DPI-MoCo: Deep Prior Image Constrained Motion Compensation Reconstruction for 4D CBCT

Dianlin Hu, Chencheng Zhang, Xuanjia Fei, Yi Yao, Yan Xi, Jin Liu, Yikun Zhang, Gouenou Coatrieux, Jean-Louis Coatrieux, Yang Chen

DPI-MoCo: Deep Prior Image Constrained Motion Compensation Reconstruction for 4D CBCT

Dianlin Hu, ChenCheng Zhang, Xuanjia Fei, Yi Yao, Yan Xi, Jin Liu, Yikun Zhang, Gouenou Coatrieux, *Senior Member*, *IEEE*, Jean Louis Coatrieux, *Life Fellow*, *IEEE*, Yang Chen, *Senior Member*, *IEEE*

***Abstract*—4D cone-beam computed tomography (CBCT) plays a critical role in adaptive radiation therapy for lung cancer. However, extremely sparse sampling projection data will cause severe streak artifacts in 4D CBCT images. Existing deep learning (DL) methods heavily rely on large labeled training datasets which are difficult to obtain in practical scenarios. Restricted by this dilemma, DL models often struggle with simultaneously retaining dynamic motions, removing streak degradations, and recovering fine details. To address the above challenging problem, we introduce a Deep Prior Image Constrained Motion Compensation framework (DPI-MoCo) that decouples the 4D CBCT reconstruction into two sub-tasks including coarse image restoration and structural detail fine-tuning. In the first stage, the proposed DPI-MoCo combines the prior image guidance, generative adversarial network, and contrastive learning to globally suppress the artifacts while maintaining the respiratory movements. After that, to further enhance the local anatomical structures, the motion estimation and compensation technique is adopted. Notably, our framework is performed without the need for paired datasets, ensuring practicality in clinical cases. In the Monte Carlo simulation dataset, the DPI-MoCo achieves competitive quantitative performance compared to the state-of-the-art (SOTA) methods. Furthermore, we test DPI-MoCo in clinical lung cancer datasets, and experiments validate that DPI-MoCo not only restores small anatomical structures and lesions but also preserves motion information.**

***Index Terms*—4D CBCT imaging, motion compensation, prior image, generative adversarial loss, contrastive learning.**

I. Introduction

ON-board cone-beam computed tomography (CBCT) mounted on a linear accelerator is an effective imaging tool in image-guided radiation therapy (IGRT) [1] because it can flexibly provide three-dimensional anatomical information and correct any changes in patient setup or target localization [2]. However, owing to the respiratory movements, CBCT will suffer from severe degradation caused by motion-induced artifacts.

Later, four-dimensional CBCT (4D CBCT) was introduced to tackle the motion-blurred artifacts. And it has played an important role in adaptive radiation therapy (ART) for lung cancer [3, 4] because of its ability to track the movements of organs and tissues in real time. However, this can also lead to severe streak artifacts in 4D CBCT because the projection data of each phase-resolved image (PRI) is extremely sparse, which undoubtedly reduces its clinical values [5]. Currently, how to preserve the dynamic changes of breathing motion while reconstructing high-quality PRIs at the same time is the main challenge in 4D CBCT. Many efforts have been devoted to solving this issue and they can be roughly grouped into three categories: iterative (IR) methods, motion compensation (MoCo) methods, and deep learning (DL)-based methods.

By incorporating a variety of regularization terms into the optimization process [6], IR methods perform better than the typical Feldkamp-David-Kress (FDK) algorithm [7]. Among them, total variation (TV) is the most widely used constraint

This work was supported in part by the National Key Research and Development Program of China (2022YFE0116700), in part by the National Key Research and Development Program of China (No. 2021ZD0113202), in part by the State Key Project of Research and Development Plan under Grant 2022YFC2401600, in part by the National Natural Science Foundation of China under Grant T2225025, in part by the Key Research and Development Programs in Jiangsu Province of China under Grant BE2021703 and BE2022768 (Corresponding authors: Yikun Zhang, Yang Chen).

Dianlin Hu, ChenCheng Zhang and Yikun Zhang are with the Laboratory of Image Science and Technology, Southeast University, Nanjing 210096, China (e-mail: dianlinhu@gmail.com, 1920497925@qq.com; yikun@seu.edu.cn).

Jin Liu is with the College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China (e-mail: liujin@ahpu.edu.cn)

Gouenou Coatrieux is with the IMT Atlantique, Inserm, LaTIM UMR1101, Brest 29000, France (e-mail:gouenou.coatrieux@imt-atlantique.fr).

Jean-Louis Coatrieux is with the Laboratoire Traitement du Signal et de l'Image, Universit ´e de Rennes 1, F-35000 Rennes, France (e-mail: jean-louis.coatrieux@univ-rennes1.fr).

Xuanjia Fei and Yi Yao are with the LinaTech Company Ltd., Suzhou, China (e-mail: xjfei@linatech.com.cn, jonathanyao@linatech.com.cn).

Yan Xi is with the Jiangsu First-imaging Medical Equipment Co., Ltd., Nantong 226100, China (e-mail: yanxi@first-imaging.com).

Yang Chen is with the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, the Laboratory of Image Science and Technology, the School of Computer Science and Engineering, and the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing 210096, China (e-mail: chenyang.list@seu.edu.cn).

function. Via minimizing the magnitudes of gradient images from PRIs, the TV method successfully promotes the 4D CBCT images in noise and artifact suppression [8, 9]. Next, to fully exploit the temporal interrelation between adjacent PRIs, some works propose 4D TV to regularize the reconstructed images in the spatial-sequential domain [10-12]. Besides, prior-image-guided TV algorithms are another popular scheme for 4D CBCT reconstruction [13, 14]. These studies are driven by the observation that the difference maps between the individual PRI and prior images reconstructed from full-sampled projection data are sparse at motion regions, which can greatly boost the final reconstruction results compared to classical TV methods [13]. Nevertheless, IR-based methods often lead to over-smoothing results and loss of small details.

The MoCo-based methods are also commonly employed for 4D CBCT reconstruction, whose key roadmap is to compensate for the respiratory motion of target PRI by applying deformation vector fields (DVFs) [15-17]. To estimate the patient-specific motion mode, the direct way is to extract the DVFs from the 4D planning CT (pCT) of the same patient and then apply the estimated DVFs to correct deformed images based on the FDK algorithm [18-20]. However, this strategy follows the assumption that the motion pattern in acquired CBCT data is identical to the ones in pCT, which is often violated in clinical scenarios. Afterward, several hybrid approaches that integrate the smoothness constraint of respiratory motion into IR optimization are proposed to gradually improve the DVFs [21-23]. Compared to [18], results demonstrate that these works achieve a boosted temporal resolution of PRIs and a more accurate tumor motion trajectory. Although MoCo-based methods can generate robust movement information, it is difficult to produce visually improved PRIs.

To mitigate streak artifacts and noise, many convolutional neural networks (CNNs) have been probed [24, 25]. One popular scheme is to take the artifact-induced images as input and produce high-quality results [26, 27]. Benefiting from the powerful feature extraction capacity, these models can provide better images with clearer edges and enhanced anatomical structures than conventional methods. However, on account of extremely sparse projection, it is hard for simple CNNs to recover satisfactory PRIs in 4D CBCT. Inspired by the PICCS [13], some studies attempt to input the PRI and prior image jointly into the CNN model and bring excellent promotions in artifact removal [28-30]. In addition, the combination of the CNN model and MoCo or IR optimization is also an effective way to boost PRIs [28, 31, 32]. It is worth noting that there is no ground truth image in clinical 4D CBCT. For this reason, the abovementioned DL-based methods cannot guarantee their worst performance when transferring the trained model on the simulated dataset to the real CBCT images. To overcome this issue, a specialized projection selection scheme is designed to construct a pseudo-paired dataset [33]. In this way, existing CNN models can be flexibly optimized on this dataset [34] and therefore get rid of the inter-domain inconsistencies that occurred in [30, 31]. However, it remains a challenging problem that simultaneously retains dynamic motions, removes streak degradations, and restores fine details.

In this study, we propose a Deep Prior Image Constrained Motion Compensation (DPI-MoCo) reconstruction framework for 4D CBCT. It decomposes the 4D CBCT into two sub-tasks including coarse image generation and structural features fine-tuning, which collectively improve the PRIs from the global and local aspects. Compared with [31], our DPI-MoCo only adapts one motion compensation step and generates satisfactory results, which has high computational efficiency. In addition, the proposed method gets rid of the labeled data, which is important for practical 4D CBCT applications. Besides, unlike [34], DPI-MoCo constructs a more complex and effective framework, including pseudo-paired dataset construction, deep model training, and motion compensation, as a result, leading to promising results. The main contributions of DPI-MoCo are four-fold.

- We propose a specialized framework for clinical 4D CBCT reconstruction, which eliminates the need for ground truth data, allowing it to be applied to various imaging conditions and vendors.
- We employ the prior-image guided artifact estimation module (PIGAE) to construct a more effective pseudo-paired dataset than [34], which not only supplies higher-quality CBCT images for training RestoreNet and RegisNet but also offers a better candidate for MoCo that can further promote the reconstruction image.
- Unlike the existing hybrid methods that combine deep learning and MoCo [31, 34], the proposed DPI-MoCo additionally designs a prior-image guided DenoiseNet after the MoCo step to facilitate streaking artifact removal at some areas with fewer changes during breathing.
- Experimental results demonstrate that on the simulated dataset, our DPI-MoCo even surpasses the existing competitive methods trained on the ground truth data. Moreover, on the clinical datasets, DPI-MoCo performs well both in stationary structure preservation and dynamic movement tracking. This proves the proposed method is promising for real 4D CBCT applications.

The rest of this paper is organized as follows. Section II presents the preliminary work related to the DPI-MoCo. The detailed descriptions of the proposed method will be given in section III. In section IV, the simulated and clinical experiments are performed. Section V will discuss some issues and conclude.

II. Background

A. Phase-Resolved Image Reconstruction

For the PRI reconstruction, the fully-sampled projection data $P \in R^{M\times N\times V}$ is first split into different phases according to the respiratory signal, where M and N are the height and width of the plat detector, respectively, and V indicates the projection numbers. Then, the PRI can be reconstructed as follows:

$$x_n = \mathrm{FDK}(P_n) \tag{1}$$

where every projection data in $P_n \in R^{M\times N\times V_n}$ belongs to the n^{th} respiratory phase and V_n is the number of projection data in P_n, x_n represents the PRI reconstructed from P_n.

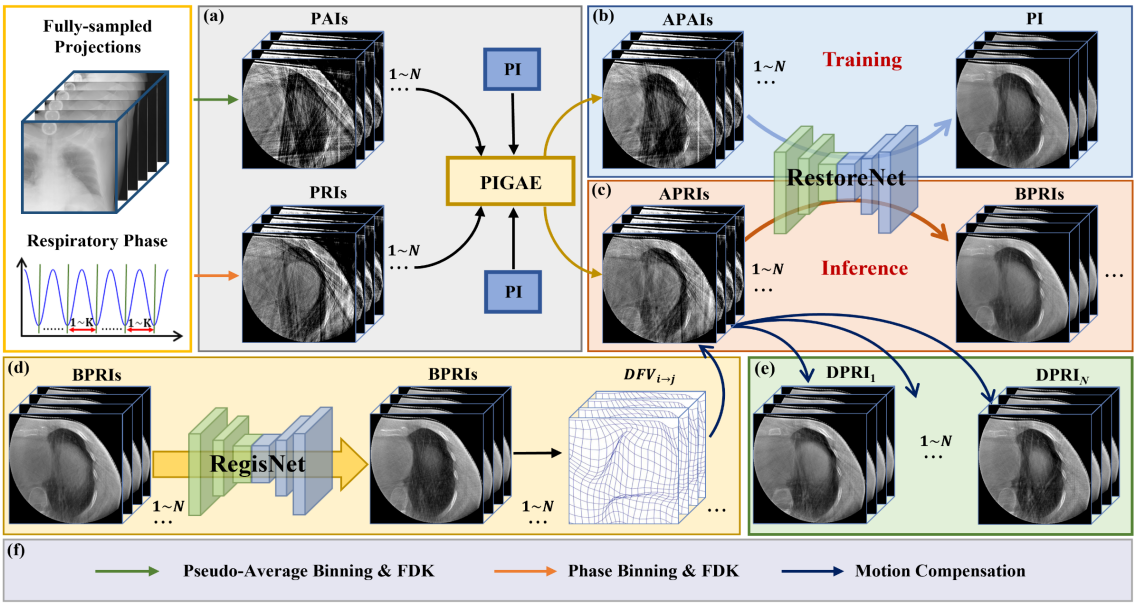

Fig 1. The overview of the proposed DPI-MoCo framework. (a) Prior-image guided artifact estimation module. (b) RestoreNet training on the pseudo-paired dataset. (c) Global artifact removal inference for APRIs. (d) Motion estimation between arbitrary BPRIs. (e) Motion compensation and denoising. (a)-(c) Coarse image generation. (d)-(e) Anatomical structure fine-tuning. PAI: Pseudo-average image, PI: Prior-Image, PRI: Phase-Resolved Image, APAI: Artifact-reduced PAI, APRI: Artifact-reduced PRI, BPRI: Boosted PRI, DPRI: Denoised PRI.

B. *Self-Contained Deep-Learning-Based Artifact Removal*

In real 4D CBCT applications, ground truth data is typically unavailable, so it is difficult to reconstruct reliable PRIs for practical scenarios. To address this matter, Madesta et al. proposed a novel projection binning scheme to build the pseudo-paired samples for instructing CNN optimization [33].

The pseudo-average projection subset P_n^a ($n \in [1, N_b]$) construction can be simply described as follows.

$$n_a \leftarrow n, r \leftarrow 1, P_n^a \leftarrow \{\} \tag{2}$$

$$P_n^a = P_n^a \cup P_{n_a,r} \tag{3}$$

$$n_a \leftarrow (n_a + 1) \bmod N_b, r \leftarrow r + 1 \tag{4}$$

$$n_a \leftarrow N_b, if\ n_a == 0 \tag{5}$$

where N_b is the total respiratory phase number, $P_{n_a,r}$ corresponds to projection that belongs to the n^{th} respiratory phase in the r^{th} breathing circle. Specifically, Eq. (2) is the initialization step, then repeat the Eqs. (3)-(5) until r exceeds the total breathing circles. Eventually, pseudo-paired samples $\{x_n^a, PI\}$ can be constructed via pseudo-average image (PAI) $x_n^a = \text{FDK}(P_n^a)$ and prior image $PI = \text{FDK}(P)$. As described above, the x_n^a not only shares similar streak artifacts with the PRI x_n, but also characterizes the same motion mode with PI. Consequently, the network trained by regarding the x_n^a as the artifact-corrupted input and PI as the high-quality ground truth is able to diminish the artifacts of x_n.

C. *Motion Compensation-Based Reconstruction*

In contrast to Eq. (1), MoCo-based methods utilize the full-sampled projection data to reconstruct the motion-free target PRI by applying DVFs. To boost the computational efficiency, Brehm et al. directly deformed the individual PRI and added them to achieve the final results, which is equal to performing a backprojection of filtered data along motion direction [16]. The MoCo reconstruction in [16] is briefly formulated as:

$$m_j = \sum_{n=1}^{N} DVF_n^j(x_n) \tag{6}$$

where DVF_n^j is a transformation that can register x_n from phase n to j via $DVF_n^j(x_n)$, and m_j presents the deformed high-quality target image.

III. Methodology

The proposed DPI-MoCo framework adopts a multi-phase processing strategy to gradually improve the PRIs (as illustrated in Fig. 1).

A. *Prior-Image Guided Artifact Estimation*

The pseudo-paired samples $\{x_n^a, PI\}$ overcome the lack of ground truth 4D CBCT images to some degree. However, as reported in [33, 34], the well-trained model obtained on this dataset still results in over-smoothing details. Following the observation that the prior image possesses sharp edges and clear structures at the stationary regions during breathing, our proposed DPI-MoCo first develops a PIGAE module (as demonstrated in Fig. 1 (a)) to promote the PRIs. Specifically, the PIGAE is described as:

$$s_n^a = PI - \text{FDK}(A_n^a PI) \tag{7}$$

$$s_n = PI - \text{FDK}(A_n PI) \tag{8}$$

where A_n^a and A_n are the system matrices. Particularly, A_n^a can project the PI along the same direction as the data in P_n^a, resulting in projection data of the same size as P_n^a. Similarly, A_n can project the PI along the same direction as the data in P_n, also resulting in projection data of the same size as P_n. Taking the PI as the ground truth, A_n^a and A_n aim to reproduce the degradations that occurred in x_n^a and x_n. Consequently, s_n^a and s_n indicate the similar streaking artifacts contained in x_n^a and x_n. Thus, the corrected results can be accessed:

$$c_n^a = x_n^a - s_n^a \tag{9}$$

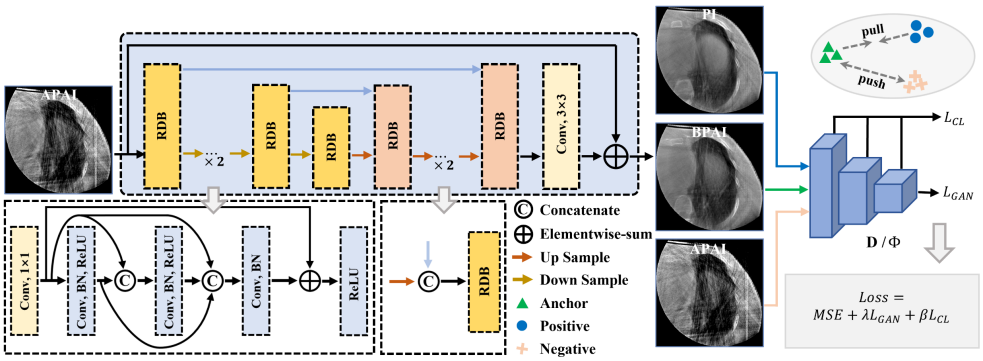

Fig.2. The architecture of the proposed RestoreNet.

$$c_n = x_n - s_n \tag{10}$$

here c_n^a and c_n are artifact-reduced PAI and PRI, denoted by APAI and ARPI, respectively. After being processed by the PIGAE module, c_n^a and c_n become more robust to bone or muscle structures. As a result, we can get the improved pseudo-average paired dataset $\{c_n^a, PI\}$.

B. *RestoreNet*

As illustrated in Fig. 1(b) and Fig. 2, the RestoreNet is adopted in this work to generate boosted PRIs (BPRIs). Usually, the RestoreNet is optimized with L_2 or L_1 loss and prefers to blur the CBCT images, which is the inherent shortcoming of CNN models [35]. To alleviate this problem, the GAN and CL techniques are simultaneously conducted to upgrade the minor feature preservation of RestoreNet in this work. In contrast to calculating the distance between the generated image and the ground truth pixel-by-pixel, GAN intends to bring the source distribution and the target counterpart closer. Various image recovery tasks have validated that GAN indeed encourages to supply of visually improved results [35, 36]. In addition, CL is another popular practice that provides realistic images [37]. There are usually three necessary elements in CL, including anchor, positive sample, and negative sample, which respectively correspond to the restored images, degraded images, and ground truth in this work. The goal of CL is to reduce the distance between the anchor and positive samples but put away the anchor from negative ones at the same time.

Generally, the GAN or CL is independently regularized on the CNN model [35, 36]. To synergistically exploit the advantages of GAN and CL, the DPI-MoCo binds them together to guide the RestoreNet. Similar to existing works [35, 38, 39], there are two loss functions that need to be optimized, which are L_D and L_G, respectively. Specifically, the loss function L_D of discriminator D is defined as follows:

$$L_D = -E[D(C(PI, H_{PI}))] + E\left[D\left(C\left(G(c_n^a), H_{G(c_n^a)}\right)\right)\right] + \rho E[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{11}$$

here, we adopt EGAN [36] to enhance the original WGAN further. Where C indicates the concatenation operation and H declares the high-frequency extraction (referred to [36]), and $\hat{x}$ means the linear interpolation between $C\left(G(c_n^a), H_{G(c_n^a)}\right)$ and $C(PI, H_{PI})$. $E[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$ represents the gradient penalty term for network regularization. ρ is a constant weighting parameter and a reasonable value can constrain the gradients to an approximate range to improve the stability of network training.

For the RestoreNet G, its objective function is:

$$L_G = MSE(G(c_n^a), PI) - \lambda E\left[D\left(C\left(G(c_n^a), H_{G(c_n^a)}\right)\right)\right] + \beta \sum_{i=1}^{I} w_i \frac{MSE(\Phi_i(G(c_n^a)), \Phi_i(PI))}{MSE(\Phi_i(G(c_n^a)), \Phi_i(c_n^a))} \tag{12}$$

where MSE is the mean square error, Φ_i and w_i $(i = 1,2, \dots, I)$ are the feature extractor and weight coefficient. λ and β are weighting parameters that correspond to adversarial loss and CL terms, respectively. Both λ and β balance the tradeoff between artifact removal and detail recovery. That means smaller λ or β cannot avoid image blurring caused by MSE but larger values may lead to some negative effects because of the instabilities of GAN or CL training processes. In most works, the pre-trained deep model developed by visual geometry group (VGG) [40] is often regarded as the feature extractor [37, 39, 41, 42]. In opposition to this, in our DPI-MoCo, the discriminator D and feature extractor Φ share the same architecture and learnable variables for better capturing intrinsic CBCT features. Particularly, the 2th, 4th, 6th, 8th convolutional layers in discriminator D (more detailed architecture can be found in [36]) serve as feature extractors. Overall, in this work, the networks D and G are trained alternatively by fixing one and updating the other, with one updating in each network optimization stage.

After training the RestoreNet, the boosted PRI (BPRI) b_n can be obtained by applying the RestoreNet on APRI c_n (as shown in Fig. 1(c)). When the ground truth images are unavailable in clinical 4D CBCT, the restored BPRIs show their advantages in terms of artifact removal than some label-dependent methods [28, 30, 31].

C. *RegisNet*

Even though RestoreNet can produce global artifact-free images, local structural details still need to be refined. MoCo is a promising way to overcome these defects by applying DVFs to utilize all PRIs. The accuracy of estimated DVFs significantly influences the performance of MoCo-based methods. As mentioned before, DVFs provided by pCT cannot precisely characterize the motion mode of CBCT images. Meanwhile, the PRIs reconstructed from the FDK algorithm will disrupt the quality of DVFs caused by streak artifacts. Recently, Zhang et al. first reconstructed initial PRIs following

the [33], and then the DVFs extracted from these were employed to compensate for the breathing motion [34]. Experimental results state that [34] outperforms [33] in tissue recovery and conventional MoCo in artifact reduction.

As exhibited in Fig. 1(d), we use RegisNet to obtain the DVFs between various respiratory phases. There are two important differences between RegisNet and [34]. One is that RegisNet receives better input compared with [34], therefore, resulting in more accurate DVFs. The other is that RegisNet establishes itself on the CNN model, so it has higher computational efficiency. The cost function of RegisNet is offered:

$$L_{RegisNet} = NCC\left(DVF_i^j(b_i), b_j\right) + \sigma Smooth(DVF_i^j) \quad (13)$$

where RegisNet takes the b_i and b_j as input and outputs DVF_i^j. NCC is the normalized cross-correlation loss and $Smooth$ represents the local smoothness constraint by minimizing the gradient of DVF_i^j [43]. σ is the parameter that controls the smoothness of the DVF. Unreasonable σ values could limit the registration performance or cause some unnatural distortions. Specifically, we use VoxelMorph [43] to implement RegisNet in this study.

D. Motion-Free Image Reconstruction

According to [16], the target phase image is reconstructed by adding all other deformed sequential results as computed in Eq. (6). For a given DVF_n^j, there are three candidates to be deformed, including x_n, c_n, and b_n. Routinely, the DVFs are conducted on the x_n to generate results. However, this scheme will leave many artifacts [34] since x_n contains severe corruptions. As for b_n, it loses detailed structures, so the target image also fails to restore them. Compared to x_n, c_n mitigates the artifacts at some stationary regions. Meanwhile, it owns more tiny features than b_n. As a result, c_n can better balance the tradeoff between artifact suppression and detail preservation. Therefore, deformed PRIs of DPI-MoCo are delivered:

$$d_j = \sum_{n=1}^{N} DVF_n^j(c_n) \quad (14)$$

Nevertheless, d_j performs well in detail recovery, it remains slight corruptions in some relatively stationary regions because of the streak artifacts in c_n. Although some advanced denoising algorithms [44-46] have been proposed for unlabeled data, they may not be suitable for application here. To improve the d_j, a specialized prior-image guided DenoiseNet is developed and its loss function is defined:

$$L_{DenoiseNet} = M(O(Y), PI) \quad (15)$$

where $Y = PI + \gamma * \epsilon$ and $\epsilon \sim N(0, \mathrm{I})$ is Gaussian noise, γ controls the noise level, M represents the metrics, O indicates the DenoiseNet. As mentioned above, the DenoiseNet is trained using PI as the ground truth again. Particularly, DenoiseNet adopts the same network architecture as RestoreNet and is also optimized by GAN and CL (identical to Eqs. (11)-(12)). After training the DenoiseNet, the denoised PRIs (DPRIs) can be obtained by $O(d_j + \gamma * \epsilon)$. Since the artifacts in d_j are small, the well-trained DenoiseNet treats them as noise and removes them from $d_j + \gamma * \epsilon$ thus realizing the improvement of d_j. The workflow of DPI-MoCo is summarized in Algorithm 1.

Algorithm 1 DPI-MoCo

Training Phase
Input: PI, ρ, λ , β, w_i, σ, γ.
Reconstruct c_n^a and c_n using Eqs. (7)-(10).
Optimize the ResotreNet with Eqs. (11)-(12).
Generate b_n via RestoreNet.
Optimize the RegisNet with Eq. (13).
Optimize the DenoiseNet with Eq. (15).

Inference Phase
Input: PI, γ, RestoreNet, RegisNet, DenoiseNet.
Reconstruct c_n using Eqs. (8)(10).
Generate b_n via RestoreNet.
Generate DVF using RegisNet and b_n.
Generate d_n using Eq. (14).
Generate DPRI using DenoiseNet($d_n + \gamma * \epsilon$).
Return: Final DPRI.

IV. EXPERIMENTS

A. Setup

1) Data Collection: There were three CBCT datasets in this work to evaluate different methods, including Monte-Carlo-based simulated dataset, LinaTech dataset, and the Elekta dataset.

The Monte-Carlo-based simulated dataset was provided by the SPARE Challenge [47]. It contained twelve patients and three of them had ground truth fully-sampled projection data for each breathing phase, and the other nine only offered down-sampled counterparts. The simulated dataset was acquired by the half-fan mode with a detector shift of 148 mm to increase the field of view. The scanned parameters were configured as follows. The distances from the X-ray source to the object and detector were 1000 mm and 1500 mm, respectively. The detector size was 768×384 and each of them covered an area of 0.776 mm^2. For one circle, 680 projections were collected. The reconstructed volume was 512×512×96 and each voxel represented the size of 0.9×0.9×1.5 mm^3. Notably, the ground truth image for individual phases was reconstructed from 680 projections. The breathing circles of this dataset are from 16 to 22.

The LinaTech dataset was acquired on the VenusX linear accelerator (LinaTech LLC, Suzhou, China) with the full-fan scanning mode. It consisted of eighteen patients collected from Pingjiang Hospital and Dengzhou Hospital, respectively. The geometry parameters were set as below. The source-to-isocenter and source-to-detector were 797 mm and 1354 mm, respectively. The detector had a size of 1408×1408 and each element stood for 0.3075×0.3075 mm^2. 360 projections were scanned via 360°. The size of the reconstructed image was 512×512×96 and everyone was 0.5×0.5×2 mm^3. And the breathing circles of different patients vary from 12 to 23.

The Elekta dataset was also given by the SPARE Challenge [47]. It has five patients with a total of 20 scans acquired with

full-fan mode. The distances from the X-ray to the rotation center and detector were 1000 mm and 1536 mm, respectively. The detector size was 512×512, each with an area of 0.8×0.8 mm^2. The reconstructed volume was 256×256×200 from 340 projections with a rotation of 200° and each voxel stood for 1×1×1 mm^3. Its breathing circles are various from 40 to 65 with a slow scanning speed.

The respiratory phases of the simulated and Elekta datasets were provided by the datasets themselves, and the breathing bins of the LinaTech dataset were extracted using [31]. Referred to most 4D CBCT-related works [28, 30, 31, 34], the respiratory circle [0%, 100%] for all datasets was divided into ten subphases (Phase 1, Phase 2, etc.) using the phase gating technique [48]. For example, Phase 1 in this work really corresponded to the phase subset [0%, 10%] and so on for the other phases.

2) Network Configuration: For the RestoreNet training, the hyper-parameter ρ in Eq. (11) was set to 10 suggested by [36, 39]. According to extensive experiments, the hyper-parameters λ and β in Eqs. (12) were selected to 0.01 and 10. Referred to [37], the w_i in Eq. (12) were respectively set to $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, and 1. The learning rate was linearly decreased from $1e^{-3}$ to $1e^{-5}$ with 50 epochs. Furthermore, the patch-based training strategy was adopted. The batchsize was 32 and each patch had a size of 128×128, which was extracted from the volumetric images. At the inference stage, RestoreNet directly took the 2D slice as input and then restored the original 3D shape.

For RegisNet optimization, the weight parameter σ in Eq. (13) was 4 and the learning rate was initially set to $1e^{-4}$ and slowly reduced to $1e^{-5}$ during 50 epochs. Both the RestoreNet and RegisNet were optimized by Adam with the setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Specifically, the RegisNet regarded the 3D image as input to generate the DVFs and the batchsize was 1.

For DenoiseNet, the noise level γ was 7×10^{-4}, and the same optimization settings as RestoreNet were applied to DenoiseNet.

Simulated and clinical datasets adopted the same hyper-parameter settings.

3) Comparison Methods & Evaluation Criteria: To validate the proposed DPI-MoCo, the respiratory phase gated FDK (Gated-FDK) [49], PICCS [13], MoCo [50], DDNet [51], CycN-Net [30], PRIOR-Net [28], and Boosting [33] were chosen as comparisons. For the simulated dataset, two patients with ground truth images were selected to train the DDNet, CycN-Net, and PRIOR-Net. The last labeled patient was performed to assess different methods. Because the Boosting and our DPI-MoCo were not dependent on the reference 4D CBCT images, the other patients were additionally employed to optimize these two algorithms.

For LinaTech and Elekta datasets, the well-trained DDNet and CycN-Net models obtained on the simulated 4D CT dataset as described in [28] were straightly applied to process corrupted PRIs. Yet the dataset-specific methods, including Boosting and the proposed DPI-MoCo, were retrained for better adaption of the new dataset. Particularly, for LinaTech dataset, twelve patients were selected for training, three patients were selected for validation, and the rest three were selected for testing. For Elekta dataset, four patients were selected for training and one patient was selected for testing. The total training time of DPI-MoCo for these three datasets was 106.9, 117.2, and 110.6 hours, respectively.

All the methods were performed on a PC (CPU was Intel(R) Core(TM) i9-10900K, 3.7 GHz, GPU was NVIDIA RTX 3090 with 24G memory).

Besides, the root mean square error (RMSE), peak signal-noise-ratio (PSNR), and structural similarity index (SSIM) were chosen for quantitative evaluation.

B. Simulated Data Results

Table I gives the quantitative evaluations of different methods in average and five selected phases. It can be seen that the FDK algorithm is greatly affected owing to the extremely

Table I
QUANTITATIVE EVALUATIONS OF DIFFERENT METHODS FOR THE SIMULATED DATASET (RMSE: mm^{-1}). BLUE AND GREEN RESPECTIVELY REPRESENT METHODS TRAINED WITH AND WITHOUT LABEL DATA.

	Metric	Gated-FDK	DDNet	CycN-Net	PRIOR-Net	Boosting	DPI-MoCo
Average	RMSE(10^{-3})	3.37	1.65	1.40	1.18	1.39	**0.76**
	PSNR	15.47	21.76	23.11	24.61	23.17	**28.45**
	SSIM	0.5400	0.8291	0.9208	**0.9347**	0.8569	0.9054
Phase 1	RMSE(10^{-3})	2.82	1.47	1.42	1.21	1.34	**0.79**
	PSNR	16.95	22.65	22.98	24.43	23.43	**28.20**
	SSIM	0.5747	0.8488	0.9218	**0.9347**	0.8650	0.9046
Phase 3	RMSE(10^{-3})	3.19	1.53	1.38	1.15	1.35	**0.72**
	PSNR	15.87	22.4	23.19	24.75	23.38	**28.97**
	SSIM	0.5597	0.8435	0.9237	**0.9382**	0.8611	0.9111
Phase 5	RMSE(10^{-3})	3.27	1.60	1.39	1.18	1.37	**0.75**
	PSNR	15.68	21.99	23.14	24.58	23.26	**28.62**
	SSIM	0.5597	0.8378	0.9202	**0.9336**	0.8602	0.9076
Phase 7	RMSE(10^{-3})	3.79	1.76	1.38	1.16	1.44	**0.76**
	PSNR	14.37	21.12	23.18	24.71	22.85	**28.23**
	SSIM	0.5167	0.8142	0.9208	**0.9346**	0.8504	0.9014
Phase 9	RMSE(10^{-3})	3.77	1.93	1.40	1.18	1.41	**0.79**
	PSNR	14.43	20.26	23.08	24.58	23.01	**28.07**
	SSIM	0.5005	0.7974	0.9189	**0.9328**	0.8509	0.8999

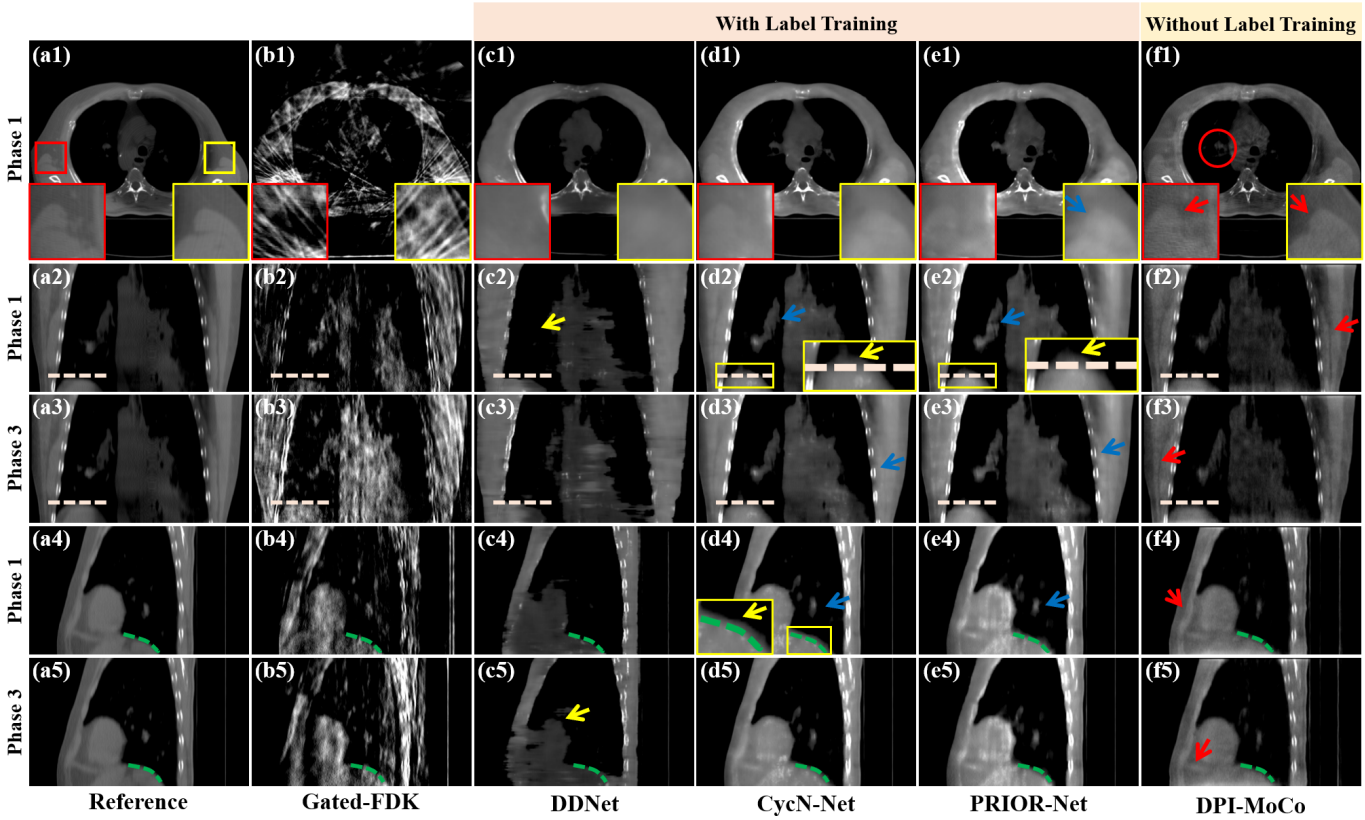

Fig. 3. Reconstructed results from selected Phases 1 and 3 on the simulated dataset for different methods, including reconstructed images and magnified regions-of-interest (ROIs). Reference images are reconstructed from full-sampled projections. (a1)-(f1) Axial results at Phase 1, (a2)-(f2) Coronal results at Phase 1, (a3)-(f3) Coronal results at Phase 3, (a4)-(f4) Sagittal results at Phase 1, (a5)-(f5) Sagittal results at Phase 3. The display window is [0.004, 0.018] mm^{-1}.

sparse-view projection data. For all DL-based methods, they significantly boost the assessments aided by the powerful feature extraction capacity. In this study, DDNet is a pure CNN model that treats the 4D CBCT reconstruction as sparse-view imaging but ignores some intrinsic properties of 4D CBCT, such as temporal correlations and prior image guidance. Consequently, it works worse than CycN-Net and PRIOP-Net in all cases. Again, benefitting from higher prior image utilization and more advanced techniques [28], PRIOR-Net brings improved scores than CycN-Net. Compared to the above three models that need labeled data for training, Boosting and our DPI-MoCo, which are built on the pseudo-label data, demonstrate competitive performance. Specifically, both being plain CNNs, Booting achieves better evaluations than DDNet. Moreover, our proposed DPI-MoCo leads to the best measurements in RMSE and PSNR. It should be noted that DDNet, CycN-Net, and PRIOR-Net use two patients to train the model, yet Boosting and DPI-MoCo are conducted with all unlabeled patients. This configuration is reasonable because it is nearly impossible to collect high-quality PRIs but can easily acquire a large amount of pseudo-labeled data in practical applications. So, as proven in Table I, Boosting, and DPI-MoCo provide a reliable alternative way for 4D CBCT reconstruction.

Fig. 3 illustrates the reconstructed results of different methods on the simulated 4D CBCT datasets at Phases 1 and 3. With an average of 68 views for reconstructing each PRI, the FDK algorithm suffers from severe streaking artifacts, making the normal tissues and organs indistinguishable. However, due to the high temporal resolution, it can approximately respond to the breathing motion in some regions (as marked by pink and green dash lines in Fig. 3(b2)-(b5)). Although DDNet shows advantages in artifact reduction over FDK (as observed in Fig. 3(c1)-(c5)), it oversmooths the results and misses some structures (as pointed out by yellow arrows in Fig. 3(c2)(c5)). Following the prior image guidance, CycN-Net and PRIOR-Net not only alleviate the excessive blurring phenomenon that occurred in DDNet but also successfully recover clearer bone structures and minor tissues (as suggested by blue arrows in Fig. 3(d2)-(e4)) while the utilization of prior images also negatively delivers a few motion-induced artifacts (as seen by yellow arrows in Fig. 3(d2)-(e2)(d4)). Last, our DPI-MoCo framework depicts promising results that can generate accurate tissue features and muscle edges (as marked by red circles and arrows in Fig. 3(f1)-(f5)) as well as hold the relatively precise motion pattern similar to FDK (as stated by pink and green dash lines in Fig. 3(f2)-(f5)).

C. Clinical Data Results

Fig. 4 exhibits the reconstructed images of different methods on the LinaTech dataset at Phases 1, 2, and 3. Apart from the streak artifacts, the down-sampled projections also produce noise to the reconstructions as noticed in Fig. 4(a1)-(a3). By minimizing the gradient maps between the restored results and prior images, PICCS can suppress most artifacts and preserve some bone features without losing dynamic information compared to FDK reconstructions (as shown by the blue arrows in Fig. 4(b1)(b3)). Because the ground truth 4D CBCT images

are inaccessible, the pre-trained DDNet and CyCN-Net on the simulated dataset have to be directly transferred to clinical counterparts. Although DDNet and CycN-Net provide more bony structures (as indicated by blue arrows in Fig. 4(c3)-(d3)), neither of them generates reliable results due to the big domain gaps (as stated by green arrows in Fig. (c1)-(d3)). Compared to DDNet and CycN-Net, Boosting is independent of labeled data, hence, bringing more guaranteed results. Regarding the PICCS as a baseline, Boosting overcomes its blocky effects but results in inaccurate movements (as indicated by the green arrow in Fig. 4(e3)). As presented by zoomed regions-of-interest (ROIs) in Fig. 4(f2)-(f3), the proposed DPI-MoCo gives impressive results, specifically in muscle recovery. Unlike Boosting, DPI-MoCo can also reflect temporal changes in lung regions.

Fig. 5 illustrates the reconstructed results of different methods on the Elekta dataset at Phase 5. Again, the pre-trained models, including DDNet and CycN-Net, suffer from severe fake artifacts caused by domain gaps. Assisted by the prior image, Boosting can restore more reliable images with clearer bones (as shown by blue arrows in Fig. 5(d1)(d3)). Moreover, the proposed method performs best in detail preservation (as observed by red arrows in Fig. 5(e1)-(e3)).

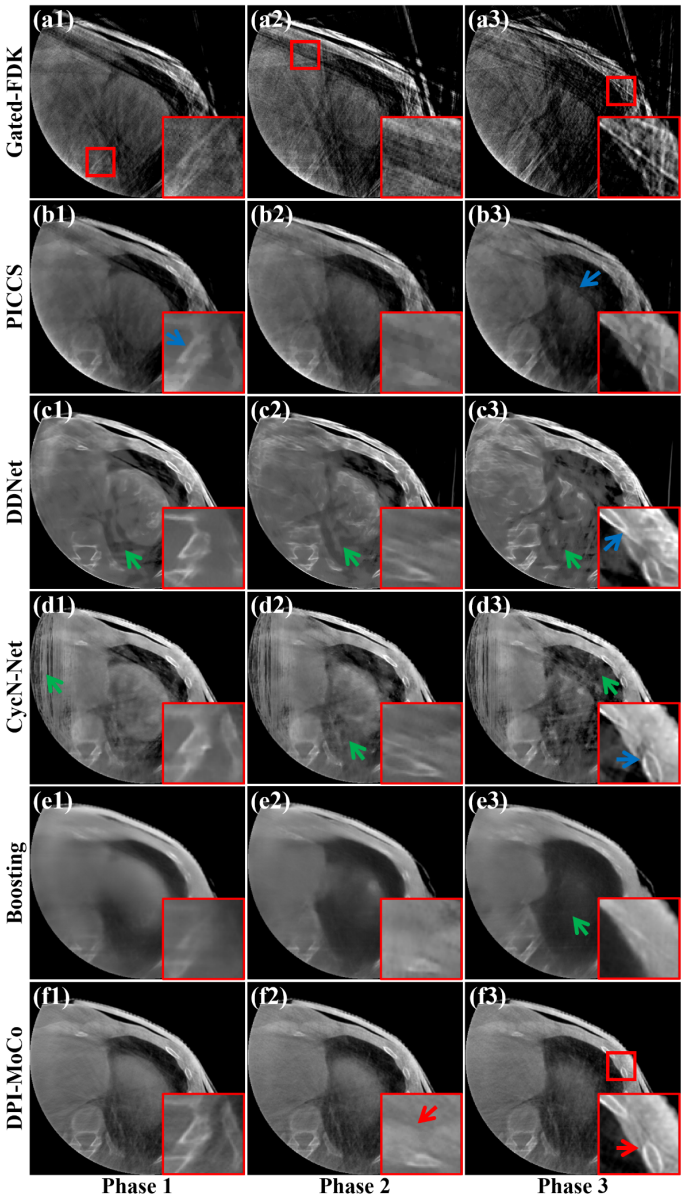

Fig. 4. Reconstructed axial results from three selected phases on the LinaTech dataset for different methods. (a1)-(f1) Reconstructed results at Phase 1, (a2)-(f2) Reconstructed results at Phase 2, (a3)-(f3) Reconstructed results at Phase 3. The display window is [0.004, 0.022] mm^{-1}.

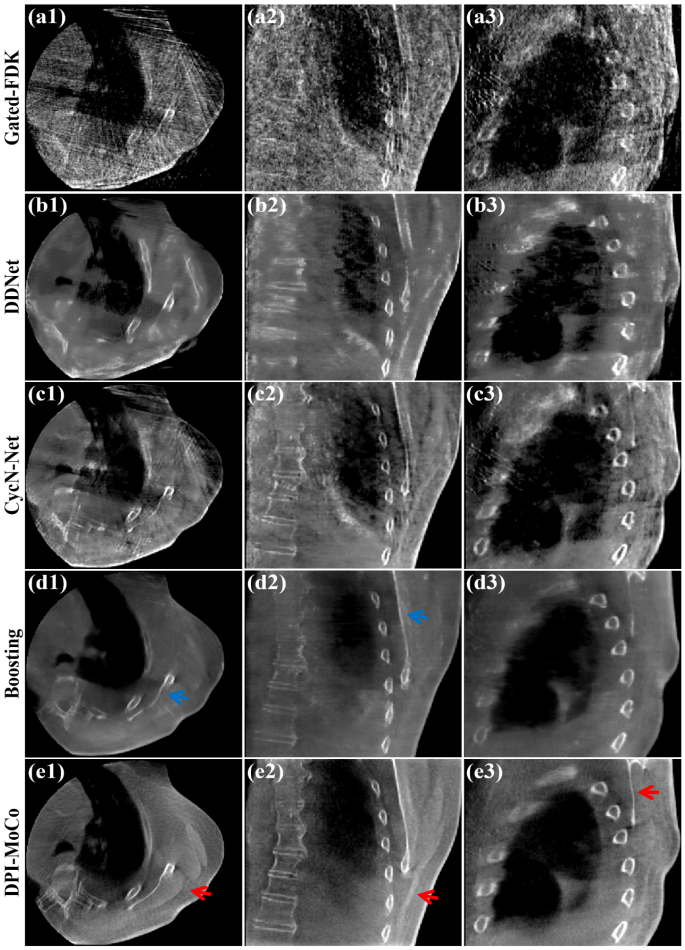

Fig. 5. Reconstructed results on the Elekta dataset for different methods at Phase 5. (a1)-(a3) Images reconstructed by Gated-FDK, (b1)-(b3) Images processed by DDNet, (c1)-(c3) Images processed by CycN-Net, (d1)-(d3) Images processed by Boosting, (e1)-(e3) Images processed by DPI-MoCo. The display window is [0.006, 0.018] mm^{-1}.

D. Analyses of Different Backbones for RestoreNet and RegisNet

To investigate the effects of different backbones for RestoreNet and RegisNet on the proposed DPI-MoCo, two additionally representative architectures, including Transformer, and Mamba (detailed configurations are listed in Table II), were performed on the simulated and Elekta datasets. Specifically, all the ResotreNets were optimized with MSE loss, and RegisNets were trained using Eq. (13) with the same hyper-parameter settings. All the results in this section were not processed by DenoiseNet.

Table III shows the average evaluations over ten phases. Overall, CNN, Transformer, and Mamba share similar quantitative assessments. Further, as observed in Figs. 6-7, although Transformer and Mamba bring more accurate bony structures (as point by red arrows in Fig. 6(c1)-(d1)), these three

backbones have resembled performance. Even though Transformer and Mamba have demonstrated remarkable performance in various medical imaging tasks, by comparing the results in Table III and Table V, it can be concluded the loss function has a greater influence than network architectures in our proposed method.

Table II
THE CONFIGURATIONS OF DIFFERENT BACKBONES FOR RESTORENET AND REGISNET.

	CNN	Transformer	Mamba
RestoreNet	Fig. 2	Restormer [52]	MambaIR [53]
RegisNet	VoxelMorph [43]	TransMorph [54]	MambaMorph [55]

Table III
AVERAGE QUANTITATIVE EVALUATIONS OF DIFFERENT BACKONES FOR RESTORENET AND REGISNET ON THE SIMULATED DATASET (RMSE: mm^{-1}).

Metric	CNN	Transformer	Mamba
RMSE(10^{-4})	8.13	8.12	**8.03**
PSNR	27.84	27.85	**27.97**
SSIM	**0.9001**	0.8976	0.8987

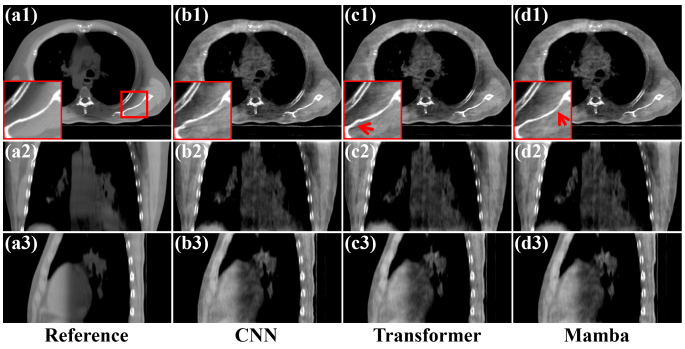

Fig. 6. Reconstructed results from Phase 2 on the simulated dataset. (a1)-(a3) Reference images reconstructed from full-sampled projections, (b1)-(b3) Images processed by CNN networks, (c1)-(c3) Images processed by Transformer networks, (d1)-(d3) Images processed by Mamba networks. The display window is [0.004, 0.014] mm^{-1}.

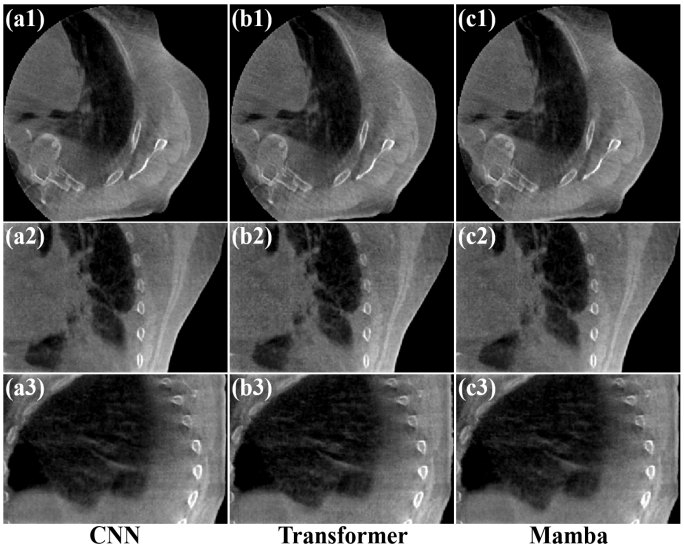

Fig. 7. Reconstructed results from Phase 2 on the Elekta dataset. (a1)-(a3) Images processed by CNN entworks, (b1)-(b3) Images processed by Transformer networks, (c1)-(c3) Images processed by Mamba networks. The display window is [0.004, 0.018] mm^{-1}.

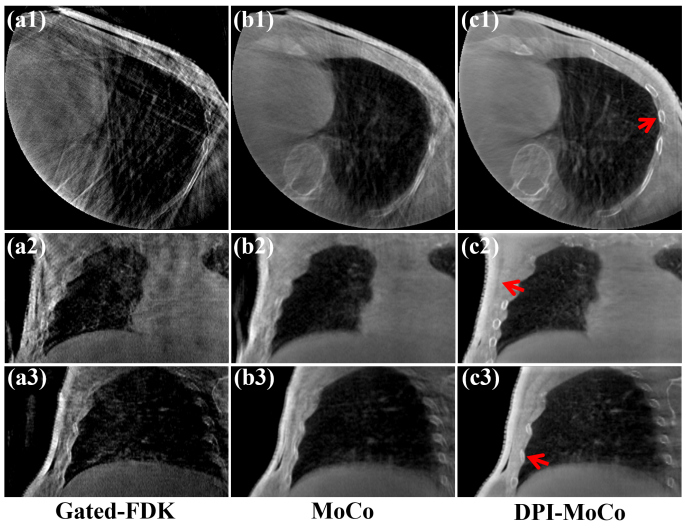

Fig. 8. Reconstruction images of different methods on the LinaTech dataset at Phase 1. (a1)-(a3) Images reconstructed by FDK, (b1)-(b2) Images reconstructed by MoCo, (c1)-(c3) Images processed by DPI-MoCo. The display window is [0.004, 0.022] mm^{-1}.

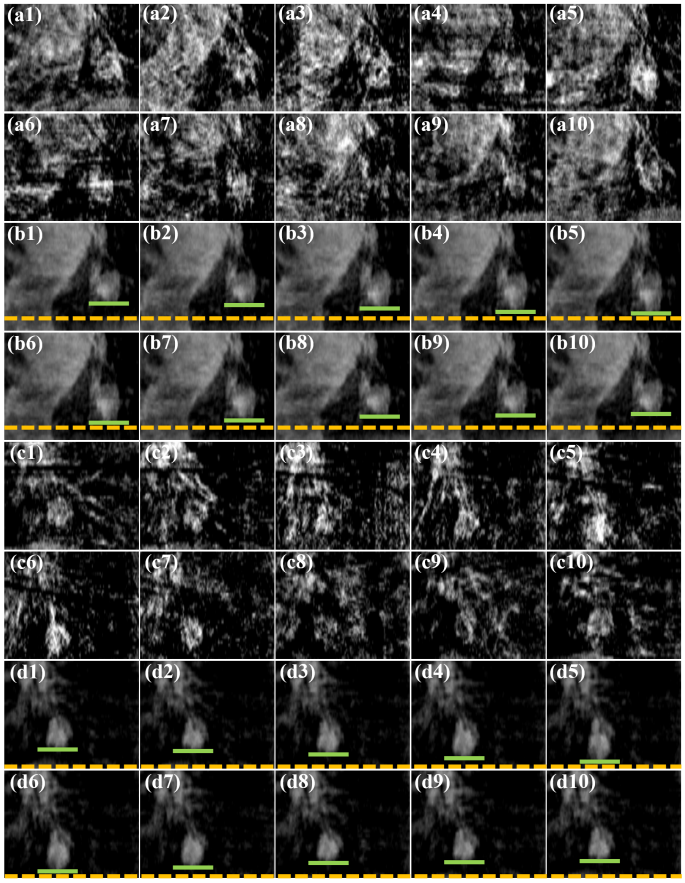

Fig. 9. Reconstructed tumor ROI results of different methods on the LinaTech dataset with continuous ten phases. (a1)-(a10) Coronal views of Gated-FDK from Phase 1 to Phase 10, (b1)-(b10) Coronal views of DPI-MoCo from Phase 1 to Phase 10, (c1)-(c10) Sagittal views of Gated-FDK from Phase 1 to Phase 10, (d1)-(d10) Sagittal views of DPI-MoCo from Phase 1 to Phase 10. The display window is [0.006, 0.2] mm^{-1}.

E. Comparison of DPI-MoCo and Conventional MoCo

Fig. 8 depicts the reconstructed results of different methods on the LinaTech dataset at Phase 1. Although the DVFs are extracted from artifact-induced PRIs, the conventional MoCo

generates higher-quality images. This claims that MoCo is more robust to dynamic imaging. With better initial images, our proposed method further boosts the bony or tissue features (as proved by the red arrows in Fig. 8(c1)-(c3)).

F. Tumor Reconstruction and Localization

The main purpose of 4D CBCT is to track the tumor movements. Therefore, the image quality of reconstructed tumors is an important criterion for evaluating 4D CBCT imaging methods. Fig. 9 illustrates the reconstructed ROI images with tumors on the LinaTech dataset with continuous ten phases. As observed in Fig. 9, it is hard for FDK to recognize lesions caused by severe artifacts. Noticeably, our DPI-MoCo greatly improves the tumor with high distinctiveness and provides obvious respiratory movements (as shown by green lines in Fig. 9).

G. Ablation Study

In this section, an ablation study was performed to probe the effectiveness of different important components used in DPI-MoCo.

Table IV
AVERAGE QUANTITATIVE EVALUATIONS OF PIGAE MODULE ON THE SIMULATED DATASET (RMSE: mm^{-1}).

Metric	w/o PIGAE	w/ PIGAE
RMSE(10^{-3})	1.35	**1.13**
PSNR	23.41	**24.99**
SSIM	0.8527	**0.8812**

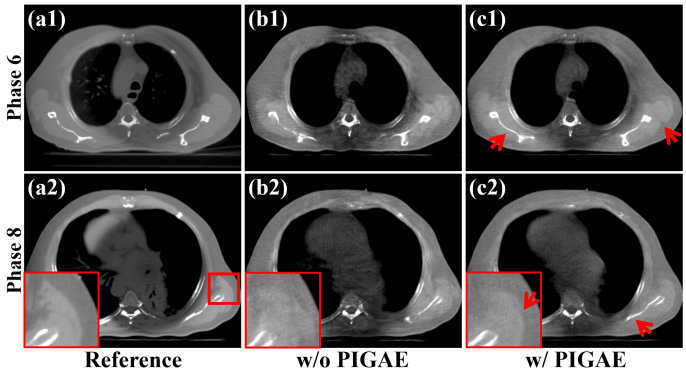

Fig. 10. Reconstructed results from Phases 6 and 8 on the simulated dataset. (a1)-(a2) Reference images reconstructed from full-sampled projections, (b1)-(b2) Images processed by RestoreNet without PIGAE module, (c1)-(c2) Images processed by RestoreNet with PIGAE module. The display window is [0.004, 0.014] mm^{-1}.

a) PIGAE Module: This part concentrates on exploring the PIGAE module with the simulated dataset. The RestoreNet trained on the pseudo-paired dataset ($\{x_n^a, PI\}$) was taken as the comparison model. The RestoreNets with or without PIGAE are trained using Eqs. (11)-(12) with the same hyper-parameter settings.

Table IV gives the average quantitative evaluations for ten phases. Assisted by the PIGAE module, BPRIs were improved greatly than without PIGAE. Moreover, it can be seen that in Fig. 10(b1)-(b2), the RestoreNet without PIGAE can remove most artifacts but lose some small details. Contrary to this, with the assistance of PIGAE, RestoreNet works well in bone and muscle preservation (as pointed out by red arrows in Fig. 10(c1)-(c2)), suggesting that the PIGAE indeed boosts the image quality in stationary areas.

b) Generative Adversarial Network and Constrative Learning: In this section, a progressive verification strategy was adopted to analyze the GAN and CL used in RestoreNet. The RestoreNet optimized with MSE was treated as the baseline model. Next, the GAN was added to the loss function to build the first comparison model. Last, CL cooperated with MSE and GAN jointly to establish the second comparison model.

Fig. 11 illustrates the reconstruction results of different loss functions on the simulated dataset. Although MSE can remove most streak artifacts, it inevitably causes blurring textures. In contrast, GAN facilitates tissue preservation (as pointed by the yellow arrows in Fig. 11(c1)-(c3)), which was identical to the observations in [35, 36]. Furthermore, after additionally adding the CL, all the results are promoted, including clearer bony structures and sharper muscle edges (as seen by yellow and red arrows in Fig. 11(d1)-(d3)).

Further, Fig. 12 explores the effects of different loss functions on the LinaTech dataset. Again, GAN enables clearer bony features (as validated by yellow arrows in Fig. 12(b1)-(b2)) and CL brings more accurate details (as indicated by red arrows in Fig. 12(c1)-(c2)).

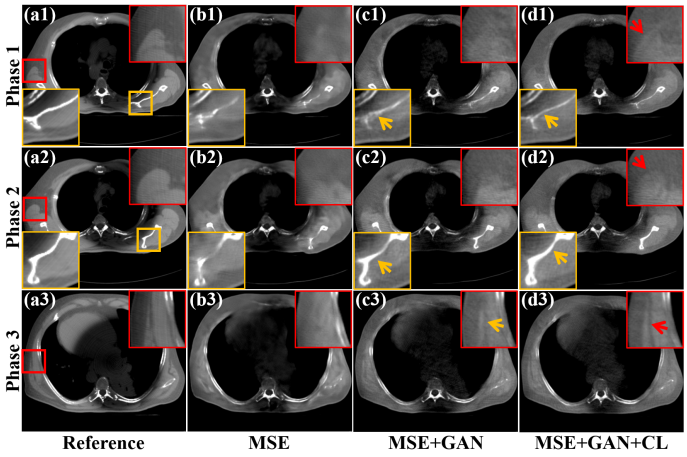

Fig. 11. Reconstructed results from Phases 1, 2, and 3 on the simulated dataset. (a1)-(a3) Reference images reconstructed from full-sampled projections, (b1)-(b3) Images optimized by MSE loss, (c1)-(c3) Images optimized by MSE+GAN loss, (d1)-(d3) Images optimized by MSE+GAN+CL loss. The display window is [0.005, 0.014] mm^{-1}.

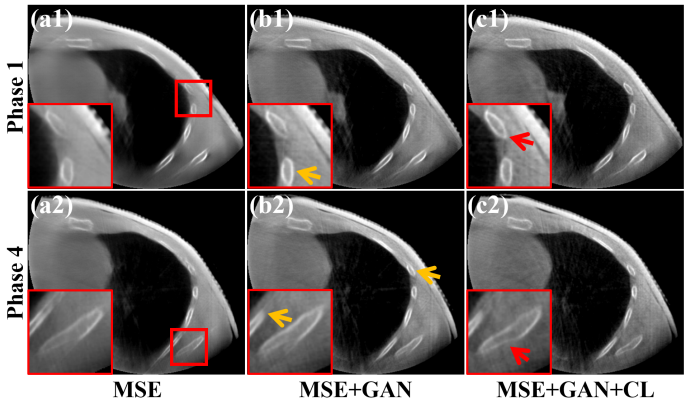

Fig. 12. Reconstructed results from Phases 1 and 4 on the LinaTech dataset. (a1)-(a2) Images optimized by MSE loss, (b1)-(c2) Images optimized by MSE+GAN loss, (c1)-(c2) Images optimized by MSE+GAN+CL loss. The display window is [0.006, 0.02] mm^{-1}.

c) MoCo Reconstruction: To examine the impact of MoCo used in the proposed framework, the images generated by RestoreNet and the results of d_n (defined in Eq. (14)) were compared on the simulated dataset.

Table V
AVERAGE QUANTITATIVE EVALUATIONS BETWEEN RESTORENET AND d_n ON THE SIMULATED DATASET (RMSE: MM^{-1}).

Metric	RestoreNet	d_n
RMSE(10^{-3})	1.13	**0.77**
PSNR	24.99	**28.36**
SSIM	0.8812	**0.9086**

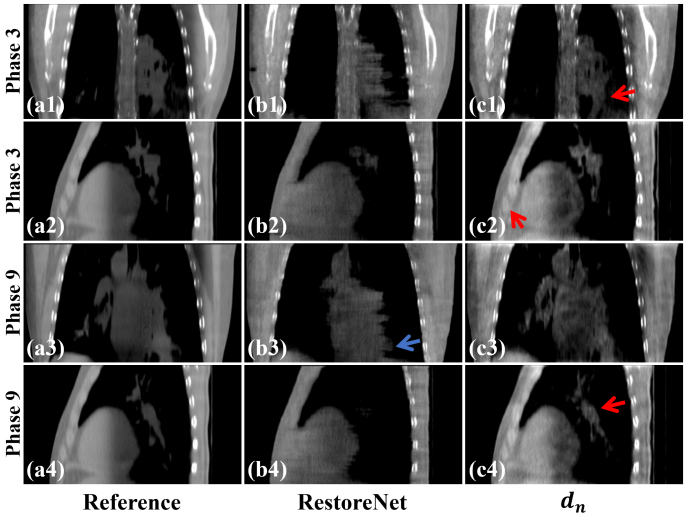

Fig. 13. Reconstructed results from Phases 3 and 9 on the simulated dataset. (a1)-(a4) Reference images reconstructed from full-sampled projections, (b1)-(b4) Images processed by RestoreNet, (c1)-(c4) Images provided by d_n. The display window is [0.004, 0.014] mm^{-1}.

The quantitative assessments are shown in Table V, which suggests that the MoCo operation can utilize more information contained in all respiratory phases to promote the 4D CBCT reconstructions. Besides, Fig. 13 demonstrates the selected coronal and sagittal views reconstructed from RestoreNet and DPI-MoCo. Compared to the original pseudo-average dataset version in [33], i.e., $\{x_n^a, PI\}$, we construct a higher quality dataset $\{c_n^a, PI\}$ as well as adopt effective loss functions, it still leads to tissue missing (as shown by the blue arrow in Fig. 13(b3)). Due to the local smoothness property of DVFs [34], the reconstructed image b_n^a can extract relatively accurate DVFs to compensate for the breathing motions. As a result, MoCo brings visually improved images with more tissues and clear edges (as marked by red arrows in Fig. 13(c1)-(c4)), which strongly states the necessity of MoCo for 4D CBCT.

d) Deformed Object Selection for MoCo: As mentioned above, there are three choices to be deformed to get the final results, including x_n, c_n, and b_n. To research the influence of different candidates, Fig. 14 provides the reconstructed images and zoomed ROIs on the LinaTech dataset at Phase 1. As expected, $\sum_{n=1}^{N} DVF_n^1(x_n)$ leads to high-contrast results but is accompanied by a few artifacts. On the contrary, $\sum_{n=1}^{N} DVF_n^1(b_n)$ works best in artifact reduction while performs worst in tissue sharpness. Last, $\sum_{n=1}^{N} DVF_n^1(c_n)$, adopted in our DPI-MoCo, gives a tradeoff between artifact suppression and detail protection (as validated by red arrows in Fig. 14(c2)).

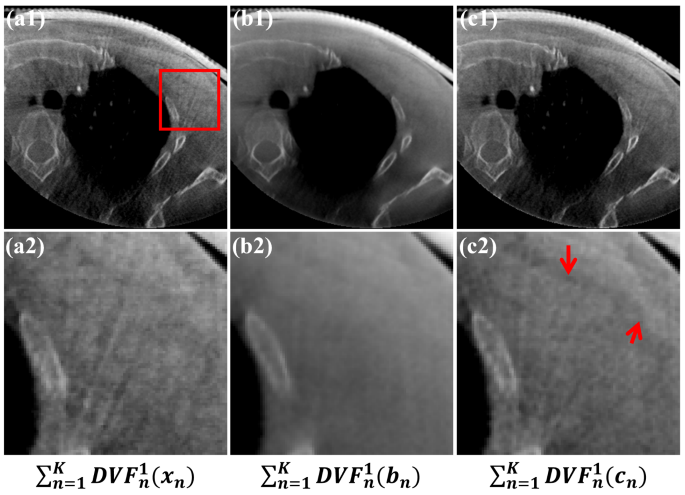

Fig. 14. Reconstruction results and zoomed ROIs on the LinaTech dataset at Phase 1, which are deformed from x_n, c_n, and b_n, respectively. The display window is [0.008, 0.02] mm^{-1}.

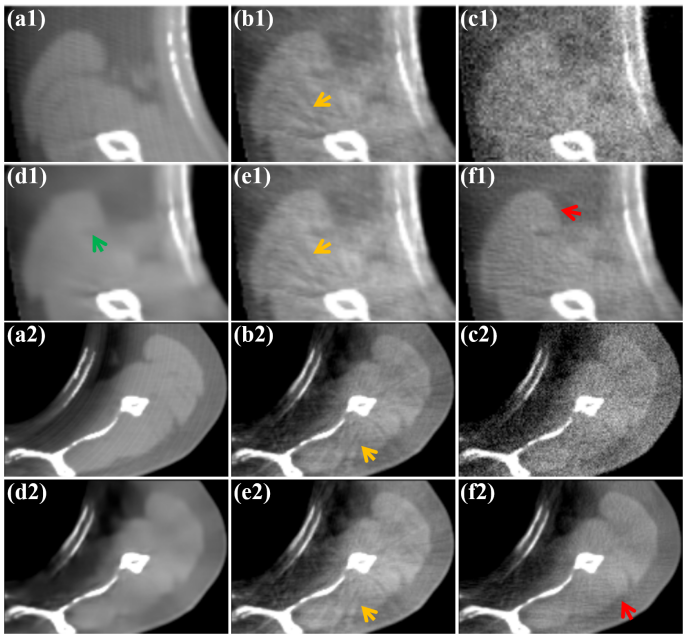

Fig. 15. Denoised ROI results on the simulated dataset of different methods at Phase 1. (a1)-(a2) Reference images reconstructed from full-sampled projections, (b1)-(b2) d_1 images, (c1)-(c2) Noisy d_1 images, (d1)-(d2) Denoised images processed by KSVD from (b1)-(b2), (e1)-(e2) Denoised images processed by Nb2Nb from (b1)-(b2), (f1)-(f2) Denoised images processed by DenoiseNet from (c1)-(c2). The display window is [0.005, 0.014] mm^{-1}.

e) DenoiseNet: To validate the prior-image guided DenoiseNet defined in Eq. (15), the KSVD [44] and Neighbor2Neighbor (Nb2Nb) [45] were selected as the comparisons, which were the representative denoising methods for compressed sensing and self-supervised learning. Fig. 15 depicts the denoised ROI results of different methods on the simulated dataset. Deformed PRI d_1 generates artifacts around bones (as shown by yellow arrows in Fig. 15(b1)-(b2)). Even removing the most of artifacts, KSVD unavoidably leads to blurring results. Meanwhile, Nb2Nb also fails to boost d_1 (as observed in Fig. 15(e1)-(e2)), which could be because Nb2Nb cannot identify the slight artifact as the noise. After adding the Gaussian noise to the d_1, the corresponding artifact also is

submerged in the noise (as indicated in Fig. 15(c1)-(c2)). When denoising the noisy d_1, the well-trained DenoiseNet not only can restore the image details but also remove the artifacts in d_1 at the same time (as proven by red arrows in Fig. 15(f1)-(f2)).

H. Computational Cost

Table VI lists the total computational cost of different methods over ten phases on one patient from the LinaTech dataset. Due to the complex forward and backward operations, PICCS needs more time than postprocessing methods, i.e., CycN-Net and Boosting. Similarly, DPI-MoCo also spends a longer time because of multi-stage processing. Generally, the proposed method could provide pretty good results within an acceptable time.

Table VI
COMPUTATIONAL COST OF DIFFERENT METHODS ON THE LINATECH DATASET (UNIT: SECOND).

Method	PICCS	CycN-Net	Boosting	DPI-MoCo
Time	1369	18	17	91

I. Hyper-Parameter Analyses

Hyper-parameters λ and β (as defined in Eq. (12)) play an important role in controlling the weights of GAN and CL. To explore the optimal settings for RestoreNet, Table VII lists the average PSNR assessment over ten phases of various hyper-parameter settings on the simulated dataset. It can be seen that λ and β values greatly affect the performance of RestoreNet. In this work, according to the quantitative evaluations, λ and β are set to 0.01 and 10, respectively.

Table VII
AVERAGE PSNR EVALUATIONS OF DIFFERENT HYPER-PARAMETERS ON THE SIMULATED DATASET.

λ \ β	0.5	1	5	10	50	100
0.0005	23.93	24.23	24.87	24.84	24.70	24.76
0.001	24.54	24.75	24.92	24.10	24.52	24.71
0.005	24.63	24.64	24.07	24.85	24.53	24.17
0.01	24.47	24.59	24.88	**24.99**	24.68	24.89
0.05	24.16	24.15	24.69	24.29	23.73	24.31
0.1	24.76	24.12	24.70	24.52	23.88	24.15

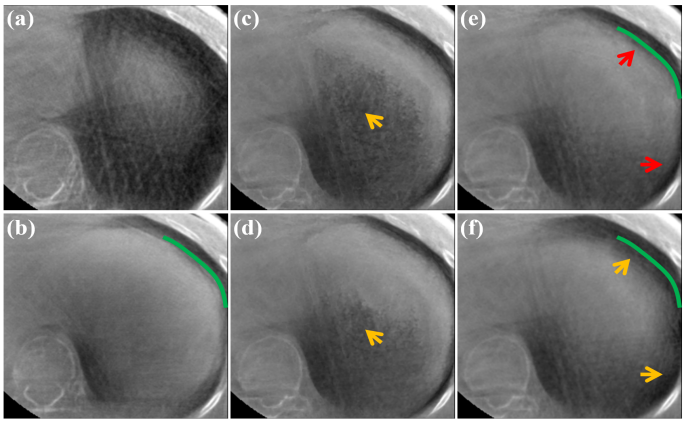

Fig. 16. Registration ROI results on the LinaTech dataset with different σ values. (a) Moving ROI image at Phase 5, (b) Fixed image at Phase 1, (c) Registrated image deformed from (a) with $\sigma = 0.1$, (d) Registrated image deformed from (a) with $\sigma = 1$, (e) Registrated image deformed from (a) with $\sigma = 4$, (f) Registrated image deformed from (a) with $\sigma = 10$. The display window is [0.004, 0.018] mm^{-1}.

Additionally, hyper-parameter σ is crucial for RegisNet optimization. It can adjust the smoothness of the DVF, which therefore influences the registration performance. Fig. 16 exhibits the registration ROI results with different σ values, where Fig. 16(a) and Fig. 16(b) are moving and fixed images, respectively. Smaller σ means less constraint on the DVF, which leads to some distortions as indicated by yellow arrows in Figs. 16(c)(d). In contrast, larger σ will produce over-smoothing DVF that cannot effectively deform the moving image to the fixed image (as observed by yellow arrows and the green line in Fig. 16(f)). Finally, σ is set to 4 to balance the tradeoff between DVF smoothness and registration performance (as suggested by red arrows and the green line in Fig. 16(e)).

V. CONCLUSION AND DISCUSSION

It remains a challenging problem to reconstruct high-quality clinical 4D CBCT images due to the extremely sparse-view projection data. To tackle this issue, traditional methods represented by PICCS and MoCo are proposed, and bring improvement in artifact reduction and detail recovery to a certain extent. Meanwhile, DL-based methods have attracted great attention in various medical imaging tasks and show superiority over conventional methods. However, two factors hamper the application of DL in real 4D CBCT. First, because the ground truth 4D CBCT images are inaccessible in real applications, some classical supervised learning methods, such as DDNet, and CycN-Net, have to be transferred from the simulated dataset to the clinical dataset. It is very difficult to make the simulated data similar to the real ones, which always results in domain gaps, so the performances of these methods are expected to decrease. Second, unlike image denoising, unsupervised learning or self-supervised learning have not been widely employed for sparse-view CBCT imaging, therefore, their applications in 4D CBCT are yet mature.

Latest, [33] offers a potential approach to overcome the above-mentioned dilemma of DL by constructing novel pseudo-average datasets. Inspired by it, this work proposes a DPI-MoCo framework to further promote PRIs. It consists of four steps, which sequentially are high-quality pseudo-average dataset construction, effective RestoreNet network training, motion estimation and compensation, and DenoiseNet optimization. Each of them can push the final results toward a more optimal solution. Particularly, the first two stages aim to provide coarse artifact-free results. Then, the next two steps concentrate on local structure fine-tuning and slight artifact removal. Unlike most existing deep models in 4D CBCT imaging, our DPI-MoCo breaks through the limitation of relying on ground truth data for training. Compared to conventional MoCo, the proposed method leads to more accurate anatomical features. Both simulated and clinical datasets validate the effectiveness of the proposed DPI-MoCo in concurrently reducing artifacts and maintaining the motion information.

Nevertheless, it still has two defects for DPI-MoCo that should be noticed. When making the pseudo-average datasets,

there is a hypothesis that x_n^a and x_n have similar streak artifacts, however, on account of the different projection sampling trajectories, x_n^a and x_n may contain slightly different degradation patterns in practical cases. Besides, the PI also has some motion artifacts in relatively stationary areas. These factors will result in a few artifacts left in the final images even though applying the DenoiseNet. At the same time, as indicated by CycN-Net and PRIOR-Net, although the utilization of prior images in the PIGAE module generates more effective pseudo-average datasets, it will negatively introduce some motion-blurred phenomena. Therefore, how to build a more reasonable pseudo-average dataset should be further explored. Very recently, the Neural Radiance Field (NeRF) has been applied to medical imaging with patient-specific mode, so the combination of NeRF and our MoCo can be considered in the future.

REFERENCE

[1] M. Oldham, D. Létourneau, L. Watt *et al.*, "Cone-beam-CT guided radiation therapy:: A model for on-line application," *Radiotherapy and Oncology*, vol. 75, no. 3, pp. 271-278, 2005.

[2] K. Srinivasan, M. Mohammadi, and J. Shepherd, "Applications of linac-mounted kilovoltage Cone-beam Computed Tomography in modern radiation therapy: A review," *Polish journal of radiology*, vol. 79, pp. 181-93, 2014, 2014.

[3] R. T. O'Brien, O. Dillon, B. Lau *et al.*, "The first-in-human implementation of adaptive 4D cone beam CT for lung cancer radiotherapy: 4DCBCT in less time with less dose," *Radiotherapy and Oncology*, vol. 161, pp. 29-34, 2021.

[4] K. K. Brock, "Adaptive Radiotherapy: Moving Into the Future," *Semin. Radiat. Oncol.*, vol. 29, no. 3, pp. 181-184, 2019.

[5] S. Leng, J. Zambelli, R. Tolakanahalli *et al.*, "Streaking Artifacts Reduction in Four-Dimensional Cone-Beam Computed Tomography," *Med. Phys.*, vol. 35, no. 6, pp. 2, 2008.

[6] S. H. Zhi, M. Kachelriess, and X. Q. Mou, "Spatiotemporal structure-aware dictionary learning-based 4D CBCT reconstruction," *Med. Phys.*, vol. 48, no. 10, pp. 6421-6436, 2021.

[7] T. Rodet, F. Noo, and M. Defrise, "The cone-beam algorithm of Feldkamp, Davis, and Kress preserves oblique line integrals," *Med. Phys.*, vol. 31, no. 7, pp. 1972-1975, 2004.

[8] J. Mascolo-Fortin, D. Matenine, L. Archambault *et al.*, "A fast 4D cone beam CT reconstruction method based on the OSC-TV algorithm," *Journal of X-Ray Science and Technology*, vol. 26, no. 2, pp. 189-208, 2018.

[9] R. Heylen, G. Schramm, P. Suetens *et al.*, "4D CBCT reconstruction with TV regularization on a dynamic software phantom," in IEEE Nuclear Science Symposium / Medical Imaging Conference (NSS/MIC), Manchester, ENGLAND, 2019.

[10] C. P. V. Christoffersen, D. Hansen, P. Poulsen *et al.*, "Registration-Based Reconstruction of Four-Dimensional Cone Beam Computed Tomography," *Ieee Transactions on Medical Imaging*, vol. 32, no. 11, pp. 2064-2077, 2013.

[11] D. C. Hansen, and T. S. Sorensen, "Fast 4D cone-beam CT from 60 s acquisitions," *Physics & Imaging in Radiation Oncology*, vol. 5, pp. 69-75, 2018.

[12] C. Mory, V. Auvray, B. Zhang *et al.*, "Cardiac C-arm computed tomography using a 3D+time ROI reconstruction method with spatial and temporal regularization," *Med. Phys.*, vol. 41, no. 2, 2014.

[13] Z. H. Qi, and G. H. Chen, "Extraction of tumor motion trajectories using PICCS-4DCBCT: A validation study," *Med. Phys.*, vol. 38, no. 10, pp. 5530-5538, 2011.

[14] G. H. Chen, J. Tang, and S. H. Leng, "Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Med. Phys.*, vol. 35, no. 2, pp. 660-663, 2008.

[15] S. H. Zhi, M. Kachelriess, and X. Q. Mou, "High-quality initial image-guided 4D CBCT reconstruction," *Med. Phys.*, vol. 47, no. 5, pp. 2099-2115, 2020.

[16] M. Brehm, P. Paysan, M. Oelhafen *et al.*, "Artifact-resistant motion estimation with a patient-specific artifact model for motion-compensated cone-beam CT," *Med. Phys.*, vol. 40, no. 10, 2013.

[17] M. Q. Chen, K. L. Cao, Y. F. Zheng *et al.*, "Motion-Compensated Mega-Voltage Cone Beam CT Using the Deformation Derived Directly From 2D Projection Images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1365-1375, 2013.

[18] T. Li, E. Schreibmann, Y. Yang *et al.*, "Motion correction for improved target localization with on-board cone-beam computed tomography," *Physics in Medicine and Biology*, vol. 51, no. 2, pp. 253-267, 2006.

[19] S. Rit, J. Wolthaus, M. van Herk *et al.*, "On-the-Fly Motion-Compensated Cone-Beam CT Using an a Priori Motion Model," *Lecture Notes in Computer Science*. pp. 729-736, 2008.

[20] S. Rit, J. Nijkamp, M. van Herk *et al.*, "Comparative study of respiratory motion correction techniques in cone-beam computed tomography," *Radiotherapy and Oncology*, vol. 100, no. 3, pp. 356-359, 2011.

[21] J. Wang, X. Gu, and T. Solberg, "High Quality 4-dimensional Cone Beam CT by Deforming Prior Images," *International Journal of Radiation Oncology Biology Physics*, vol. 84, no. 3, pp. S739-S739, 2012.

[22] J. Wang, and X. Gu, "Simultaneous Motion Estimation and Image Reconstruction (SMEIR) for 4D Cone-Beam CT," *Med. Phys.*, vol. 40, no. 6, 2013.

[23] J. L. Liu, X. Zhang, X. Q. Zhang *et al.*, "5D respiratory motion model based image reconstruction algorithm for 4D cone-beam computed tomography," *Inverse Problems*, vol. 31, no. 11, 2015.

[24] L. Y. Chao, Z. W. Wang, H. B. Zhang *et al.*, "Sparse-view cone beam CT reconstruction using dual CNNs in projection domain and image domain," *Neurocomputing*, vol. 493, pp. 536-547, Jul, 2022.

[25] S. Kim, J. Ahn, B. Kim *et al.*, "Convolutional neural network–based metal and streak artifacts reduction in dental CT images with sparse-view sampling scheme," *Med. Phys.*, vol. 49, no. 9, pp. 6253-6277, 2022.

[26] Z. R. Jiang, Y. X. Chen, Y. W. Zhang *et al.*, "Augmentation of CBCT Reconstructed From Under-Sampled Projections Using Deep Learning," *IEEE Transactions on Medical Imaging*, vol. 38, no. 11, pp. 2705-2715, 2019.

[27] A. Lahiri, G. Maliakal, M. L. Klasky *et al.*, "Sparse-View Cone Beam CT Reconstruction Using Data-Consistent Supervised and Adversarial Learning From Scarce Training Data," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 13-28, 2023.

[28] D. L. Hu, Y. K. Zhang, J. Liu *et al.*, "PRIOR: Prior-Regularized Iterative Optimization Reconstruction For 4D CBCT," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5551-5562, 2022.

[29] Z. R. Jiang, Z. Y. Zhang, Y. S. Chang *et al.*, "Enhancement of 4-D Cone-Beam Computed Tomography (4D-CBCT) Using a Dual-Encoder Convolutional Neural Network (DeCNN)," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 6, no. 2, pp. 222-230, Feb, 2022.

[30] S. H. Zhi, M. Kachelriess, F. Pan *et al.*, "CycN-Net: A Convolutional Neural Network Specialized for 4D CBCT Images Refinement," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3054-3064, 2021.

[31] P. F. Yang, X. Ge, T. F. Y. Tsui *et al.*, "Four-Dimensional Cone Beam CT Imaging Using a Single Routine Scan via Deep Learning," *Ieee Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1495-1508, May, 2023.

[32] G. Y. Chen, Y. S. Zhao, Q. Huang *et al.*, "4D-AirNet: a temporally-resolved CBCT slice reconstruction method synergizing analytical and iterative method with deep learning," *Physics in Medicine and Biology*, vol. 65, no. 17, 2020.

[33] F. Madesta, T. Sentker, T. Gauer *et al.*, "Self-contained deep learning-based boosting of 4D cone-beam CT reconstruction," *Med. Phys.*, vol. 47, no. 11, pp. 5619-5631, 2020.

[34] Z. H. Zhang, J. M. Liu, D. S. Yang *et al.*, "Deep learning-based motion compensation for four-dimensional cone-beam computed tomography (4D-CBCT) reconstruction," *Med. Phys.*, vol. 50, no. 2, pp. 808-820, 2023.

[35] Y. Dong, Y. H. Liu, H. Zhang *et al.*, "FD-GAN: Generative Adversarial Networks with Fusion-Discriminator for Single Image Dehazing," in 34th AAAI Conference on Artificial Intelligence New York, NY, 2020, pp. 10729-10736.

[36] D. Hu, Y. Zhang, J. Liu *et al.*, "SPECIAL: Single-Shot Projection Error Correction Integrated Adversarial Learning for Limited-Angle CT," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 734-746, 2021.

[37] H. Wu, Y. Qu, S. Lin *et al.*, "Contrastive Learning for Compact Single Image Dehazing," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10546-10555.

[38] G. Yang, S. M. Yu, H. Dong *et al.*, "DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1310-1321, 2018.
[39] Q. Yang, P. Yan, Y. Zhang *et al.*, "Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348-1357, 2018.
[40] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[41] D. Hu, Y. Zhang, W. Li *et al.*, "SEA-Net: Structure-Enhanced Attention Network for Limited-Angle CBCT Reconstruction of Clinical Projection Data," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-13, 2023.
[42] D. Hu, Y. Zhang, J. Liu *et al.*, "DIOR: Deep Iterative Optimization-Based Residual-Learning for Limited-Angle CT Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1778-1790, 2022.
[43] G. Balakrishnan, A. Zhao, M. R. Sabuncu *et al.*, "An Unsupervised Learning Model for Deformable Medical Image Registration," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9252-9260.
[44] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311-4322, 2006.
[45] T. Huang, S. J. Li, X. Jia *et al.*, "Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images," *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 14776-14785, 2021.
[46] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep Image Prior," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1867-1888, Jul, 2020.
[47] C.-C. Shieh, Y. Gonzalez, B. Li *et al.*, "SPARE: Sparse-view reconstruction challenge for 4D cone-beam CT from a 1-min scan," *Med. Phys.*, vol. 46, no. 9, pp. 3799-3811, 2019.
[48] M. Brehm, P. Paysan, M. Oelhafen *et al.*, "Self-adapting cyclic registration for motion-compensated cone-beam CT in image-guided radiation therapy," *Med. Phys.*, vol. 39, no. 12, pp. 7603-7618, Dec, 2012.
[49] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *J. Opt. Soc. Am. A*, vol. 1, no. 6, pp. 612-619, 1984/06/01, 1984.
[50] M. J. Riblett, G. E. Christensen, E. Weiss *et al.*, "Data-driven respiratory motion compensation for four-dimensional cone-beam computed tomography (4D-CBCT) using groupwise deformable registration," *Med. Phys.*, vol. 45, no. 10, pp. 4471-4482, 2018/10/01, 2018.
[51] Z. Zhang, X. Liang, X. Dong *et al.*, "A Sparse-View CT Reconstruction Method Based on Combination of DenseNet and Deconvolution," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1407-1417, 2018.
[52] S. W. Zamir, A. Arora, S. Khan *et al.*, "Restormer: Efficient Transformer for High-Resolution Image Restoration," *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5718-5729, 2022.
[53] H. Guo, J. Li, T. Dai *et al.*, "Mambair: A simple baseline for image restoration with state-space model," *arXiv preprint arXiv:2402.15648*, 2024.
[54] J. Y. Chen, E. C. Frey, Y. F. He *et al.*, "TransMorph: Transformer for unsupervised medical image registration," *Med. Image Anal.*, vol. 82, pp. 34, Nov, 2022.
[55] T. Guo, Y. Wang, and C. Meng, "Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration," *arXiv preprint arXiv:2401.13934*, 2024.