of interest (ROIs), such as the brain, breast, knee, and prostate [1], [2], [4], [8]. Despite this, the harmonization of intervertebral disc (IVD) MRI has received relatively less attention compared to brain and breast MRI or CT scans [9], [10]. Nonetheless, MRI variability is also crucial in clinical outcomes in the lumbar spine. For instance, in image-guided surgery (IGS), this variability is vital. Specifically, the accuracy of IVD segmentation for visualization of the surgical target is increased by vendor-invariant MR images in the IGS, thereby improving the safety and precision of surgery [11].

Radiomics, which was recently highlighted in medical image analysis, is a numerical analytical method used to uncover underlying features (e.g., inter-pixel relationships) in a radiologic image [12]. It plays a crucial role in supporting clinical diagnosis and prognosis, as radiomic features (RFs) extracted from medical images can reveal informative characteristics not visible to the human eye. However, RF extraction is influenced by various image acquisition factors, including modality, vendor, acquisition protocol, and site [13], [14], [15], [16]. Therefore, harmonizing MRI intensity is imperative in radiomics to achieve reliable outcomes and ensure reproducibility [15], [16]. Moreover, the features extracted through radiomics are highly high-dimensional. Some RFs among all features (hundreds and thousands of features) may be redundant for specific studies. In other words, some RFs do not reflect relevant meanings in the medical images used in particular studies. Thus, a feature selection procedure is necessary because this procedure (dimension reduction) alleviates the curse of dimensionality [17] and influences the performance of DL networks based on the type of RF selected.

For these reasons, we propose a DL model to harmonize multi-vendor lumbar spine MRI, aiming to enhance the reproducibility of radiomics and the accuracy of IVD segmentation predictions.

B. RELATED WORKS

This section reviews two approaches for medical image harmonization, based on [3], [4], [8], and [18]: the conventional statistical approach and the DL-based approach. Specifically, the DL-based approach focuses on harmonization for main tasks such as classification and segmentation [3].

1) CONVENTIONAL APPROACH

Over the past few decades, conventional statistical approaches have been widely applied in medical imaging. An archetypal method is Nyúl intensity normalization [19], a histogram-matching technique where the intensities of the target image are linearly mapped to histogram landmarks (e.g., intensity percentiles) derived from standard (reference) images. The Z-score method is a statistical measure that first subtracts the mean intensity of an image from each pixel value and then divides it by its corresponding standard deviation [20]. WhiteStripe is a method that normalizes pixel intensity based on the intensity values of normal-appearing white matter (NAWM) in the brain [5]. Weisenfeld and Warfield [21] used Kullback-Leibler divergence (KLD) to minimize the disparity between the adjusted (standardized or normalized) and target images for brain segmentation. Additionally, Jäger et al. [22] proposed a method using a joint probability density function for mapping image intensities.

2) DL-BASED APPROACH

Because of the limitations of conventional methods, as mentioned in the research background section, DL-based harmonization methods have been extensively explored for medical images. For instance, Selim et al. [23] introduced a framework for standardizing CT images using a generative adversarial network (GAN), named STAN-CT. They further enhanced this framework with an alternative training strategy and ensemble approach in their GANai model, testing the RF stability [24]. Xu et al. [25] proposed a GAN-based style transfer network. The harmonized images produced by their network were utilized for data augmentation to improve U-Net segmentation performance. However, the networks proposed in these studies were designed for intra-device datasets (using the same scanner with different acquisition parameters) and not for multi-vendor CT scans. CT scans, as mentioned previously, are relatively less affected by vendor-induced variability compared to MR images. Accordingly, MRI harmonization for clinical purposes has been recently introduced for various ROIs. For example, DeSilvio et al. [26] utilized GAN to normalize prostate MRI data for cancer detection. Guan et al. [27] proposed an

attention-guided deep domain adaptation (AD2A) framework for harmonizing multi-site structural MRI to identify brain disorders. Furthermore, some studies have adopted a cycle-consistent adversarial network (CycleGAN) approach. Gao et al. [28] implemented CycleGAN with a many-to-one weak-paired strategy to harmonize multi-center MRI with unpaired datasets, evaluating their framework by differentiating high-grade glioma (HGG) from lower-grade glioma (LGG). Modanwal et al. [4] also utilized CycleGAN for intensity normalization in breast MRI, investigating segmentation performance with varying fields of view (FOV) in the PatchGAN discriminator.

Recently, Šušteršič et al. [29] experimented with a classification network using segmented results from U-Net for IVD MRI in axial and sagittal plane views. They used the same dataset as ours, examining the segmentation performance with this dataset. Consequently, our proposed method was compared with the U-Net used in [29].

C. CONTRIBUTIONS

The primary aim of this work is to achieve accurate segmentation of IVD by harmonizing multi-vendor MRI. DL is a promising approach for image harmonization and segmentation. However, DL-based automated segmentation typically requires MR images that are invariant across different vendors, along with their corresponding ground truths (GTs) – data that have been manually labeled by annotators. This manual annotation process is time-consuming and labor-intensive. Moreover, for segmented results from various new datasets, DL-based segmentation networks must be trained separately for each dataset, as depicted in Fig. 1. This approach necessitates the use of manually segmented GTs from these new datasets for training purposes. To address this challenge, we propose a DL framework that harmonizes multi-vendor MRI for robust IVD segmentation. This framework allows for the omission of additional segmentation training and manual annotation when introducing new datasets, as illustrated in Fig. 1.

The proposed DL framework is stated as follows:

1) The framework integrates a CycleGAN [30]-based network with a U-net [31]-based segmentation network to harmonize various MRIs, thereby facilitating robust IVD segmentation. This framework incorporates two key strategies: the RF loss function and the dice loss function. While radiomics is typically employed in medical imaging systems to aid clinical decisions, in this study, it serves as a learning strategy within DL approaches. Additionally, a radiomics-based metric evaluates the harmonization performance of our framework.

2) There are two reasons for adopting a CycleGAN-based model. For a general reason, DL models for image harmonization show improved performance with scan-rescan datasets, which are images scanned under different protocols of the same patient at different times. Our datasets were unpaired because of the requirement for diverse publicly

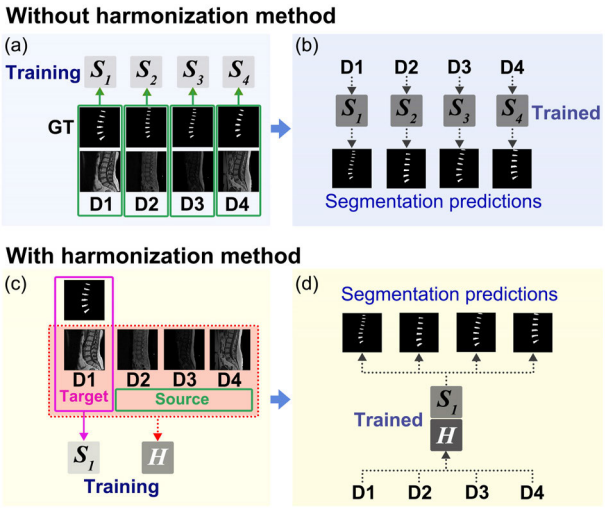

FIGURE 1. Process to acquire the segmented predictions with or without harmonization across multiple datasets (D1, D2, D3, and D4) acquired from different vendors under multi-scan protocols. The left part of the illustration, (a) and (c), presents the training process for different datasets (D1, D2, D3, and D4) without harmonization (a) and with harmonization strategy (c). The right part, (b) and (d), shows the process to obtain the segmentation predictions. Without a harmonization network, each segmentation network (S_1, S_2, S_3, and S_4) should be trained on each dataset and its GTs separately. On the other hand, the harmonization method enables the omission of additional segmentation training and manual annotation for new datasets.

available datasets. Obtaining these paired images would be challenging in clinical settings. Hence, a CycleGAN-based model, known for its efficacy with unpaired datasets, was chosen, drawing inspiration from [32], where CycleGAN was used in a super-resolution network with unpaired low- and high-resolution images. For a specific reason in our study, the cyclic nature of this model enhances segmentation performance. The segmentation network supports the harmonization process of the CycleGAN-based network. Conversely, the CycleGAN-based method is crucial for extracting features pertinent to IVD segmentation and applying these features to source images using the dice loss functions, as illustrated in Fig. 2 and Fig. 4. If the harmonization network is not cyclic, the cycle-consistency and harmonization dice loss functions cannot be utilized. Their effectiveness for robust segmentation has been demonstrated.

3) The segmentation network boosts the CycleGAN-based model to enhance features crucial for IVD segmentation. As a result, the proposed model achieves accurate segmentation predictions, outperforming state-of-the-art techniques like nnU-Net [33], even with new datasets that include disease information.

II. METHODS AND MATERIALS

A. DATASET

We collected publicly available lumbar spine MRI datasets to acquire multi-vendor data from 544 patients [34], [35] and 218 patients with a history of low back pain from four different hospitals [36]. Sudirman et al. [34] collected MRI scans with complete requirements from a collaborating

TABLE 1. Data information.

	Manufacturer	SBS (mm)	ST (mm)	PS (mm)	TR (ms)	TE (ms)	MFS (T)	ETL	FA (deg)	Num of slices
Target	Siemens[1]	4.8	4	0.729-0.875	620-680	9.2	1.5	3	150	7686
Source	Philips[1]	4.4-5.0	4	0.469-0.781	384.1-713.2	8.0-10.5	1.5	4-6	90	2941
	Philips[2]	4.0-4.4	4	0.313-0.742	408.0-1122.0	8.0-20.0	1.5	3-6	90	1997
	Philips[3]	4.4	4	0.313-0.474	400.0-562.5	10.0	3	5-8	90	604
	Philips[4]	4.4	4	0.344-0.625	426.5-615.7	10.0-15.0	3	5-7	90	462
	Philips[5]	4.4	4	0.329-0.392	522.2-610.7	10.0	3	6-7	90	135
	Philips[6]				Missing					778
	Philips[7]				Missing					120
	Siemens[1]	4.8	4	0.438-0.906	300-799	9.2	1.5	3	150	311
	Siemens[2]	4.4	4	0.664-0.804	500-632	7.9-11.0	1.5	6	120	1733
	Siemens[3]				Missing					101
	Public data [36][39]				Missing					2195

SBS: Spacing between Slices / ST: Slice Thickness / PS: Pixel Spacing / TR: Repetition Time / TE: Echo Time / MFS: Magnetic Field Strength / ETL: Echo Train Length / FA: Flip Angle
[Model] Philips[1]: Intera / Philips[2]: Gyroscan Intera / Philips[3]: Ingenia CX / Philips[4]: Ingenia / Philips[5]: Ingenia Elition X / Philips[6]: Achieva / Philips[7]: Elition / Siemens[1]: MAGNETOM Essenza / Siemens[2]: MAGNETOM Amira / Siemens[3]: Avanto

hospital and experts, and the accuracy and consistency of these scans were demonstrated using the metrics developed by [37]. Al-Kafri et al. [37] further investigated a segmentation network for disease classification, demonstrating the completeness of their dataset. Khalil et al. [38] organized a multi-vendor/modal lumbar spine MRI database with accurate ground truth data, aiming for robust automated segmentation, as detailed in [35]. The dataset for the SPIDER challenge, collected by Van der Graaf et al. [36], originated from four different centers to enhance the diagnostic studies of lumbar spine MRI. In their study, Van der Graaf et al. [39] provided GTs for vertebrae, intervertebral discs (IVDs), and spinal canals. They also compared the segmentation performance between the AI algorithm used in their work and the nnU-Net, using their dataset [36].

Additionally, MRI slices from 604 patients were acquired at Seoul National University Bundang Hospital (SNUBH; Seongnam-si, Republic of Korea). The use of this data was approved by the Institutional Review Board of SNUBH (B-2304-822-101). Among the collected datasets, T1-weighted (T1-w) sagittal slices were used in this study. Detailed information about the T1-w MRI dataset is presented in Table 1. The scan parameters mentioned in Table 1 are known to affect the signal-to-noise ratio (SNR), contrast, and image resolution [40], [41], [42], even when using the same device. Consequently, T1-w MRI obtained with the Siemens MAGNETOM Essenza scanner, under almost identical scan protocols, was chosen as the target image (notably, these images were without lumbar spine disease). On the other hand, the source images were acquired from various vendors and under different acquisition parameters, with some including a history of spinal disease.

B. EXPLANATION OF RADIOMIC FEATURES

Radiomics typically employs six classes of features [12], [43]: (1) First-order statistics, which are derived from the image histogram representing the distribution of pixel intensities; (2) Shape-based features that characterize the geometric properties of the ROI; (3) Gray Level Co-occurrence Matrix (GLCM), indicating the distribution of co-occurring pixel pairs at a preset alignment; (4) Gray Level Run Length Matrix (GLRLM), providing spatial information about lengths of consecutive pixels sharing the same intensity; (5) Gray Level Size Zone Matrix (GLSZM), detailing the size of zones where pixels have identical gray levels; and (6) Gray Level Dependence Matrix (GLDM), quantifying the dependency of gray levels (the frequency of neighboring pixels having the same value as the central pixel). Each class encompasses a variety of features. For instance, using the PyRadiomics open-source Python package (version 3.0.1) [44], 24 different features can be extracted from the GLCM, including correlation, difference entropy, joint energy, contrast, and homogeneity.

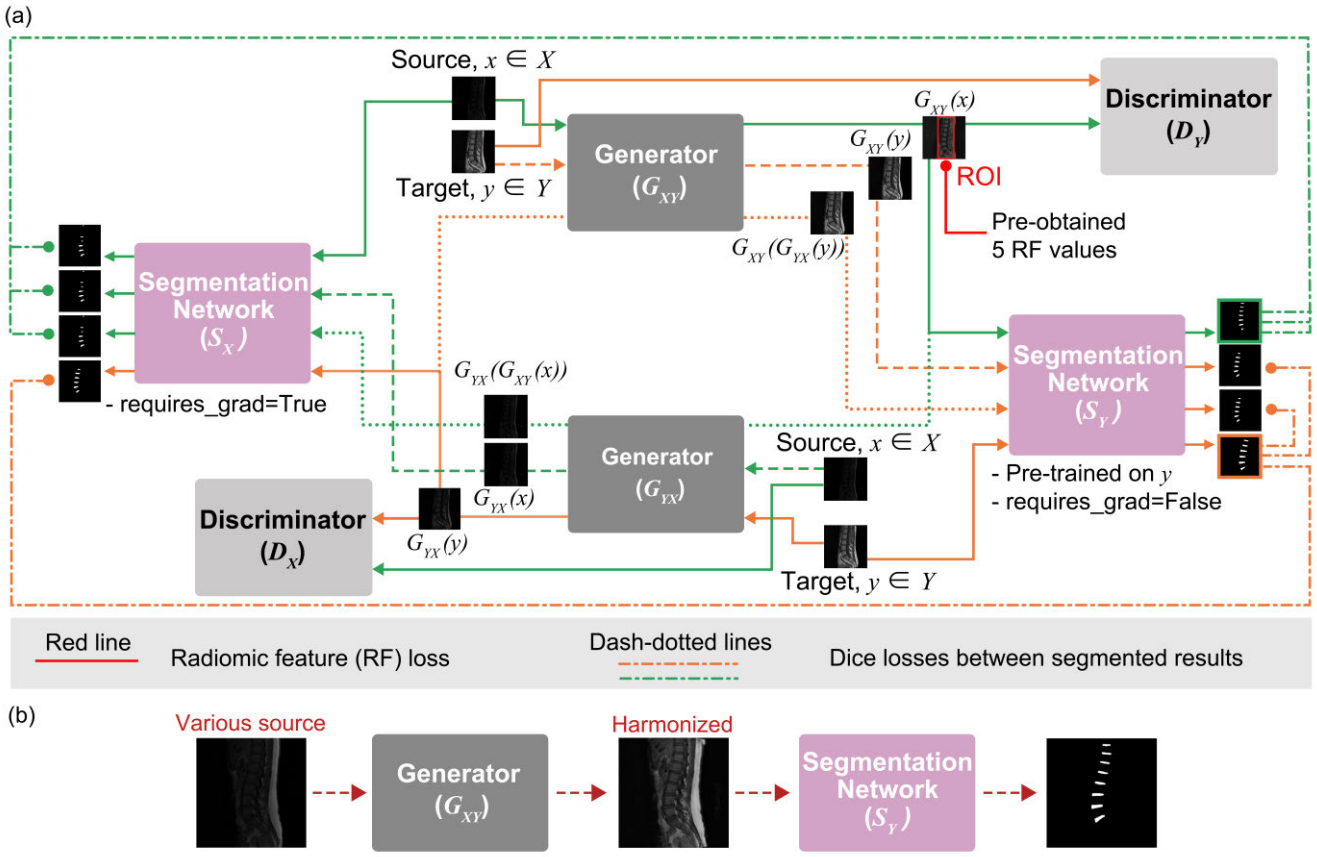

FIGURE 2. Overall architecture of the proposed model. (a) The proposed model combines a CycleGAN-based model and two segmentation networks (S_Y, S_X). The G_{XY} maps the source image x to target image y for our goal. The segmentation networks enhance the features for disc segmentation while training the CycleGAN. The S_Y was pre-trained on the target images in advance, and the S_X is not pre-trained but incorporated for cycle balance. The pre-trained S_Y provides the GTs (green and orange boxes) to calculate the dice loss functions (dash-dotted lines). The RF loss is calculated from the ROI of the harmonized image by comparing the RF values of ROI and the pre-obtained RF values from the target images (red line). (b) The G_{XY} trained in (a) can harmonize diverse MR images to predict the segmented results via the S_Y trained on the target images.

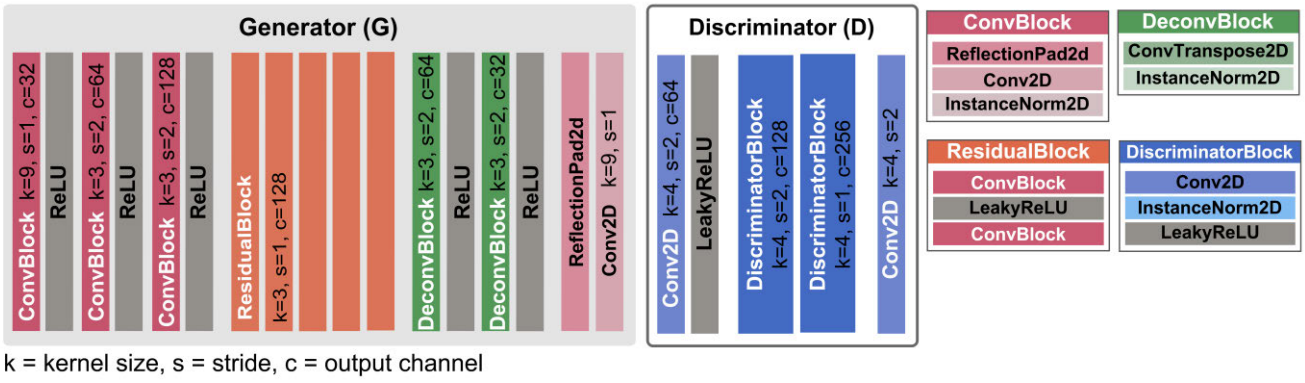

FIGURE 3. Detailed architectures of the generator and discriminator in the proposed model.

C. NETWORK ARCHITECTURE

A CycleGAN-based network is employed in this study, where G_{XY} learns a mapping from the source domain X to the target domain Y for optimal IVD segmentation, as illustrated in Fig. 2. x ($\in X$) represents the source images (i.e., multi-vendor images). By contrast, y ($\in Y$) indicates the target images (i.e., those acquired by an identical vendor and acquisition parameter). The architecture of the generators, G_{XY} and its inverse mapping G_{YX}, are adapted from the style transfer (ST) network proposed by [45], and two PatchGAN discriminators [46], D_Y and D_X, are utilized for adversarial learning. The segmentation networks (S_Y, S_X) use the architecture provided by the Segmentation Models Pytorch (SMP) library (version 0.3.3) [47].

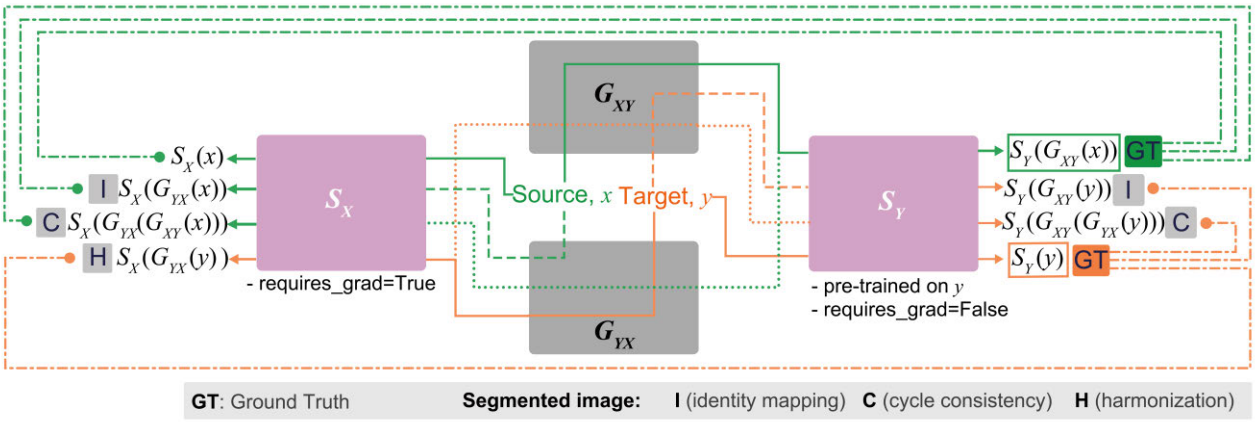

FIGURE 4. Schematic diagram of calculating dice loss functions (dash-dotted lines) during harmonization. The propagations of the source and target image are represented by green and orange colored lines, respectively. The green and orange boxes (GT) act as ground truths. Each GT is compared with each segmented prediction (gray box) whose anatomic structure is the same as each GT's. 'I' is a segmented image by identity mapping, and 'C' and 'H' denote the segmented results from the image reconstructed by cycle-consistent mapping and the harmonized image, respectively.

1) GENERATOR

The two generators have the same architecture as that of the ST network. This network primarily consists of three blocks, as shown in Fig. 3: ConvBlock, ResidualBlock, and DeconvBlock. The ConvBlock includes a ReflectionPad, a convolutional layer (either 3×3 or 9×9), and instance normalization, followed by a ReLU activation function. The DeconvBlock, similarly, consists of a 3×3 convolutional layer with a stride of 1/2, followed by instance normalization and a ReLU activation function. The ResidualBlock is composed sequentially of a ConvBlock, a LeakyReLU activation function, and another ConvBlock.

2) DISCRIMINATOR

The two discriminators have the same architecture, which is primarily composed of two DiscriminatorBlocks. Each of these blocks includes 4×4 convolutional layers, instance normalization, and LeakyReLU layers. At the beginning of the discriminator architecture, there is an initial layer that includes a 4×4 convolutional layer followed by a LeakyReLU layer. The architecture concludes with a final layer that consists solely of a 4×4 convolutional layer. Each discriminator competes with its generator network. In our model, the receptive field size is set to 34×34.

3) SEGMENTATION NETWORK

Segmentation networks employ architectures from SMP, which allow for various combinations of encoders with pre-trained weights and decoders. We conducted experiments with different combinations to identify the most effective model. These combinations included ResNet34, ResNet50, and ResNet101 [48] as encoders with pre-trained weights from 'ImageNet', and U-net, U-net++ [49], and DeepLabV3 [50] as decoders, each with a sigmoid activation function. Note that the S_Y is pre-trained on the target images and then frozen upon integration into our proposed model. This enables the S_Y provide the GTs (illustrated as green and orange boxes in Fig. 2 and Fig. 4) for calculating the dice losses (represented by dash-dotted lines in Fig. 2 and Fig. 4). The S_Y also facilitates the evaluation of the segmentation performance of our method during the testing phase, as shown in Fig. 6. The S_X shares the same architecture as the S_Y. However, unlike the S_Y, it is not pre-trained and instead learns to improve segmentation performance on the source images during the training of our proposed model. To maintain cycle balance, both networks are incorporated into our model.

D. LOSS FUNCTIONS

We employed the original CycleGAN [30] loss functions, supplemented with dice and RF loss functions to harmonize various MRI and enhance the RFs associated with IVD segmentation.

1) ADVERSERIAL LOSS

Both generators, G_{XY} and G_{YX}, learn using an adversarial strategy. For example, the adversarial loss function for G_{XY} and its discriminator D_Y can be formulated as follows:

$$\mathcal{L}_{GAN}(G_{XY}, D_Y) = \mathbb{E}_{y \sim P_Y}[log\, D_Y(y)] + \mathbb{E}_{x \sim P_X} \left[\log(1 - D_Y(G_{XY}(x))\right] \quad (1)$$

where P_X and P_Y denote the data distribution of domains X and Y, respectively, and G_{XY} and D_Y are engaged in a min-max game. Specifically, G_{XY} attempts to synthesize the source image x as indistinguishable from the target image y and then fools D_Y to classify the harmonized image $G_{XY}(x)$ as real. On the other hand, D_Y attempts to correctly distinguish between the real image y and the artificial image $G_{XY}(x)$. Similarly, the generator G_{YX} and its discriminator D_X undergo a parallel optimization process.

2) CYCLE CONSISTENCY LOSS

For cycle consistency between the original images (e.g., x) and the reconstructed images (e.g., $G_{YX}(G_{XY}(x))$, we implement both forward (i.e., $x \rightarrow G_{XY}(x) \rightarrow G_{YX}(G_{XY}(x)) \approx x$)

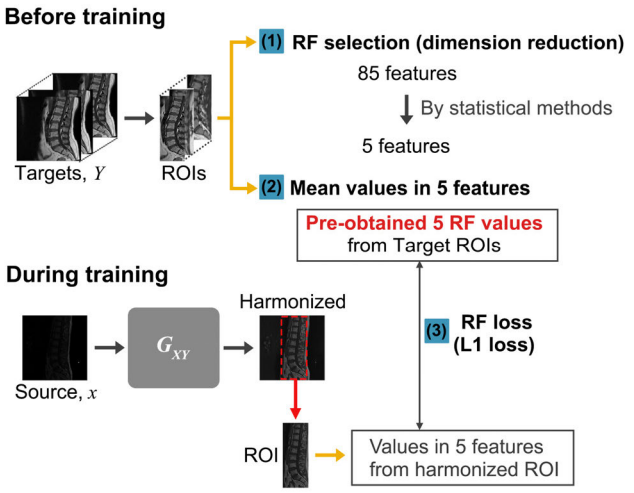

FIGURE 5. Procedure for RF loss function. Before training, (1) the features representing the target images are selected among all features using statistical methods. (2) The mean values of the selected RFs are pre-obtained from the ROIs in the target (pre-obtained 5 RF values). During training, (3) the RF values of the ROI in the harmonized image are compared with the pre-obtained RF values from (2).

and backward cycle consistency (i.e., $y \rightarrow G_{YX}(y) \rightarrow G_{XY}(G_{YX}(y)) \approx y$) losses:

$$\mathcal{L}_{cyc}(G_{XY}, G_{YX}) = \mathbb{E}_{x\sim P_X}\left[\|G_{YX}(G_{XY}(x)) - x\|_1\right] + \mathbb{E}_{y\sim P_Y}\left[\|G_{XY}(G_{YX}(y)) - y\|_1\right] \quad (2)$$

where the loss function is the L1 norm introduced in the original CycleGAN [30].

3) IDENTITY MAPPING LOSS

The identity mapping loss function is employed to support harmonized images to retain their original contents (anatomic structures) and avoid color variations:

$$\mathcal{L}_{id}(G_{XY}, G_{YX}) = \mathbb{E}_{x\sim P_X}\left[\|G_{YX}(x) - x\|_1\right] + \mathbb{E}_{y\sim P_Y}\left[\|G_{XY}(y) - y\|_1\right] \quad (3)$$

4) DICE LOSS IN S_Y, S_X

In our proposed model, we incorporate dice loss functions, including identity mapping, cycle-consistency, and harmonization dice loss functions. These functions are instrumental in improving the image features that are crucial for accurate IVD segmentation. As depicted in Fig. 4, the predicted results $S_Y(G_{XY}(x))$ and $S_Y(y)$ serve as the GT (referred to as 'GT' and represented by green and orange boxes). These GTs are then compared with the segmented images (I, C, and H, denoted by the gray box in Fig. 4) that correspond to the same anatomic structures as the GTs to compute the dice losses (dash-dotted lines).

5) RADIOMIC FEATURE (RF) LOSS

The RF loss function is utilized in the G_{XY} network to harmonize the input images in terms of the RFs. The overall procedure for the RF loss function is depicted in Fig. 5. Before the training process, we select informative RFs among all the available features using the statistical methods described in Section III-B, where five RFs are selected. Subsequently, we calculate the values of these selected RFs from 1000 slices within the ROI in the target images. These calculated RF values are then averaged over the 1000 target images, yielding pre-obtained values for the five RFs (named pre-obtained 5 RF values). During the training phase, we calculate the values of the selected RFs from the images harmonized by our model and compare them with the pre-obtained RF values. In essence, the L1 loss function is employed to measure the difference in the selected RFs between the target and harmonized images at each iteration and minimize this difference during the training process.

III. EXPERIMENT SET-UP

The overall workflow is visualized in Fig. 6. In the pre-set phase, we begin by selecting the optimal segmentation network for the target images, as detailed in Section III-C-I. Subsequently, the chosen segmentation network (S_Y) undergoes pre-training on the target images, and this pre-trained network is integrated into our proposed harmonization model. The RF selection procedure, as elaborated in Section III-B, is employed to configure the RF loss function. During the training phase, our proposed model learns to harmonize a diverse range of source images to map them to the target images. In the testing phase, the resultant harmonized images are fed into the pre-trained S_Y to assess the segmentation performance. All architectures were implemented in PyTorch 1.9.0, and the training process was performed with a single NVIDIA GeForce RTX 3090 Ti, 64 GB of RAM, and a 13th Gen Intel(R) Core (TM) i9-13900KF.

A. DATASETS FOR EXPERIMENTS

1) DATASET FOR SEGMENTATION NETWORK (S_Y)

We used 1230 sagittal slices of target images from 169 healthy patients to identify the optimal segmentation network among the SMP models and to pre-train the selected network for integration into the harmonization network. The number of slices (patients) divided into training, validation, and test subsets was 1028 (140), 100 (14), and 102 (15), respectively. The input images were resized to a spatial size of 320 × 320, matching the original image resolution of the target MRI. Min-max normalization (0-255) was applied to the images, and they were converted into 3-channel images to accommodate the use of pre-trained weights from ImageNet in the encoder (ResNet). Regarding GTs, all datasets for this segmentation network underwent manual annotation by two clinical experts at SNUBH, employing the 3D slicer software [51].

2) DATASET FOR HARMONIZATION NETWORK

Detailed information regarding the dataset, which includes target and source images, is presented in Table 1. The division of MRI slices for the training and validation subsets followed an 8:2 ratio. The same image preprocessing procedure as

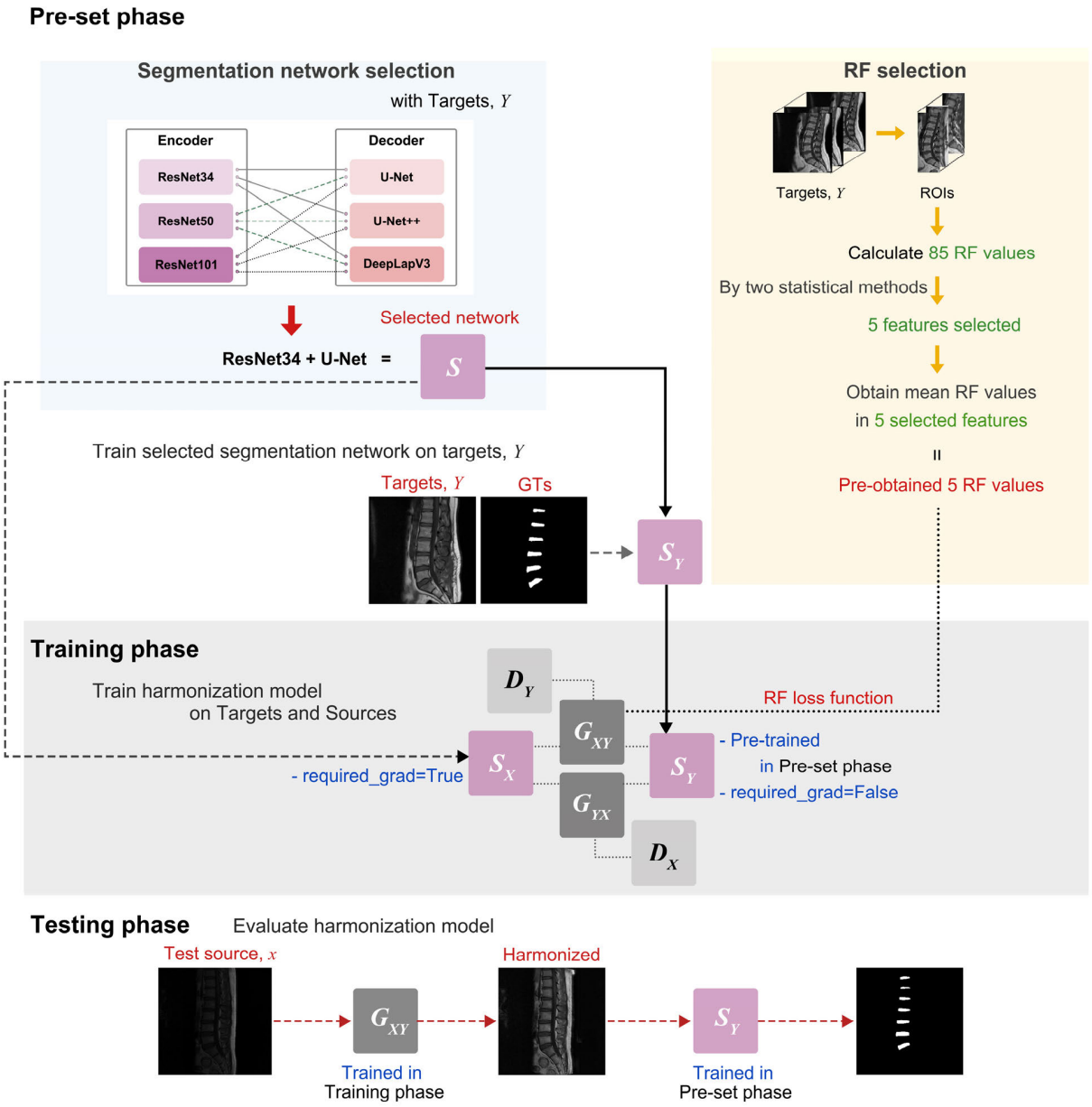

FIGURE 6. Flowchart of the work process. For the pre-set, the optimal segmentation network S on the target images is first selected among various models (encoder + decoder). ResNet34+U-Net achieved optimal performance in this study. The selected S is trained on the target images and incorporated into the harmonization network for training. In addition, features relevant to our study are selected by statistical methods. Five features were selected from the target ROIs in this study. Five RF values are measured (Pre-obtained 5 RF values) for the RF loss function. In the training phase, the proposed harmonization model is trained on the target and source images with two strategies (segmentation networks (dice loss functions) and RF loss function). In the testing phase, the segmentation performance on the harmonized images is evaluated via the pre-trained S_Y (red dotted line).

described for the S_Y dataset was applied to this dataset but the original gray-scale images (one channel) were used as input. When inputting into the segmentation network while training the proposed model, the images were converted into three channels. The test subset (100 weakly paired sets of targets and sources) for harmonization evaluation was organized to reduce the structural disparity between each other, that is, the target and source images for the test were paired to have similar anatomic structures. The ROI was cropped to a pixel size of 120×320 and rescaled using min-max normalization (0-255). In contrast, the test subset (source) for segmentation assessment consisted of 100 slices from 70 patients, with half of these containing disc disease information, to test the robustness to variability.

B. FEATURE SELECTION

This procedure aims to select RFs that contain the representative properties of the target images. Prior to feature selection, the ROI (lumbar spine) was cropped with a pixel size of 120×320 and rescaled using min-max normalization (0-255). Five features were selected from a total of 85 features (using

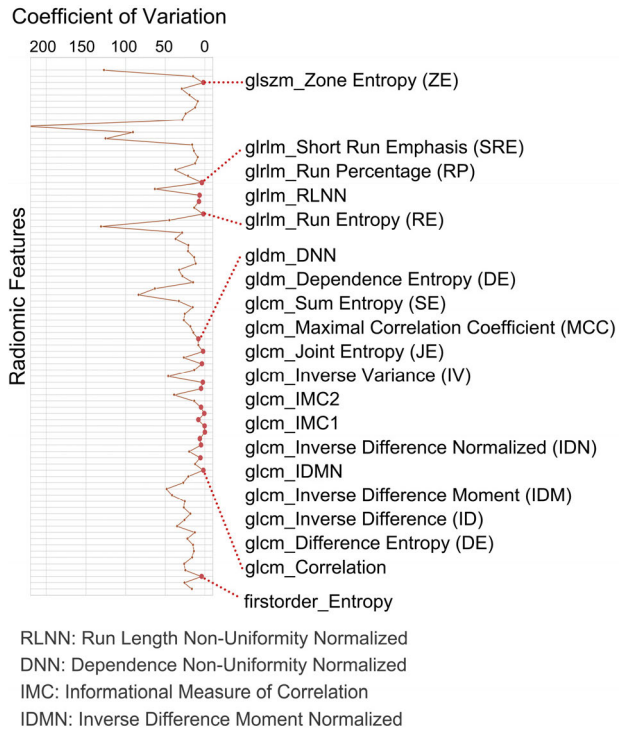

FIGURE 7. Feature Selection Step 1. The top 20 RFs with minor variances (red dots) are selected by the CV.

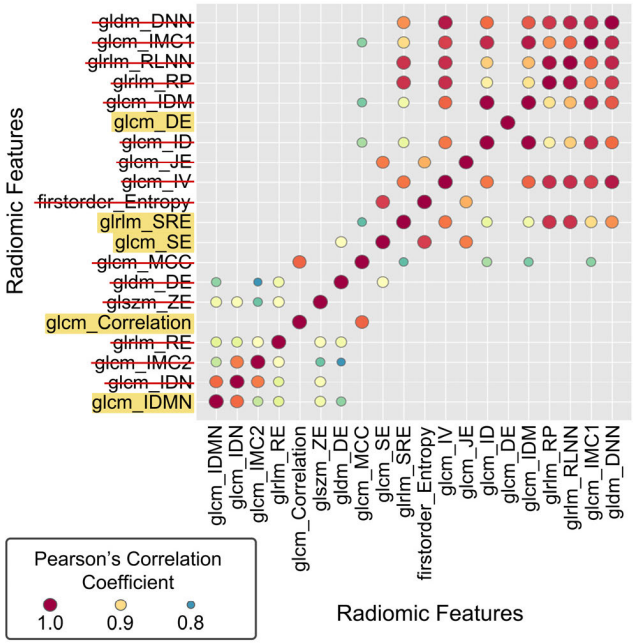

FIGURE 8. Feature Selection Step 2. The PCC between the RFs selected in Step 1 is visualized with colored dots. From the glcm_IDMN with the smallest CV, the features with high PCC (>0.8) are sequentially excluded. (From bottom to top in the y-axis, the CV becomes higher) (red-line: excluded features; highlighted in yellow: selected features).

the default settings in PyRadiomics) for the RF loss function and RF evaluation, following a two-step process:

Step 1: Coefficient of Variance (CV) Calculation [52].

The CV was calculated to identify RFs with small variances using 1000 target images. The results of this step are illustrated in Fig. 7, where the red dots represent the top 20 RFs with the lowest variances.

Step 2: Pearson's Correlation Coefficient (PCC) Calculation [53].

Subsequently, the PCC was calculated between the 20 top RFs to identify and exclude redundant features. Some RFs contain similar texture information, and it is important to eliminate those highly correlated with other features. RFs with a correlation coefficient above 0.8 were excluded from the initial 20 top-ranked RFs, as depicted in Fig. 8.

Ultimately, the RF loss function and evaluation in this study were based on five selected features: glcm_IDMN (Inverse Difference Moment Normalized), glcm_Correlation, glcm_SE (Sum Entropy), glrlm_SRE (Short Run Emphasis), and glcm_DE (Difference Entropy). Explanations for the measurement of each feature are presented in Table 6. The open-source package used for this analysis was PyRadiomics, version 3.0.1, with the default settings.

C. TRAINING AND EXPERIMENT PROCESS

1) PRE-TRAINING SEGMENTATION NETWORK (S_Y)

We investigated various encoder-decoder combinations to determine the best segmentation network for the target images. Prior to its integration into the proposed model, the selected segmentation network, S_Y underwent pre-training on the target images. Hyperparameter optimization was carried out with a range of parameters, including optimization algorithms such as ADAM [54] and Stochastic Gradient Descent (SGD) with momentum parameters set to $\beta_1 = 0.5$, $\beta_2 = 0.999$, and learning rates of 0.001, 0.005, and 0.01. The training process was conducted over a varying number of epochs, spanning up to 400, while employing mini-batch sizes of 8 and 16. Furthermore, the weight ranges of the dice loss and intersection-over-union (IoU) loss functions were explored, with values ranging from 0 to 1 or 10.

2) TRAINING HARMONIZATION MODEL

The input images $x(y)$ first propagated through the G_{XY} (G_{YX}) network with a batch size of four. This network mapped the domain X (Y) to the domain Y (X), as shown in Fig. 2. Then, this harmonization network generated the images G_{XY} (x) (G_{YX} (y)) by the following components: discriminator network (D_Y (D_X)) penalty (adversarial constraint), cycle consistency, identity mapping, three dice loss functions, and an RF loss function. During the discriminator training, the input image $y(x)$ and harmonized images G_{XY} (x) (G_{YX} (y)) were classified as real and fake, respectively.

3) PROPOSED MODEL OPTIMIZATION

The optimal hyperparameters of the proposed model were explored. The ADAM with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ was used, and a learning rate ranging from 1×10^6 to 1×10^4 was examined throughout up to 300 epochs. Concerning the weights assigned to the loss

TABLE 2. Segmentation performance of diverse combinations of encoders and decoders [mean, 95% confidence intervals].

Model	Encoder	Decoder	Trainable Param.	Dice	IoU	F1 score	precision	recall
ResNet34 + U-Net	ResNet34	U-Net	24.4M	**0.927, 0.005**	**0.864, 0.008**	**0.926, 0.005**	0.929, 0.007	**0.926, 0.009**
ResNet34 + U-Net++	ResNet34	U-Net++	26.1M	0.921, 0.005	0.854, 0.008	0.921, 0.005	0.938, 0.007	0.907, 0.010
ResNet34 + DeepLabV3	ResNet34	DeepLabV3	26.0M	0.921, 0.005	0.854, 0.008	0.920, 0.005	0.937, 0.006	0.906, 0.009
ResNet50 + U-Net	ResNet50	U-Net	32.5M	0.922, 0.005	0.855, 0.008	0.922, 0.005	0.936, 0.007	0.910, 0.010
ResNet50 + U-Net++	ResNet50	U-Net++	49.0M	0.925, 0.005	0.861, 0.008	0.925, 0.005	**0.941, 0.007**	0.911, 0.010
ResNet50 + DeepLabV3	ResNet50	DeepLabV3	39.6M	0.911, 0.005	0.838, 0.008	0.910, 0.005	0.920, 0.007	0.905, 0.008
ResNet101 + U-Net	ResNet101	U-Net	51.5M	0.923, 0.005	0.858, 0.008	0.923, 0.005	0.939, 0.007	0.910, 0.009
ResNet101 + U-Net++	ResNet101	U-Net++	68.0M	0.923, 0.005	0.857, 0.009	0.922, 0.006	0.937, 0.008	0.911, 0.010
ResNet101 + DeepLabV3	ResNet101	DeepLabV3	58.6M	0.913, 0.005	0.842, 0.008	0.922, 0.006	0.922,0.007	0.908, 0.008

functions, the optimized hyperparameters for the adversarial, cycle consistency, and identity mapping losses were all set to 10. Additionally, the dice loss functions in the S_Y and S_X had hyperparameters of 20 and 10, respectively, while the weight of the RF loss function was set to 10.

4) EFFECTIVENESS OF SEGMENTATION MODELS

To ensure cycle balance, two segmentation networks were utilized, and all possible dice loss functions were employed, as depicted in Fig. 4. We conducted an ablation study to evaluate the effectiveness of the segmentation network (only S_X) and all dice loss functions.

D. EVALUATION METHODS

The proposed model was evaluated in terms of both segmentation and harmonization performance. Widely used metrics for semantic segmentation (Dice coefficient, IoU, F1 score, precision, and recall) were employed for segmentation evaluation. The test dataset for segmentation prediction consisted of diverse source images, maintaining a 5:5 ratio between slices from healthy patients and those with disc disease to examine the robustness to variability. The performance of the proposed model was compared with several other methods. These included the modified CycleGAN harmonization model from [4], the U-Net model from [29], models listed in Table 2, and the nnU-Net from [33].

Nyúl intensity normalization [19] and the modified CycleGAN harmonization model investigated in [4] were compared with our model for the harmonization test. Nyúl normalized images were acquired using Nyúl intensity normalization on the source images. This normalization began by estimating the averages of the landmarks (percentiles) on the histograms of 3000 target images. Subsequently, the source images were mapped to these landmarks to undergo normalization or harmonization. The modified CycleGAN model [4] was trained using our training dataset. The proposed model and these other methods were assessed by measuring the style feature disparity from the target images. Thus, a prerequisite for the evaluation is to set the test dataset (target and source images) to be paired to share similar anatomical structures. Therefore, 100 slices of the reference images for evaluation were manually selected from the target images, which retained similar anatomical structures to the test source images. This process mitigated structural differences between the datasets, as the datasets used for evaluation were not originally paired. We quantified the relative error of the selected RF values between the reference (target) and the images harmonized by the various harmonization methods.

IV. RESULTS

A. OPTIMAL SEGMENTATION NETWORK

Diverse combinations of encoders and decoders from SMP were compared. The objective was to identify the network that exhibited the most accurate segmentation results for the target images to integrate into the proposed network. It was found that the ResNet34 (encoder) + U-Net (decoder) network outperformed all other combinations across all metrics (Dice = 0.927, IoU = 0.864, F1 score = 0.926, and recall = 0.926), with the exception of precision, as presented in Table 2. Therefore, ResNet34+U-Net was incorporated into the harmonization network.

B. ABLATION STUDY ON SEGMENTATION NETWORKS

The effectiveness of each segmentation network in the CycleGAN-based model (named 'Ours') was investigated: (1) Ours without both S_X and S_Y ($-S_X - S_Y$) and (2) Ours without S_X ($-S_X + S_Y$). The test dataset comprised various source images, including slices with lumbar spine disease, to demonstrate the model robustness to variability. These test source images were harmonized through the harmonization network and subsequently evaluated using the pre-trained S_Y network. The proposed model without S_Y was not examined because the S_Y network provides GTs during the training of the harmonization model, making it necessary for the S_X to

TABLE 3. Segmentation performance of proposed model with or without segmentation networks (S_X and S_Y) [mean, 95% confidence intervals].

Network	S_X	S_Y	Dice	IoU	F1 score	precision	recall
- S_X - S_Y			0.897, 0.009	0.815, 0.014	0.897, 0.009	0.894, 0.041	0.852, 0.040
- S_X + S_Y		✓	0.898, 0.008	0.818, 0.013	0.897, 0.009	0.893, 0.040	0.855, 0.040
+ S_X + S_Y	✓	✓	0.918, 0.006	0.850, 0.010	0.917, 0.006	0.940, 0.006	0.897, 0.009

TABLE 4. Segmentation performance of proposed model with or without dice loss functions [mean, 95% confidence intervals].

Dice loss functions			Dice	IoU	F1 score	precision	recall
Cycle consistency	Harmonization	Identity mapping					
✓	✓		0.906, 0.006	0.830, 0.009	0.906, 0.006	0.917, 0.006	0.875, 0.010
✓		✓	0.906, 0.006	0.829, 0.038	0.906, 0.006	0.915, 0.006	0.878, 0.009
	✓	✓	0.903, 0.006	0.824, 0.009	0.903, 0.006	0.916, 0.006	0.874, 0.010
✓	✓	✓	0.918, 0.005	0.850, 0.008	0.917, 0.006	0.940, 0.005	0.897, 0.009

work in conjunction with the S_Y. The proposed model with both segmentation networks ($+S_X + S_Y$) achieved remarkable scores for all metrics (Dice = 0.918, IoU = 0.850, F1 score = 0.917, precision = 0.940, and recall = 0.897), as shown in Table 3.

C. ABLATION STUDY ON DICE LOSS FUNCTIONS

An ablation study was conducted to assess the efficacy of three distinct dice loss functions: identity mapping, cycle consistency, and harmonization dice loss function, as depicted in Fig. 4. This experiment showed that the proposed harmonization model, equipped with all three dice losses, achieved the best results across all metrics in Table 4.

D. COMPARISON WITH OTHER METHODS ON SEGMENTATION PERFORMANCE

We compared the segmentation performance of the proposed model with other methods, including:

1. A modified CycleGAN model examined by Modanwal et al. [4] for intensity normalization in breast MRI.
2. The basic U-Net used for IVD segmentation in disc hernia diagnosis by Šušteršič et al. [29], which utilized the same public dataset as ours to train the segmentation network.
3. ResNet101+U-Net [55].
4. ResNet34+DeepLabV3.
5. The widely applied nnU-Net, a state-of-the-art technique for 2D/3D segmentation tasks [33].

In a similar manner to our study procedure, we first trained the modified CycleGAN harmonization model [4] using our dataset and then evaluated it using the S_Y. In [55], the segmentation performance of the vertebral body in MRI was investigated by comparing multiple models (e.g., U-Net, VGG+U-Net, ResNet+U-Net, and ResNet+SegNet) for disease classification. Among these, ResNet101+U-Net outperformed other models. Similarly, for our target dataset, ResNet34+U-Net emerged as the optimal model. Therefore, we tested whether ResNet+U-Net outperformed other models using multi-vendor MRI. Furthermore, all the models listed in Table 2 were compared (see Table 7), where ResNet34+DeepLabV3 showed the highest scores. ResNet34+U-Net outperformed other models (combinations of encoders and decoders in Table 2) with the target dataset (Table 2), whereas ResNet34+DeepLabV3 achieved better segmentation performance with diverse source datasets (Table BI). All segmentation models were trained on the target dataset and tested on the test source images.

The effectiveness of the RF loss function and segmentation networks in the proposed model is also examined. 'Ours (+RF-Seg)' denotes the proposed model with the RF loss function but without the segmentation networks, whereas 'Ours (-RF+Seg)' indicates the model with the segmentation networks but without the RF loss function. As shown in Table 5, the proposed model with the RF loss function and segmentation networks exhibited outstanding performance in terms of statistical metrics. Furthermore, compared to other methods, our proposed model with both strategies ('Ours (+RF+Seg)') yielded the highest scores in Dice, IoU, F1 score, and recall (Dice = 0.920, IoU = 0.853, F1 score = 0.920, and recall = 0.902). Visual results of the segmented images are shown in Fig. 9, highlighting the proposed model robustness to variability in MRI for IVD segmentation.

E. VISUAL RESULT ON ANATOMIC PRESERVATION

We ensured that the resultant images harmonized via the proposed model preserved the anatomical structures of the original (source) images after harmonization using the cutting

TABLE 5. Comparison of segmentation performance with other methods [mean, 95% confidence intervals].

Model	Dice	IoU	F1 score	precision	recall
Modanwal *et al.* [4] → S_y	0.869, 0.016	0.775, 0.020	0.869, 0.016	0.873, 0.039	0.872, 0.010
U-Net [29]	0.489, 0.075	0.408, 0.068	0.488, 0.075	0.937, 0.004	0.421, 0.072
ResNet101+U-Net [55]	0.709, 0.058	0.613, 0.056	0.709, 0.058	**0.960, 0.006**	0.639,0.060
ResNet34+DeepLabV3	0.862, 0.015	0.764, 0.020	0.862, 0.015	0.909, 0.006	0.830, 0.023
nnU-Net [33]	0.907, 0.009	0.832, 0.014	0.907, 0.009	0.945, 0.004	0.875, 0.015
Ours (+RF -Seg)	0.914, 0.006	0.842, 0.010	0.913, 0.006	0.935, 0.006	0.895, 0.010
Ours (-RF +Seg)	0.918, 0.006	0.850, 0.010	0.917, 0.006	0.940, 0.006	0.897, 0.009
Ours (+RF +Seg)	**0.920, 0.005**	**0.853, 0.008**	**0.920, 0.006**	0.940, 0.005	**0.902, 0.008**

and weaving method, as demonstrated in Fig. 10 (last column labeled 'Geometry Matching'). It was observed that the style of the target images was effectively transferred to various source images, and the content of the source images, even when scanned by different vendors, was retained after harmonization.

F. COMPARISON WITH OTHER HARMONIZATION METHODS ON RADIOMICS

The proposed model was numerically compared with other harmonization methods, including Nyúl intensity normalization [19] and the modified CycleGAN model investigated in [4]. Relative errors, in comparison to the RF values of the target images (actual values), were measured for five selected RFs (glcm_IDMN, glcm_Correlation, glcm_SE, glrlm_SRE, and glcm_DE), which are the archetypal features of the target image. In Fig 11, a darker color (representing a relative error ≈ 0) indicates that the RFs of the test images closely match the feature values of the target images. Our model exhibited minor disparities in RF values compared to other harmonization methods, particularly in glcm_IDMN, glcm_Correlation, glcm_SE, and glcm_DE, as depicted in Fig. 11. Visual results from multiple vendors are also compared in Fig. 12. The resultant images harmonized by our proposed model effectively adopted the style of the target images and demonstrated better consistent quality, even with different vendors.

V. DISCUSSION

The harmonization of medical images has been emphasized because MRI features are vendor-dependent, hindering accurate diagnosis and prognosis at clinical sites. Therefore, we propose a framework to harmonize multi-vendor MRI (source image) and map them to the target image for accurate IVD segmentation, using a novel radiomics approach. Radiomics is typically used to analyze the texture or shape of an ROI in medical imaging for diagnosis and prognosis. In this study, we utilized the radiomics approach in reverse, employing it as a loss function (RF loss function) to enhance the harmonization model in terms of radiomics. Additionally, dice loss functions were incorporated into the harmonization network to improve the features related to robust IVD segmentation.

The proposed DL model consists of two modules: a CycleGAN-based model for data harmonization and a segmentation network for enhancing predictions. In this study, we adopted the segmentation architecture provided by the SMP library, but our proposed model can be seamlessly integrated with any other segmentation network developed elsewhere. Furthermore, training a segmentation network on new datasets and annotation for GTs is optional with our proposed model, as explained in Fig. 1. Consequently, our model enables direct segmentation prediction on various MRI images after harmonization.

We conducted an ablation experiment to demonstrate the effectiveness of our proposed strategies (segmentation networks, all dice loss functions, and RF loss function) on segmentation performance. This experiment conclusively showed that all segmentation networks and loss functions contributed significantly to the proposed harmonization model.

Finally, we compared the segmentation performance of the proposed model with that of other methods, including the state-of-the-art technique, nnU-Net. Target images were acquired from identical vendors, scan parameters, and healthy patients. In contrast, the test source images contained considerable variability induced by different vendors, scan parameters, and disease information. Despite these variabilities, our model consistently outperformed other models in segmentation evaluation.

To compare harmonization performance in radiomics, we manually curated a dataset in which the content of source images closely resembled that of the target images, mitigating structural disparities. The proposed model was compared with the conventional histogram matching method, Nyúl intensity normalization [19], and CycleGAN-based harmonization model investigated in [4] by measuring relative errors compared to the target images. Our proposed model exhibited the best harmonization performance both quantitatively and qualitatively.

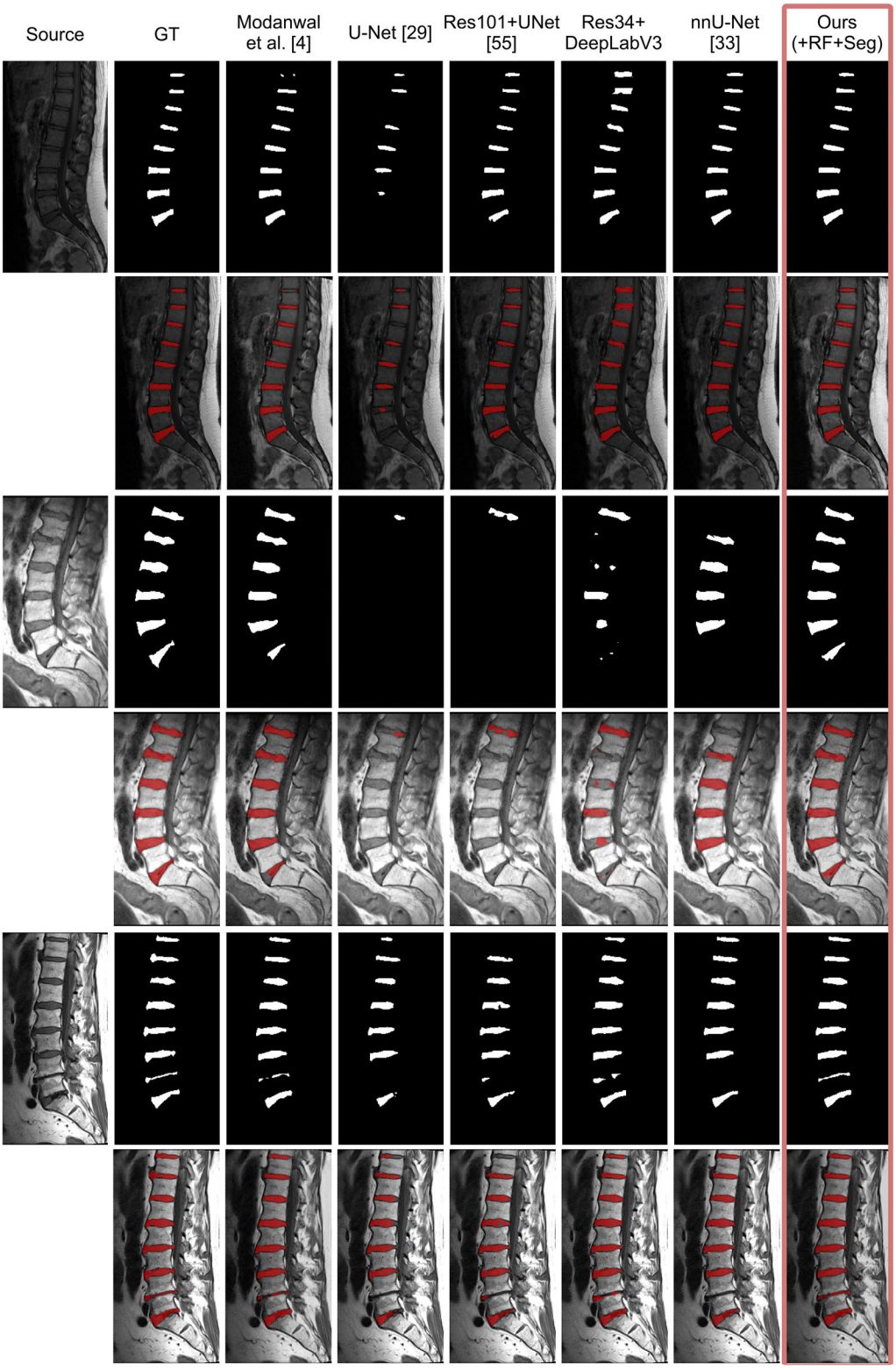

FIGURE 9. Comparison of segmented visual results of our model (+RF+Seg) and other methods on test source images acquired from different vendors. The 'GT' was manually annotated from the source image. The image below the GT is the overlaid image with the source and GT. The segmented results via each method are shown in each upper row. The overlaid images are presented in each bottom row. Note that in the overlaid images in the third and last column, the MR images are the harmonized images via the model in [4] and our proposed model. The resultant images in the last rows are the segmented predictions of the MR images with disc disease.

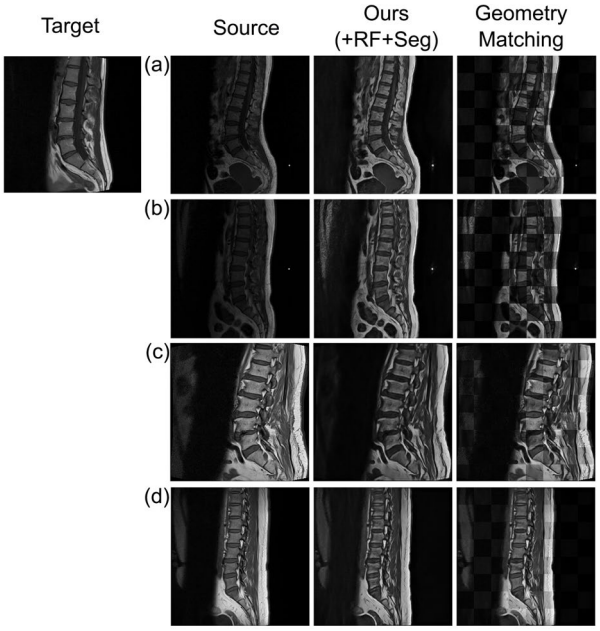

FIGURE 10. Visual results of harmonization and geometric conservation of the harmonized images via the proposed model. Each source image was scanned by different vendors: (a) Philips Gyroscan, (b) Philips Intera, (c) Siemens MAGNETOM Essenza, and (d) Siemens MAGNETOM Amira. The harmonized images are presented in the middle column ('Ours (+RF+Seg)'). The comparison of anatomic contents between the source and harmonized images is shown in the last column ('Geometry Matching').

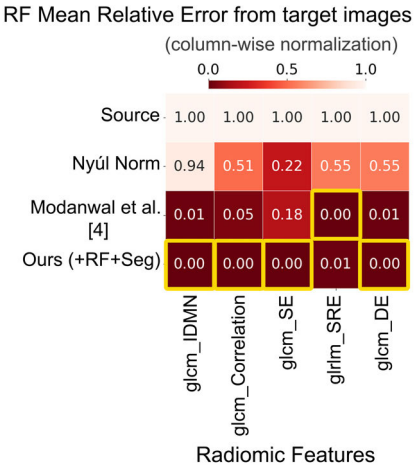

FIGURE 11. RF relative error comparison between the source images, other methods (Nyúl normalization and Modanwal et al. [4]), and our model in five selected RFs. A darker color indicates that the RF of the test images is more similar to the target images. The yellow lined boxes denote the highest similarity to the target images.

Our study has some limitations. The segmentation architectures provided by SMP were employed, and with this segmentation network, the proposed model achieved good segmentation performance. However, if we could employ a better segmentation network in the proposed harmonization model, the resultant prediction would be more robust to the MRI variability. Additionally, our source images were diverse, but collecting more datasets would improve the model reproducibility with new data. For feature selection,

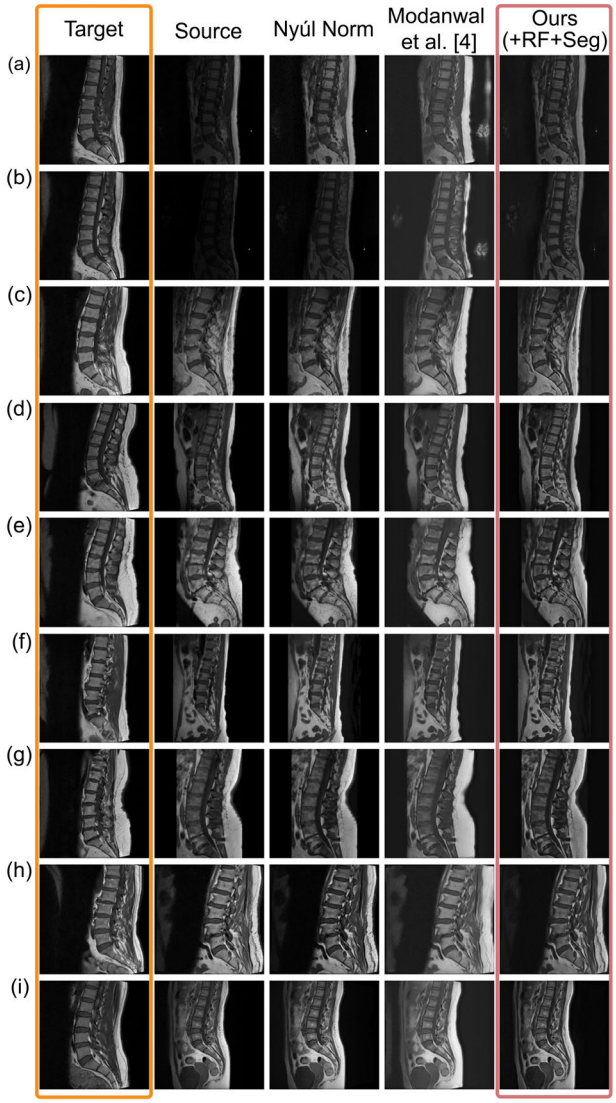

FIGURE 12. Visual result comparison of harmonization. Each image was acquired by different vendors. (a) Philips Intera, (b) Philips Gyroscan, (c) Philips Ingenia, (d) Philips Ingenis CX, (e) Philips Eliton, (f) Philips Eliton X, (g) Philips Achieva, (h) Siemens MAGNETOM Essenza, and (i) Siemens MAGNETOM Amira.

we used two methods to obtain representative features of the target images: (1) Step 1: CV and (2) Step 2: PCC. However, there are several other methods for dimension reduction, such as principal component analysis (PCA), recursive feature elimination (RFE), and least absolute shrinkage and selection operator (LASSO) [17], [53], [56]. Each of these methods has its advantages and limitations. PCC, as used in this study, is sensitive to user-defined thresholds and detects only linear dependencies, potentially affecting model performance [17], [53]. Therefore, testing our model with various feature selection methods and carefully setting thresholds could yield optimal results.

In this study, we have proposed a harmonization DL model for accurate segmentation, demonstrating superior performance compared to other methods. We anticipate that this

TABLE 6. Explanation on five selected features.

Class of feature	Type of feature	Measure	Equation	
GLCM $P(i,j)$	IDMN	Local homogeneity	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+(\frac{k^2}{N_g^2})}$ $k=\lvert i-j\rvert$	$p_{x-y}(k)=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)$ $k=\lvert i-j\rvert$
	Correlation	Linear dependency of gray level to their respective voxels in GLCM	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij-\mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}$	$p(i,j)=\frac{P(i,j)}{\sum P(i,j)}$
	SumEntropy	Sum of neighborhood intensity value differences	$\sum_{k=2}^{2N_g} p_{x+y}(k)log_2(p_{x+y}(k)+\epsilon)$ $k=i+j$	$p_{x+y}(k)=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)$ $k=i+j$
	DifferenceEntropy	Uncertainty/randomness in neighborhood intensity value difference	$\sum_{k=0}^{N_g-1} p_{x-y}(k)log_2(p_{x-y}(k)+\epsilon)$ $k=\lvert i-j\rvert$	
GLRLM $P(i,j\lvert\theta)$	ShortRunEmphasis	Distribution of short **run lengths***	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\lvert\theta)}{j^2}}{N_r(\theta)}$	$p(i,j\lvert\theta)=\frac{P(i,j\lvert\theta)}{N_r(\theta)}$

[Note] More detailed definitions in the PyRadiomics documentation (https://pyradiomics.readthedocs.io/en/latest/index.html).
$\boldsymbol{N_g}$: The number of discrete intensity gray levels in the image / $\boldsymbol{N_r}$**:** The number of discrete run lengths in the image / $\boldsymbol{N_r(\theta)}$**:** The number of runs in the image along angle, $\boldsymbol{\theta}$
run length*: The number of consecutive pixels that have the same gray level value.

method will be practical in supporting clinical decisions and can be applied to other ROIs, such as the brain and breast. For example, Modanwal et al. [4] utilized a modified CycleGAN network to segment breast MRI. However, they experimented with single-to-single data harmonization (GE Healthcare to Siemens scanner). Our method can be applied to multi-vendor breast MRIs. To date, our current study focused on the segmentation performance of the overall lumbar spine structure. However, future work should assess the accuracy of disease-specific regions (e.g., herniated discs) for clinical applications. We are currently in discussions with clinical experts at SNUBH for a new study on clinical practice, including spinal disease classification. For example, Šušteršič et al. [29] used the basic U-Net model to segment the L4 and L5 lumbar spines for disc disease classification. Compared to their segmentation performance, our proposed model showed improved performance, as presented in Table 5. Therefore, the accuracy of the disease classification could be improved even with multi-vendor MRI using our proposed method. Furthermore, for a simpler approach in a classification study, integrating a DL-based classification model into our harmonization model instead of the segmentation network used in this work could enable direct disease classification without the need for segmentation. Additionally, accurately reconstructing a 3D spine model from each segmented vertebral body from CT and intervertebral disc from MRI is crucial for IGS [57], [58]. Therefore, we plan to investigate translation between different modalities (e.g., CT to MRI) to segment the intervertebral disc directly from CT images. In a preliminary study, we applied the same framework used in this study to convert CT scans to MR images and simultaneously segment the vertebral body and intervertebral disc using multi-vendor CT and MRI. The visual results (segmentation predictions) were promising but quantitative evaluation requires a paired dataset consisting of CT and MRI from the same patients. If our proposed MRI harmonization network can translate from multi-vendor CT to MRI, we can reconstruct a 3D spine model solely from CT images using a single harmonization network, eliminating the need for MRI.

VI. CONCLUSION

We demonstrated the feasibility of the proposed DL-based method for harmonizing multi-vendor MRI using radiomics. The segmentation networks were integrated into a harmonization model to enhance features related to IVD segmentation using dice loss functions. The newly designed RF loss function in this study also contributed to harmonizing image features. These two strategies significantly improved the accuracy of IVD segmentation in the face of variability induced by different vendors, scan parameters, and diseases. The performance of our model surpassed that of other methods, including state-of-the-art techniques like nnU-Net.

TABLE 7. Comparison of various segmentation networks on test source images [mean, 95% confidence intervals].

Model	Dice	IoU	F1 score	precision	recall
ResNet34 + U-Net	0.603, 0.077	0.531, 0.072	0.603, 0.077	**0.971, 0.005**	0.551, 0.075
ResNet34 + U-Net++	0.594, 0.075	0.520, 0.071	0.596, 0.076	0.930, 0.024	0.555, 0.077
ResNet34 + DeepLabV3	**0.862, 0.015**	**0.764, 0.020**	**0.862, 0.015**	0.909, 0.006	**0.830, 0.023**
ResNet50 + U-Net	0.693, 0.055	0.588, 0.056	0.693, 0.055	0.947, 0.020	0.615, 0.060
ResNet50 + U-Net++	0.714, 0.061	0.627, 0.060	0.714, 0.061	0.967, 0.004	0.649, 0.060
ResNet50 + DeepLabV3	0.813, 0.034	0.711, 0.037	0.813, 0.034	0.910, 0.007	0.775, 0.042
ResNet101 + U-Net	0.709, 0.058	0.613, 0.056	0.709, 0.058	0.960, 0.006	0.639, 0.060
ResNet101 + U-Net++	0.678, 0.058	0.574, 0.056	0.678, 0.058	0.962, 0.006	0.597, 0.060
ResNet101 + DeepLabV3	0.654, 0.067	0.568, 0.065	0.654, 0.068	0.940, 0.008	0.607, 0.070

In summary, the proposed model has the potential to benefit clinical practice, including medical imaging segmentation, disease diagnosis, and treatment planning, even with diverse MRIs in various ROIs (e.g., brain and breast). For future work, we will explore spine disease classification using our harmonization method to overcome misdiagnosis caused by MRI variability. For precise and safe spine surgical navigation, we plan to investigate the translation from multi-vendor CT scans to a target MR image, allowing the segmentation of the vertebral body and intervertebral disc exclusively from CT scans. This approach has the potential to save time, cost, and effort by eliminating the need for both CT and MRI data acquisition.

APPENDIX A
SELECTED RADIOMIC FEATURES

Five radiomic features, which acted as the archetypal features of the target images, were selected using statistical methods, as described in Section III-B. Detailed descriptions of these features are presented in Table 6.

APPENDIX B
SEGMENTATION PERFORMANCE OF ALL COMBINATIONS (ENCODER+DECODER)

Diverse combinations of encoders and decoders (from SMP) in Table 2 performed segmentation using the test source dataset. Table 7 shows that ResNet34+DeepLabV3 exhibited the best performance.

REFERENCES

[1] B. E. Dewey, C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, P. C. M. van Zijl, and J. L. Prince, "DeepHarmony: A deep learning approach to contrast harmonization across scanner changes," *Magn. Reson. Imag.*, vol. 64, pp. 160–170, Dec. 2019, doi: 10.1016/j.mri.2019.05.041.

[2] Z. Hu, Q. Zhuang, Y. Xiao, G. Wu, Z. Shi, L. Chen, Y. Wang, and J. Yu, "MIL normalization—Prerequisites for accurate MRI radiomics analysis," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104403, doi: 10.1016/j.compbiomed.2021.104403.

[3] Y. Nan, J. D. Ser, S. Walsh, C. Schonlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, and B. Ernst, "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions," *Inf. Fusion*, vol. 82, pp. 99–122, Jun. 2022, doi: 10.1016/j.inffus.2022.01.001.

[4] G. Modanwal, A. Vellal, and M. A. Mazurowski, "Normalization of breast MRIs using cycle-consistent generative adversarial networks," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106225, doi: 10.1016/j.cmpb.2021.106225.

[5] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage, Clin.*, vol. 6, pp. 9–19, Jan. 2014, doi: 10.1016/j.nicl.2014.08.008.

[6] E. Stamoulou, C. Spanakis, G. C. Manikis, G. Karanasiou, G. Grigoriadis, T. Foukakis, M. Tsiknakis, D. I. Fotiadis, and K. Marias, "Harmonization strategies in multicenter MRI-based radiomics," *J. Imag.*, vol. 8, no. 11, p. 303, Nov. 2022, doi: 10.3390/jimaging8110303.

[7] Y. Li, S. Ammari, C. Balleyguier, N. Lassau, and E. Chouzenoux, "Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features," *Cancers*, vol. 13, no. 12, p. 3000, Jun. 2021, doi: 10.3390/cancers13123000.

[8] N. Robitaille, A. Mouiha, B. Crépeault, F. Valdivia, and S. Duchesne, "Tissue-based MRI intensity standardization: Application to multicentric datasets," *Int. J. Biomed. Imag.*, vol. 2012, May 2012, Art. no. 347120, doi: 10.1155/2012/347120.

[9] M. Vania, D. Mureja, and D. Lee, "Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels," *J. Comput. Des. Eng.*, vol. 6, no. 2, pp. 224–232, Apr. 2019, doi: 10.1016/j.jcde.2018.05.002.

[10] C. Kim, O. Bekar, H. Seo, S.-M. Park, and D. Lee, "Computed tomography vertebral segmentation from multi-vendor scanner data," *J. Comput. Design Eng.*, vol. 9, no. 5, pp. 1650–1664, Sep. 2022, doi: 10.1093/jcde/qwac072.

[11] C. A. Linte, J. White, R. Eagleson, G. M. Guiraudon, and T. M. Peters, "Virtual and augmented medical imaging environments: Enabling technology for minimally invasive cardiac interventional guidance," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 25–47, 2010, doi: 10.1109/RBME.2010.2082522.

[12] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, and E. Neri, "A deep look into radiomics," *La Radiologia medica*, vol. 126, no. 10, pp. 1296–1311, Jul. 2021, doi: 10.1007/s11547-021-01389-x.

[13] K. Fatania, F. Mohamud, A. Clark, M. Nix, S. C. Short, J. O'Connor, A. F. Scarsbrook, and S. Currie, "Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—A systematic review," *Eur. Radiol.*, vol. 32, no. 10, pp. 7014–7025, Apr. 2022, doi: 10.1007/s00330-022-08807-2.

[14] A. Zwanenburg, M. Vallieres, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and M. Bogowicz, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.

[15] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, and L. H. Schwartz, "Reproducibility of radiomics for deciphering tumor phenotype with imaging," *Sci. Rep.*, vol. 6, no. 1, p. 23428, Mar. 2016, doi: 10.1038/srep23428.

[16] S. Fiset, M. L. Welch, J. Weiss, M. Pintilie, J. L. Conway, M. Milosevic, A. Fyles, A. Traverso, D. Jaffray, U. Metser, J. Xie, and K. Han, "Repeatability and reproducibility of MRI-based radiomic features in cervical cancer," *Radiotherapy Oncol.*, vol. 135, pp. 107–114, Jun. 2019, doi: 10.1016/j.radonc.2019.03.001.

[17] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014, doi: 10.1007/s12021-013-9204-3.

[18] A. Fateh, M. Fateh, and V. Abolghasemi, "Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning," *Inf. Sci.*, vol. 581, pp. 479–494, Dec. 2021, doi: 10.1016/j.ins.2021.09.051.

[19] L. G. Nyúl and J. K. Udupa, "On standardizing the MR image intensity scale," *Magn. Reson. Med.*, vol. 42, no. 6, pp. 1072–1081, Nov. 1999, doi: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[20] A. Chaddad, M. J. Kucharczyk, P. Daniel, S. Sabri, B. J. Jean-Claude, T. Niazi, and B. Abdulkarim, "Radiomics in glioblastoma: Current status and challenges facing clinical implementation," *Frontiers Oncol.*, vol. 9, p. 374, May 2019, doi: 10.3389/fonc.2019.00374.

[21] N. L. Weisenfeld and S. K. Warfield, "Normalization of joint image-intensity statistics in MRI using the Kullback–Leibler divergence," in *Proc. IEEE 2nd Int. Symp. Biomed. Imag. (ISBI)*, Arlington, VA, USA, Apr. 2004, pp. 101–104, doi: 10.1109/ISBI.2004.1398484.

[22] F. Jager, Y. Deuerling-Zheng, B. Frericks, F. Wacker, and J. Hornegger, "A new method for MRI intensity standardization with application to lesion detection in the brain," in *Proc. VMV*, Aachen, Germany, 2006, p. 269.

[23] M. Selim, J. Zhang, B. Fei, G. Q. Zhang, and J. Chen, "STAN-CT: Standardizing CT image using generative adversarial networks," in *Proc. AMIA Annu. Symp.*, 2020, p. 1100. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075475/

[24] G. Liang, S. Fouladvand, J. Zhang, M. A. Brooks, N. Jacobs, and J. Chen, "GANai: Standardizing CT images using generative adversarial network with alternative improvement," in *Proc. IEEE Int. Conf. Healtc. Inform. (ICHI)*, Xi'an, China, Jun. 2019, pp. 1–11, doi: 10.1109/ICHI.2019.8904763.

[25] Y. Xu, Y. Li, and B.-S. Shin, "Medical image processing with contextual style transfer," *Hum.-Centric Comput. Inf. Sci.*, vol. 10, no. 1, pp. 1–16, Nov. 2020, doi: 10.1186/s13673-020-00251-9.

[26] T. DeSilvio, S. Moroianu, I. Bhattacharya, A. Seetharaman, G. Sonn, and M. Rusu, "Intensity normalization of prostate MRIs using conditional generative adversarial networks for cancer detection," *Proc. SPIE*, vol. 11597, pp. 121–126, Feb. 2021, doi: 10.1117/12.2582297.

[27] H. Guan, Y. Liu, E. Yang, P.-T. Yap, D. Shen, and M. Liu, "Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102076, doi: 10.1016/j.media.2021.102076.

[28] Y. Gao, Y. Liu, Y. Wang, Z. Shi, and J. Yu, "A universal intensity standardization method based on a many-to-one weak-paired cycle generative adversarial network for magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2059–2069, Sep. 2019, doi: 10.1109/TMI.2019.2894692.

[29] T. Sustersic, V. Rankovic, V. Milovanovic, V. Kovacevic, L. Rasulic, and N. Filipovic, "A deep learning model for automatic detection and classification of disc herniation in magnetic resonance images," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 6036–6046, Dec. 2022, doi: 10.1109/JBHI.2022.3209585.

[30] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232, doi: 10.48550/arXiv.1703.10593.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[32] S. Maeda, "Unpaired image super-resolution using pseudo-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 288–297.

[33] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[34] S. Sudirman et al., "Lumbar spine MRI dataset," *Mendeley Data V2*, Apr. 2019, doi: 10.17632/k57fr854j2.2.

[35] Y. A. Khalil, "Lumbar vertebral body and intervertebral disc segmentation in multi-scanner and multi-modal MRI—A ground truth database," OSF, Jul. 2023, doi: 10.17605/OSF.IO/QX5RT.

[36] J. W. van der Graaf et al., "SPIDER—Lumbar spine segmentation in MR images: A dataset and a public benchmark," Zenodo, Nov. 2023, doi: 10.5281/zenodo.10159290.

[37] A. S. Al-Kafri, S. Sudirman, A. Hussain, D. Al-Jumeily, F. Natalia, H. Meidia, N. Afriliana, W. Al-Rashdan, M. Bashtawi, and M. Al-Jumaily, "Boundary delineation of MRI images for lumbar spinal stenosis detection through semantic segmentation using deep neural networks," *IEEE Access*, vol. 7, pp. 43487–43501, 2019, doi: 10.1109/ACCESS.2019.2908002.

[38] Y. A. Khalil, E. A. Becherucci, J. S. Kirschke, D. C. Karampinos, M. Breeuwer, T. Baum, and N. Sollmann, "Multi-scanner and multi-modal lumbar vertebral body and intervertebral disc segmentation database," *Sci. Data*, vol. 9, no. 1, p. 97, Mar. 2022, doi: 10.1038/s41597-022-01222-8.

[39] J. W. van der Graaf, M. L. van Hooff, C. F. M. Buckens, M. Rutten, J. L. C. van Susante, R. Jan Kroeze, M. de Kleuver, B. van Ginneken, and N. Lessmann, "Lumbar spine segmentation in MR images: A dataset and a public benchmark," 2023, *arXiv:2306.12217*.

[40] G. Sze, Y. Kawamura, C. Negishi, R. T. Constable, M. Merriam, K. Oshio, and F. Jolesz, "Fast spin-echo MR imaging of the cervical spine: Influence of echo train length and echo spacing on image contrast and quality," *AJNR Am. J. Neuroradiol.*, vol. 14, no. 5, pp. 1203–1213, 1993. [Online]. Available: https://www.ajnr.org/content/14/5/1203

[41] N. M. Major and M. W. Anderson, "Basic principles of musculoskeletal MRI," in *Musculoskeletal MRI*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2020, ch. 1, pp. 1–22.

[42] S. Abdulla, and C. Clarke, "MR imaging," in *FRCR Physics Notes: Medical Imaging Physics for the First FRCR Examination*, 3rd ed. London, U.K.: Radiology Cafe, 2020, ch. 5, pp. 139–202.

[43] V. Parekh and M. A. Jacobs, "Radiomics: A new application from established techniques," *Exp. Rev. Precis. Med. Drug Develop.*, vol. 1, no. 2, pp. 207–226, Mar. 2016, doi: 10.1080/23808993.2016.1164013.

[44] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Oct. 2017, doi: 10.1158/0008-5472.can-17-0339.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711, doi: 10.1007/978-3-319-46475-6_43.

[46] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[47] P. Yakubovskiy. (2022). *Segmentation Models Pytorch*. GitHub Repository. Accessed: May 2, 2023. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[49] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[50] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[51] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J. C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, and J. Buatti, "3D slicer as an image computing platform for the quantitative imaging network," *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.

[52] P. Chirra, P. Leo, M. Yim, B. N. Bloch, A. R. Rastinehad, A. Purysko, M. Rosen, A. Madabhushi, and S. E. Viswanath, "Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI," *J. Med. Imag.*, vol. 6, no. 2, p. 1, Jun. 2019, Art. no. 024502, doi: 10.1117/1.jmi.6.2.024502.

[53] J. D. Shur, S. J. Doran, S. Kumar, D. Dafydd, K. Downey, J. P. B. O'Connor, N. Papanikolaou, C. Messiou, D.-M. Koh, and M. R. Orton, "Radiomics in oncology: A practical guide," *RadioGraphics*, vol. 41, no. 6, pp. 1717–1732, Oct. 2021, doi: 10.1148/rg.2021210037.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[55] R. F. Masood, I. A. Taj, M. B. Khan, M. A. Qureshi, and T. Hassan, "Deep learning based vertebral body segmentation with extraction of spinal measurements and disorder disease classification," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103230, doi: 10.1016/j.bspc.2021.103230.

[56] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[57] B. Felix, S. B. Kalatar, B. Moatz, C. Hofstetter, M. Karsy, R. Parr, and W. Gibby, "Augmented reality spine surgery navigation: Increasing pedicle screw insertion accuracy for both open and minimally invasive spine surgeries," *Spine*, vol. 47, no. 12, pp. 865–872, Jun. 2022, doi: 10.1097/brs.0000000000004338.

[58] K. McCloskey, R. Turlip, H. S. Ahmad, Y. G. Ghenbot, D. Chauhan, and J. W. Yoon, "Virtual and augmented reality in spine surgery: A systematic review," *World Neurosurg.*, vol. 173, pp. 96–107, May 2023, doi: 10.1016/j.wneu.2023.02.068.

CHAEWOO KIM received the B.S. degree in physics from Sung Kyun Kwan University, Suwon, Republic of Korea, in 2010, the M.S. degree in physics from Yonsei University, Seoul, Republic of Korea, in 2014, and the M.Sc. degree in cognitive and computational neuroscience from The University of Sheffield, Sheffield, U.K., in 2017. She is currently pursuing the Ph.D. degree in AI robotics with the Seoul National University of Science and Technology, Seoul.

From 2014 to 2015, she was a Research Assistant with the Samsung Medical Center, Seoul. Her research interests include deep learning methods for clinical applications and computational neuroscience.

Ms. Kim's awards and honors include The University of Sheffield South Korea Postgraduate Merit Scholarship and the Full Scholarship from Brain Korea (BK) 21 Project for Physics.

SANG-MIN PARK received the B.S. and M.D. degrees from the College of Medicine, Chung-Ang University, Seoul, Republic of Korea, in 2010, and the M.S. degree in orthopedic science from the College of Medicine, Chung-Ang University, in 2014. He is currently pursuing the Ph.D. degree in clinical medical sciences with the College of Medicine, Seoul National University, Seoul.

From 2017 to 2021, he was an Assistant, and since 2021, he has been an Associate Professor with the Spine Center and the Department of Orthopaedic Surgery, Seoul National University College of Medicine; and Seoul National University Bundang Hospital, respectively. He is the author of three books and more than 80 articles. He holds two patents. His research interests and specialties include degenerative lumbar spine surgery, minimally invasive spine surgery, endoscopic spine surgery, osteoporosis, spine tumor, mixed reality, and medical twin.

Dr. Park received more than ten honors and awards. His recent honors and awards include the Best Video Presentation Award and the Best Scientific Award (Clinical) at the 67th Annual Congress of the Korean Orthopaedic Association 2023. He is an Editor of the *Journal of the Korean Orthopaedic Association* (JKOA), *Asian Spine Journal* (ASJ), and *Journal of Advanced Spine Surgery* (JASS).

SANGHOON LEE received the B.S. and M.D. degrees from the College of Medicine, Yonsei University, Seoul, Republic of Korea, in 2018. He is currently pursuing the M.S. degree with the Department of Orthopaedic Surgery, College of Medicine, Seoul National University, Seoul.

DEUKHEE LEE (Member, IEEE) received the B.S. degree in mechanical engineering from Han Yang University, Seoul, Republic of Korea, in 2000, the M.S. degree in mechanical engineering from Seoul National University, Seoul, in 2003, and the Ph.D. degree in mechanical engineering from The University of Tokyo, Japan, in 2008.

He is currently a Principal Researcher with the Korea Institute of Science and Technology, Seoul, and the Yonsei-KIST Convergence Research Institute, Yonsei University, Seoul. He is the author of more than 20 articles and holds more than 40 patents. His research interests include medical image analysis, surgical navigation, and medical robotics.

Dr. Lee is a Committee Member of the Society for Computational Design Engineering. He received more than eight honors and awards. His recent honors and awards include the Best Scientific Award at KIST in 2021. He is an Associate Editor of the *Journal of Computational Design and Engineering*.