Prompting Lipschitz-constrained network for multiple-in-one sparse-view CT reconstruction

Baoshun Shi, *Senior Member, IEEE*, Ke Jiang, Qiusheng Lian, Xinran Yu, and Huazhu Fu, *Senior Member, IEEE*

***Abstract*— Despite significant advancements in deep learning-based sparse-view computed tomography (SVCT) reconstruction algorithms, these methods still encounter two primary limitations: (*i*) It is challenging to explicitly prove that the prior networks of deep unfolding algorithms satisfy Lipschitz constraints due to their empirically designed nature. (*ii*) The substantial storage costs of training a separate model for each setting in the case of multiple views hinder practical clinical applications. To address these issues, we elaborate an explicitly provable Lipschitz-constrained network, dubbed LipNet, and integrate an explicit prompt module to provide discriminative knowledge of different sparse sampling settings, enabling the treatment of multiple sparse view configurations within a single model. Furthermore, we develop a storage-saving deep unfolding framework for multiple-in-one SVCT reconstruction, termed PromptCT, which embeds LipNet as its prior network to ensure the convergence of its corresponding iterative algorithm. In simulated and real data experiments, PromptCT outperforms benchmark reconstruction algorithms in multiple-in-one SVCT reconstruction, achieving higher-quality reconstructions with lower storage costs. On the theoretical side, we explicitly demonstrate that LipNet satisfies boundary property, further proving its Lipschitz continuity and subsequently analyzing the convergence of the proposed iterative algorithms. The data and code are publicly available at `https://github.com/shibaoshun/PromptCT`.**

***Index Terms*— Sparse-view computed tomography, deep unfolding network, convergence analysis, prompt learning.**

I. Introduction

X-RAY computed tomography (CT) is a crucial diagnostic technology in medical imaging. However, prolonged or excessive exposure to X-ray radiation may pose potential risks of radiation-related diseases [1]. Sparse-view CT (SVCT) serves as an effective solution, reducing human radiation exposure by acquiring partial projection data through equidistant sampling over the full scanning range [2], [3]. This approach not only shortens scanning time but also mitigates motion artifacts caused by shaking, heartbeat, and respiration [4]. Nevertheless, the absence of projection data at certain angles may result in severe global streak artifacts in CT images reconstructed using filtered back-projection (FBP), potentially compromising critical tissue details and hindering clinical diagnosis. Effectively reconstructing high-quality CT images from sparse-view projection data remains a significant challenge [5].

Traditional iterative algorithms typically tackle the SVCT reconstruction problem by imposing various hand-crafted priors. Due to the insufficient characterization of image features by hand-crafted priors and numerous iterations required for algorithm convergence, these algorithms suffer from low-quality reconstruction and heavy computational burden. Inspired by the significant success of deep learning (DL) in medical image reconstruction, deep neural networks (DNNs) have been applied to the SVCT reconstruction task [6]. Despite the high-quality reconstructions achieved by these data-driven methods, the model architectures of DNNs typically lack interpretability, thereby hindering further theoretical analysis [7]–[10].

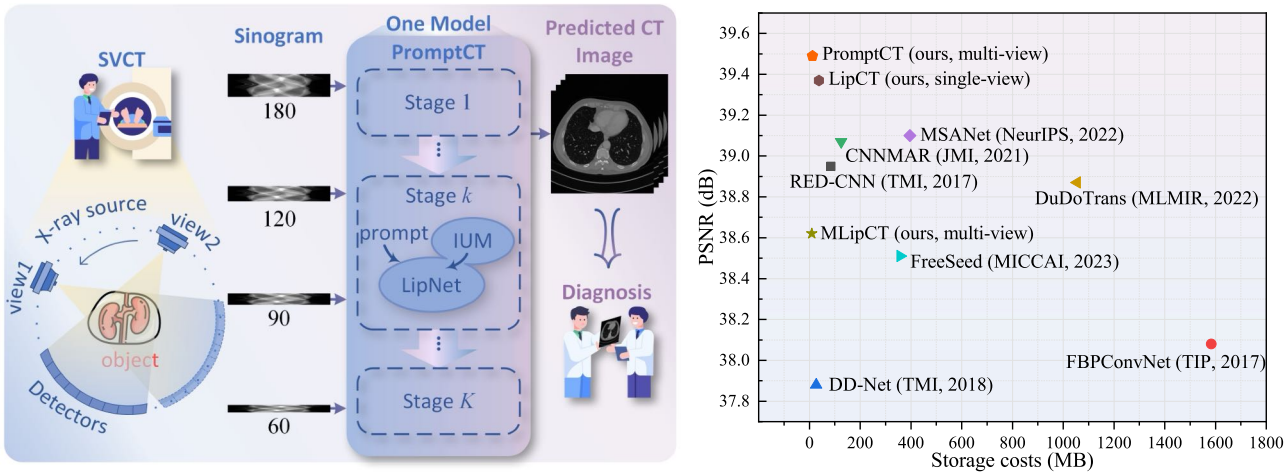
Recently, a promising research direction for interpretable networks is the ”iterative theory + deep learning” scheme, which can be divided into plug-and-play (PnP) reconstruction methods [11], [12] and deep unfolding reconstruction methods [13]. On the theoretical side, if the PnP iterative algorithms are unrolled, the convergence analysis of the PnP algorithms can be transferred to analyze the stability of the corresponding unrolled algorithms [14], i.e., the convergent PnP algorithms lead to stable unrolled algorithms, with their performance becoming increasingly stable as the stage increases [14]. However, the theoretical analysis of existing PnP or deep unfolding methods still faces several limitations [15], [16]. In particular, the selection and implementation of denoisers play a critical role in PnP reconstruction algorithm convergence [17]–[19]. In fact, to ensure algorithm convergence, the use of bounded denoisers necessitates that the gradients of data fidelity terms are bounded [20]. However, incomplete data, noise interference, and computational complexity often render closed-form solutions to optimization problems infeasible in most medical imaging tasks, potentially limiting the direct application of bounded denoisers in proving algorithm convergence. On the other hand, traditional deep unfolding algorithms often rely

This work was supported by the National Natural Science Foundation of China under Grants 62371414 and 62571473, by the Hebei Natural Science Foundation under Grant F2025203070, and by the Hebei Key Laboratory Project under Grants 202250701010046. The authors thank the anonymous reviewers for constructive suggestions. *(Baoshun Shi and Ke Jiang contributed equally to this work.)(Corresponding authors: Baoshun Shi and Qiusheng Lian.)*

Baoshun Shi, Ke Jiang, and Qiusheng Lian are with the School of Information Science and Engineering, Yanshan University, Qinhuang Dao, 066004, Hebei province, China (e-mail: shibaoshun@ysu.edu.cn, lianqs@ysu.edu.cn).

Xinran Yu is with the Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China.

Huazhu Fu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632.



(a) (b)

Fig. 1: (a) The proposed deep unfolding PromptCT enables SVCT reconstruction across varying numbers of projection data, e.g., 60, 90, 120, and 180, using a single model to aid in medical diagnosis and treatment, where each stage incorporates the image update module (IUM) and the prompt-based LipNet that satisfies the Lipschitz constraint. (b) We compare the average PSNR (dB) and storage cost (MB) of the proposed SVCT method and the existing SVCT methods under four sampling views. It reveals that the proposed strategy significantly saves storage costs while improving reconstruction performance. Notably, MLipCT is the multi-view model of the backbone architecture, i.e., LipCT, without explicit prompt module, and PromptCT is the multi-view model with explicit prompt module.

on empirically designed prior networks, which perform well in practical applications but are difficult to explicitly prove satisfying boundary properties or Lipschitz conditions. These conditions are common prerequisites for ensuring algorithm convergence.

On the practical side, there exists a range of sparse-view sampling strategies tailored to specific clinical requirements [21]. The distribution of artifacts in the reconstructed CT images varies under different sampling views or under-sampling rates. In recent years, existing DL-based SVCT reconstruction methods typically handle these sparse sampling configurations individually by training the separate model for each specific sparse-view setting [22]–[24]. Although this "one-model-for-one-setting" approach demonstrates excellent experimental performance, it requires substantial storage costs, and the flexibility of the single model limits clinical applications. Inspired by prompt learning, all-in-one methods typically train a general model for various tasks, but they require extensive datasets [25]. However, acquiring large amounts of paired CT data is impractical in medical imaging.

To tackle the aforementioned issues, we propose a prompting Lipschitz-constrained network for multiple-in-one SVCT reconstruction. This "one-model-for-multi-view" strategy aims to develop a universal multi-view model for multiple sampling views in the SVCT reconstruction task, thereby generating high-quality CT images to assist doctors in diagnosis. In theory, the proposed Lipschitz-constrained network, serving as a prior network, cleverly integrates the two theories of boundary property and Lipschitz constraint, which successfully circumvents the stringent restrictions on the data fidelity terms while ensuring the convergence of iterative algorithms. Particularly, we introduce sparse sampling masks as explicit prompts to develop a flexible single model that is trained to handle the SVCT reconstruction with different sampling ratios. The contributions of this paper are as follows:

- We propose an explicitly provable Lipschitz-constrained sparse representation model-driven network, termed LipNet, constructed from a deep unfolding sparse representation framework that satisfies boundary property. Within LipNet, we elaborate a constant-generating sub-network (CGNet) with an explicit prompt module to determine view-aware and spatial-variant thresholds.
- We exploit sinogram sampling masks as explicit prompts to distinguish view information from different sparse-view settings. Even using thousands of samples, the explicit prompt module can classify mixed feature information, enabling the reconstruction network to handle multiple sparse sampling settings within a single universal model.
- We devise the so-called CGNet composed of the shallow feature extraction module, swin Transformer block (STB), and spatial frequency block (SFB) to adaptively generate the proportional constants of thresholds by extracting local, regional, and global information from representation coefficients, thereby improving the reconstruction performance and representation ability.
- As depicted in Fig. 1, we construct a storage-saving deep unfolding network for multiple-in-one SVCT reconstruction, called PromptCT, which incorporates the proximal gradient descent technique for algorithm optimization. We replace the proximal operator in PromptCT with the well-designed LipNet, providing a clear working mechanism and convergence analysis. Comprehensive experiments, including synthetic-to-real generalization, finely substantiate that PromptCT achieves outstanding performance across various sparse-view settings using a multi-view model, surpassing existing SVCT reconstruction methods. On the theoretical side, we explicitly demonstrate that LipNet satisfies boundary property and Lipschitz continuous, and further prove the convergence of the

proposed iterative algorithm.

The remainder of the manuscript is structured as follows. In Section II, we provide a brief overview of related works. Section III introduces the prompting Lipschitz-constrained network and its theoretical analysis. Section IV focuses on introducing the deep unfolding SVCT reconstruction network and its theoretical analysis. Section V showcases the experimental evaluations to validate the superiority of the proposed network. Section VI summarizes our work and outlines potential directions for future research.

II. Related Work

A. SVCT reconstruction

In general, existing DL-based SVCT reconstruction approaches can be roughly grouped into two categories, i.e., data-driven methods and model-driven methods [26]. Specifically, data-driven SVCT reconstruction methods rely on a large number of input-output data pairs for training, directly learning the mapping relationship and feature representations of the reconstruction task from the data [27], [28]. Although these methods, which do not rely on prior knowledge or require modeling of physical processes, often exhibit strong flexibility, the black-box nature of DNNs typically results in insufficient model interpretability for these data-driven networks.

In recent years, many efforts have focused on incorporating the DNNs into the traditional iterative framework through PnP strategy [29] or deep unfolding strategy [2], [13], [30], to improve the interpretability and representation ability of the network. In theory, common PnP frameworks combining iterative reconstruction algorithms with pre-trained Gaussian denoisers can achieve promising performance in a wide range of imaging tasks. However, the solutions provided by these frameworks remain suboptimal due to the non-adaptive deep Gaussian denoiser. In contrast, the deep unfolding method replaces these pre-trained deep denoisers of iterative algorithms with DNNs through end-to-end learning [31], [32]. These model-driven networks combine the interpretability of iterative algorithms with the powerful reconstruction ability of learning-based approaches.

B. Prompt learning

Prompt learning [33] is an emerging technique in the fields of natural language processing and machine learning, experiencing rapid development and garnering widespread attention. Initially, prompt learning involves studying how to introduce additional texts (i.e., prompts) as inputs to pre-trained large language models to obtain the desired outputs. With further research, prompt-based approaches aim to provide contextual information to the model for fine-tuning towards target tasks, enabling parameters to adapt the model more effectively, thus offering significant flexibility. In the field of image restoration, some works utilize learnable prompts to distinguish different image restoration tasks and demonstrate their effectiveness in low-level vision tasks [25]. For example, Gao et al. [34] employed degradation-specific information to dynamically guide restoration networks using learned prompts, while Zhang et al. [35] proposed a two-stage approach that encompasses task-specific knowledge collection and component-oriented knowledge integration. Further advancements came from Park et al. [36], who employed adaptive discriminative filters, and Zhu et al. [37], who concentrated on learning both general and weather-specific features to improve performance in weather-degraded conditions. Learnable prompts require a large number of samples for effective feature discrimination, thus this strategy is not applicable in medical image reconstruction tasks since it is difficult to acquire massive CT images. Therefore, we propose using explicit prompts instead of learnable prompts to efficiently embed discriminative information into the data flow of the reconstruction network only using thousands of samples.

III. The Proposed Prompting Lipschitz-constrained Network

A. LipNet: when bounded network meets Lipschitz-constrained network

The convergence proofs of the PnP algorithms can be transferred to analyze the stability of deep unfolding methods [14]. Therefore, the properties of denoisers in PnP algorithms can be extended to the prior networks in deep unfolding algorithms. Based on this fact, we extend the bounded denoiser to the bounded network and generalize the concept of Lipschitz constraint to the Lipschitz-constrained network, both of which can serve as prior networks extended to deep unfolding algorithms, ensuring the convergence of their iterative algorithms. The definitions of bounded network and Lipschitz-constrained network are provided as follows.

Definition 1 (Bounded network): The procedure of a DNN denoted as $\mathcal{D}_B(\cdot;\sigma)$ with a input parameter σ: $\mathbb{R}^n \to \mathbb{R}^n$ such that for any input $\boldsymbol{x} \in \mathbb{R}^n$, the following inequality holds

$$||\mathcal{D}_B(\boldsymbol{x};\sigma) - \boldsymbol{x}||_2^2 \leq \sigma^2 C \tag{1}$$

where σ is the noise level and the universal constant C is independent of σ. Under the condition of bounded networks, the diminishing noise level condition is crucial for the convergence analysis of the algorithm.

Definition 2 (Lipschitz-constrained network): The procedure of a DNN denoted as $\mathcal{D}_L(\cdot)$ is Lipschitz continuous with a constant $L > 0$ if

$$||\mathcal{D}_L(\boldsymbol{x}_1) - \mathcal{D}_L(\boldsymbol{x}_2)||_2 \leq L||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2 \tag{2}$$

for any input $\boldsymbol{x}_1, \boldsymbol{x}_2 \to \mathbb{R}^n$.

In PnP reconstruction algorithms, employing a bounded denoiser to demonstrate algorithm convergence requires ensuring that the gradient of the data fidelity term is bounded [11], [38]. However, the optimization problem containing the data fidelity term demands solving closed-form solutions, which poses practical challenges for complex imaging scenarios such as the SVCT reconstruction task. On the other hand, although imposing the assumption that denoisers satisfy Lipschitz constraints can guarantee the convergence of PnP reconstruction algorithms, explicitly proving that denoisers adhere to Lipschitz constraints remains a challenge. Some efforts focus on imposing constraints on filter weights, but these

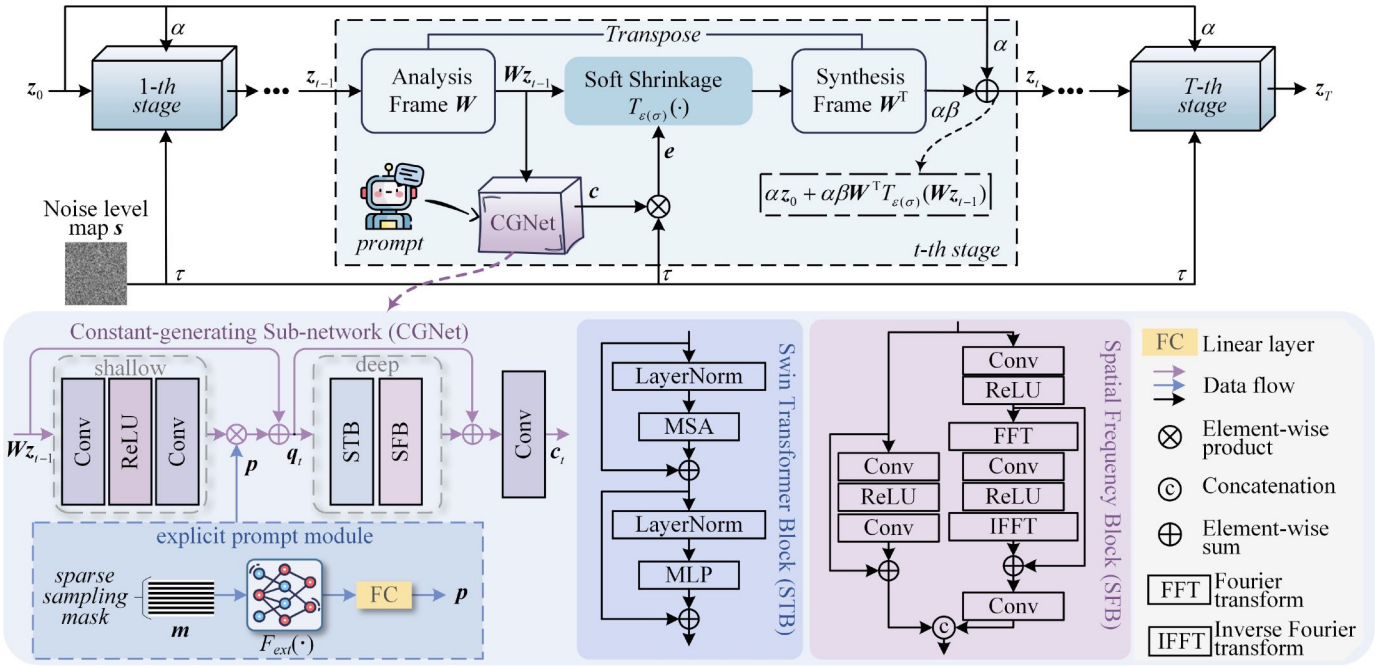


Fig. 2: The network architecture of LipNet. Each stage of this network consists of an analysis frame $\boldsymbol{W}$, a soft shrinkage operation, a synthesis frame $\boldsymbol{W}^{\mathrm{T}}$, and a constant-generating sub-network (CGNet), forming a single-layer sparse representation model-driven architecture. The element-wise products of the generated proportional constants $\boldsymbol{c}$ and the elements in the input noisy map $\boldsymbol{s}$ are used as the thresholds $\boldsymbol{e}$ for shrinking frame coefficients. The network architecture of CGNet comprises the shallow feature extraction module, the deep feature extraction module, and the explicit prompt module, utilized for generating spatial-variant proportional constants. Among them, the shallow feature extraction contains two 3×3 convolutional layers, and the deep feature extraction consists of STB and SFB for extracting local, regional, and global feature information, respectively. In addition, we embed view prompts between these two feature extraction modules to guide the differentiation of multiple sparse sampling views.

approaches reduce the representation ability of the network, thereby leading to a degradation in reconstruction performance [17]. In this paper, we integrate the theories of boundary properties and Lipschitz constraints to construct a network that explicitly satisfies Lipschitz constraints by leveraging bounded networks. This approach effectively mitigates the constraints associated with the two major theories, thereby avoiding the need for closed-form solutions and making it more suitable for SVCT reconstruction tasks.

As a popular representation model, the sparse representation model, which can promote the sparsity of images over some sparsifying frames, is commonly used to construct provable networks and has been proven effective in imaging fields [11], [39]. Specially, a frame $\boldsymbol{W}$ equipped with the tight property $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}$ is called a tight frame, whose inverse transform is its transpose. Tight frames can be roughly classified into analytical tight frames [40] and data-driven tight frames [41]. Imaging algorithms using analytical tight frames are often faster, but their reconstructions are inferior compared with those using data-driven tight frames. Data-driven tight frames, being adaptive to the data, can capture more structural information compared to analytical ones. In general, traditional tight frame learning methods often update the tight frames and their corresponding frame coefficients via a time-consuming alternating optimization strategy, which hinders the incorporation of modern deep learning techniques [42]. Existing bounded denoising networks are predominantly constructed using tight frames [12], [38], which must adhere to the strict constraint $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}$. This requirement imposes substantial limitations on their flexibility and representation ability.

To address this limitation, we cleverly relax the strict tight frame constraint and propose a novel approach that constructs bounded networks using general sparsifying frames instead of tight frames. Specifically, we propose to learn both the analysis frame $\boldsymbol{W}$ and the synthesis frame $\boldsymbol{W}^{\mathrm{T}}$ in an end-to-end supervised learning manner. To further enhance the representation ability of these sparsifying frames, we design a CGNet that not only improves their adaptability but also eliminates the need for laborious threshold tuning. This approach leverages the benefits of frame-based representation while overcoming the drawbacks of tight frame constraints, providing a more flexible and effective framework for image reconstruction tasks. Furthermore, instead of strict sparsity, we explore approximate sparsity, which is widely applied in the sparsifying transform learning [43], [44]. Assuming that the input image is approximately sparse under the learnable sparsifying frame $\boldsymbol{W}$, the optimization problem for filtering can be formulated as

$$\min_{\boldsymbol{z}} \frac{1}{2}||\boldsymbol{z}_0 - \boldsymbol{z}||_2^2 + \mu||\boldsymbol{W}\boldsymbol{z}||_1 \tag{3}$$

where $\boldsymbol{z}_0$ is the underlying image, $\boldsymbol{z}$ is the introduced auxiliary

variable, μ denotes the trade-off parameter, $||\cdot||_1$ represents the l_1 norm, and $\boldsymbol{W}$ is the learnable sparsifying frame or filter. Notably, compared with previous tight frame networks [11], [19], [38], our filter does not require the tight constraint. By leveraging an auxiliary variable $\boldsymbol{u}$, Eqn. (3) can be rewritten as

$$\min_{\boldsymbol{z},\boldsymbol{u}} \frac{1}{2}||\boldsymbol{z}_0 - \boldsymbol{z}||_2^2 + \mu||\boldsymbol{u}||_1, \ s.t. \ \boldsymbol{W}\boldsymbol{z} = \boldsymbol{u}. \tag{4}$$

Subsequently, we solve the problem defined in Eqn. (4) by using the Half Quadratic Splitting (HQS) solver. Thus, a sequence of individual subproblems emerges (for the t-th iteration):

$$\boldsymbol{z}_t = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}||\boldsymbol{z}_0 - \boldsymbol{z}||_2^2 + \frac{\beta}{2}||\boldsymbol{W}\boldsymbol{z} - \boldsymbol{u}_{t-1}||_2^2 \right\} \tag{5}$$

and

$$\boldsymbol{u}_t = \arg\min_{\boldsymbol{u}} \left\{ \frac{\beta}{2}||\boldsymbol{W}\boldsymbol{z}_t - \boldsymbol{u}||_2^2 + \mu||\boldsymbol{u}||_1 \right\} \tag{6}$$

where β stands for a penalty parameter. The iterative solutions of subproblems for $\boldsymbol{z}$ and $\boldsymbol{u}$ are defined as

$$\boldsymbol{z}_t = \frac{\boldsymbol{z}_0 + \beta\boldsymbol{W}^{\mathrm{T}}\boldsymbol{u}_{t-1}}{\boldsymbol{I} + \beta\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}} \tag{7}$$

and

$$\boldsymbol{u}_t = T_{\varepsilon(\sigma)}(\boldsymbol{W}\boldsymbol{z}_t) \tag{8}$$

where $T_{\varepsilon(\sigma)}(\cdot)$ is the soft thresholding operator $soft_{\varepsilon(\sigma)} = sign(u)\max(|u| - \varepsilon(\sigma), 0)$ and the threshold $\varepsilon(\sigma)$ is correlated with the noise strandard deviation σ. Based on Eqn. (8), we have $\boldsymbol{u}_{t-1} = T_{\varepsilon(\sigma)}(\boldsymbol{W}\boldsymbol{z}_{t-1})$. Substituting this equation into Eqn. (7), we can recast Eqn. (7) as

$$\boldsymbol{z}_t = \frac{\boldsymbol{z}_0 + \beta\boldsymbol{W}^{\mathrm{T}}T_{\varepsilon(\sigma)}(\boldsymbol{W}\boldsymbol{z}_{t-1})}{\boldsymbol{I} + \beta\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}}. \tag{9}$$

Define $\alpha = \frac{1}{\boldsymbol{I}+\beta\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}}$ as a constant less than 1, the proposed iterative updated rule is summarized as follows

$$\boldsymbol{z}_t = \alpha\boldsymbol{z}_0 + \alpha\beta\boldsymbol{W}^{\mathrm{T}}T_{\varepsilon(\sigma)}(\boldsymbol{W}\boldsymbol{z}_{t-1}). \tag{10}$$

In summary, the problem defined in Eqn. (3) can be solved via iterating the inversion and filter steps alternatively. Considering the aforementioned solving process, the proposed LipNet incorporates a parameter σ and is modeled as a function $\mathcal{D}_\theta(\cdot;\sigma)$, defined as $\boldsymbol{z}_T = \mathcal{D}_\theta(\boldsymbol{z}_0;\sigma)$, where θ represents the set of learnable parameters within this DNN. Based on the interpretability of the sparse representation model and the strong learning ability of DNNs, the proposed LipNet is constructed as a deep unfolding network in the model- and data-driven manner by unfolding the iteration defined in Eqn. (10). The architecture of our proposed LipNet is illustrated in Fig. 2. In LipNet, each stage corresponds rigorously to each iteration of the HQS algorithm. Moreover, in order to enhance the representation ability of the network, we elaborate a constant-generating sub-network based on explicit prompts to generate adaptive thresholds, and the concrete network architecture is introduced in Section III-B.

B. The prompt-based constant-generating sub-network

The thresholds $\varepsilon(\sigma)$ in Eqn. (10), which can be learned in a supervised learning manner, are crucial for the representation ability of sparse representation model-driven networks [12]. Based on the so-called universal threshold theorem, the thresholds of soft shrinkage are proportional to the noise level, i.e., $\varepsilon_i = c_i \cdot \sigma$. Here ε_i is the threshold for the i-th channel of the feature map, and c_i is the corresponding proportional constant. Since image structures are varying in the spatial domain, the corresponding threshold should be spatially varying. In order to generate spatially varying thresholds, we elaborate the so-called CGNet which can generate proportional constants. The thresholds can be determined by the product of the proportional constants and the noise levels. Formally, the thresholds can be defined as

$$\boldsymbol{e} = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_i, \cdots]^{\mathrm{T}} = \boldsymbol{c} \otimes \boldsymbol{s} \tag{11}$$

where $\otimes$ denotes the element-wise product, and $\boldsymbol{s}$ is the noise level map obtained by stretching the noise level σ. The size of $\boldsymbol{s}$ is the same as that of learned proportional constant vector $\boldsymbol{c}$. To avoid arbitrary values, the elements of $\boldsymbol{c}$ are limited to $[c_{min}, c_{max}]$. By doing so, the thresholds are instance-optimal and spatial-variant. Since the image quality improves with each iteration, σ should gradually decrease [38]. In this paper, we introduce the learnable parameter $\tau \in (0, 1)$, such that $\sigma_t = \tau\sigma_{t-1}$.

The overall architecture of CGNet: To provide a clearer understanding of CGNet, we present its overall architecture in Fig. 2. Specifically, the proposed CGNet consists of the shallow feature extraction module, the deep feature extraction module, the explicit prompt module, and the 3×3 convolutional layer, aiming to generate spatial-variant proportional constants to enhance the representation ability of the overall network. In shallow feature extraction, stacked convolutional layers perform filtering operations on the input to extract local features and structural information. Mathematically, the shallow feature extraction operator $\mathcal{F}_{shallow}(\cdot)$ can be expressed as follows:

$$\mathcal{F}_{shallow}(\cdot) = Conv(ReLU(Conv(\cdot))) \tag{12}$$

where $ReLU(\cdot)$ denotes the rectified linear unit activation layer and $Conv(\cdot)$ indicates the 3×3 convolutional layer. Building upon this, we introduce an explicit prompt module to encode specific image knowledge and sampling information, embedding it between shallow and deep feature extraction, thus enabling accurate discrimination of different sparse sampling settings. We define the prompt information obtained from the explicit prompt module as $\boldsymbol{p}$. At the t-th stage, the feature refinement process can be represented mathematically as follows:

$$\boldsymbol{q}_t = \mathcal{F}_{shallow}(\boldsymbol{W}\boldsymbol{z}_{t-1}) \otimes \boldsymbol{p} + \boldsymbol{W}\boldsymbol{z}_{t-1} \tag{13}$$

where $\boldsymbol{W}\boldsymbol{z}_{t-1}$ is the coefficient vector and $\boldsymbol{q}_t$ denotes the feature map obtained through shallow feature extraction. Furthermore, in deep feature extraction, we employ the STB to compensate for the limitations of the receptive field imposed

by convolutional operations, thereby capturing regional information on features and effectively extracting deep features. Subsequently, the SFB is employed to extract more comprehensive and detailed features. The process of extracting deep feature can be mathematically expressed as follows:

$$\mathcal{F}_{deep}(\cdot) = \mathcal{F}_{SFB}(\mathcal{F}_{STB}(\cdot)) \tag{14}$$

where $\mathcal{F}_{deep}(\cdot)$ is the deep feature extraction operator, $\mathcal{F}_{STB}(\cdot)$ represents the procedure of the Swin Transformer block that captures global dependencies and long-range interactions in the feature space, and $\mathcal{F}_{SFB}(\cdot)$ denotes the spatial frequency block, which enhances the feature maps by emphasizing important spatial and frequency components to refine and strengthen the feature representation. Specifically, the fast fourier convolution for global information extraction in the frequency domain branch, and CNN-based residual modules for enhanced local feature representation in the spatial domain branch. Finally, the global skip connection combines shallow features with deep features to further enhance the model stability and the representation ability.

To adaptively generate the thresholds and adapt to different sparse-view settings, a proportional constant vector is generated at each stage of the network. This constant vector plays a crucial role in balancing the contributions of different spatially varying features during the iterative process. The proportional constant-generating process at the t-th stage can be formulated as:

$$\boldsymbol{c}_t = Conv(\mathcal{F}_{deep}(\mathcal{F}_{shallow}(\boldsymbol{W}\boldsymbol{z}_{t-1}) \otimes \boldsymbol{p} + \boldsymbol{W}\boldsymbol{z}_{t-1})) + \boldsymbol{q}_t) \tag{15}$$

where $\boldsymbol{c}_t$ represents the corresponding output proportional constant map.

Explicit prompt module: Prompt learning techniques have been widely employed in general image restoration as effective tools for various restoration tasks [25], each equipped with learnable prompts to interact with input images or latent features. Given the difficulty in acquiring a sufficient number of paired images in practical scenarios, learning prompts that can effectively discriminate between a large number of fine-grained sparse sampling is very challenging. To address this issue, we propose a more applicable explicit prompt module that learns discriminative information from input explicit prompts. Specifically, considering the differences in CT image distribution under different view numbers in the SVCT reconstruction task, we introduce a prompt module to effectively guide the generation of view-aware adaptive thresholds for different sparse view numbers. The essential difference in the sparse sampling configurations stems from the sampling distribution of projection data. Hence, we adopt a binary down-sampling matrix mask $\boldsymbol{M} \in \mathbb{R}^{N_p \times N_b}$ with 0 indicating the missing region, represented by the vector $\boldsymbol{m}$, naturally encoding the sampling information to construct explicit view prompts. We employ three convolutional layers to extract features related to sparse sampling from the down-sampling matrix, then apply full connection layer to convert the features into the suitable shape ($1 \times$ channel or dimension) for the corresponding module outputs. Formally, the explicit prompt module can be formulated by

$$\boldsymbol{p} = FC(F_{ext}(\boldsymbol{m})) \tag{16}$$

where $\boldsymbol{p}$ indicates the prompt vector, $FC(\cdot)$ refers to the fully connected layer operation, and $F_{ext}(\cdot)$ denotes the CNN operation. We use convolution layers with a kernel size of 3×3 to extract feature information from the sampling settings and apply the ReLU activation function to enhance the ability of the network to model nonlinearities. By incorporating this prompt-based approach, LipNet can adaptively generate prompt features that carry discriminative information specific to each sampling setting, enabling universal SVCT reconstruction with a single model.

C. Theoretical analysis of LipNet

In this section, in order to demonstrate that LipNet satisfies the boundary property and Lipschitz constraint, we first present **Lemma 1** that proves the boundary property of LipNet, and then prove that LipNet satisfies the Lipschitz continuity. Based on **Lemma 1**, **Theorem 1** claims LipNet is a bounded network. Furthermore, **Theorem 2** claims the Lipschitz continuity of LipNet based on **Theorem 1**. The detailed proofs of **Lemma 1**, **Theorem 1**, and **Theorem 2** can be found in the supplementary material[1].

Lemma 1: For any input $\boldsymbol{x} \in \mathbb{R}^n$ and some universal constant $L = \tau_0 c_{max}^2$ (Here, c_{max} denotes the maximum element of proportional constant vector $\boldsymbol{c}$) independent of M, we can get

$$||\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x} - \boldsymbol{W}^{\mathrm{T}}T_{\varepsilon(\sigma)}(\boldsymbol{W}\boldsymbol{x})||_2^2 \leqslant M\sigma^2 L \tag{17}$$

where M denotes the dimension of the threshold vector $\boldsymbol{e}$ and σ indicates the input noise level.

Theorem 1 (Boundary property of LipNet): For any input $\boldsymbol{x} \in \mathbb{R}^n$, the proposed LipNet $\mathcal{D}_\theta(\cdot;\sigma)$ based on **Lemma 1** is bounded such that

$$||\boldsymbol{x} - \mathcal{D}_\theta(\boldsymbol{x};\sigma)||_2^2 \leqslant N\sigma^2 \tag{18}$$

for some universal constant N independent of σ.

Theorem 2 (Lipschitz continuity of LipNet): Based on **Theorem 1**, the proposed LipNet $\mathcal{D}_\theta(\cdot;\sigma)$ is υ-Lipschitz continuous. This means that there exists a $\upsilon > 0$ such that for all $\boldsymbol{x}_1, \boldsymbol{x}_2$, the following relationship holds

$$||\mathcal{D}_\theta(\boldsymbol{x}_1;\sigma) - \mathcal{D}_\theta(\boldsymbol{x}_2;\sigma)||_2^2 \leqslant \upsilon||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2^2. \tag{19}$$

IV. The Proposed PromptCT for SVCT Reconstruction

A. Problem formulation and deep unfolding network

In fan-beam SVCT system configuration, sparse-view projection data $\boldsymbol{Y} \in \mathbb{R}^{N_p \times N_b}$ (a.k.a raw projection or sinogram), obtained through uniform sampling via a 360-degree rotation of the detector, where N_p and N_b respectively represent the number of sparse-view projections and detector bins. The SVCT reconstruction aims to recover the underlying CT image from the observed incomplete projection data, for which we

[1] https://github.com/shibaoshun/PromptCT

can formulate the corresponding minimization problem as follows:

$$\min_{\boldsymbol{x}} \frac{1}{2}||\boldsymbol{y}-\mathcal{P}\boldsymbol{x}||_2^2+\lambda R(\boldsymbol{x}) \tag{20}$$

where $\mathcal{P}$ is the forward projection operator, specifically the Radon transform under sparse-view conditions. We define the under-sampling data as $\boldsymbol{y} = vec(\boldsymbol{Y}) \in \mathbb{R}^{N_p N_b}$ using the vector operator. The first term is known as the data-fidelity term, which encourages consistency between the recovered CT image and the projection data. The second term represents the regularization term, imposing desired properties on the image to be reconstructed. The parameter λ is a regularization parameter that pursues the trade-off between these two terms.

The proximal gradient descent algorithm applied to the optimization problem in Eqn. (20) yields the following iterations

$$\begin{cases} \boldsymbol{x}^{(k+0.5)} = \boldsymbol{x}^{(k)} + \eta\mathcal{P}^{\mathrm{T}}(\boldsymbol{y}-\mathcal{P}\boldsymbol{x}^{(k)}) & (21) \\ \boldsymbol{x}^{(k+1)} = prox_{\eta R}(\boldsymbol{x}^{(k+0.5)}) & (22) \end{cases}$$

where $\eta > 0$ is the step size and $prox_{\eta R}(\cdot)$ is a proximal operator related to the prior form $R(\cdot)$ about $\boldsymbol{x}$. The fundamental idea behind the deep unfolding approach is to fix a maximum number of iterations denoted as K, and declare $\boldsymbol{x}^{(K)}$ as our estimate for $\boldsymbol{x}$. We refer to Eqn. (21) as the image update module (IUM) and consider replacing $prox_{\eta R}(\cdot)$ with a trainable prior network $\mathcal{D}_\theta(\cdot;\sigma)$, which gives the iteration map

$$\boldsymbol{x}^{(k+1)} = \mathcal{D}_\theta(\boldsymbol{x}^{(k+0.5)};\sigma). \tag{23}$$

The prior network $\mathcal{D}_\theta(\cdot;\sigma)$ defined in Eqn. (23) is crucial for enhancing reconstruction quality. Therefore, we employ the well-designed LipNet as the prior network to improve the reconstruction performance through strong representation ability. Figure. 3 shows the overall network architecture of the proposed deep unfolding proximal gradient descent network for the SVCT reconstruction task. Specifically, PromptCT consists of K stages corresponding to K iterations of the iterative algorithm, and each stage contains IUM and LipNet.

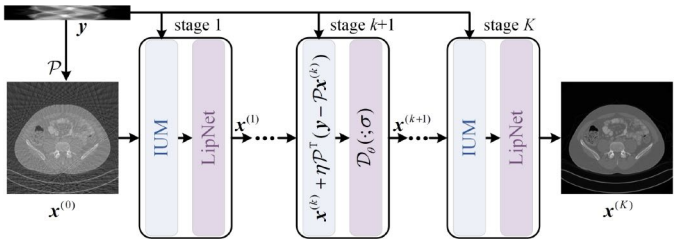


Fig. 3: The proposed deep unfolding proximal gradient descent network architecture (i.e., PromptCT) consists of the IUM and the prompting Lipschitz-constrained network $\mathcal{D}_\theta(\cdot;\sigma)$ (i.e., LipNet).

B. Loss Functions

As mentioned earlier, we design a prompt-based deep unfolding network for the SVCT reconstruction task. For the training process of PromptCT, in order to remove artifacts while preserving the global structures of the output CT image, we supervise the restored CT image $\boldsymbol{x}^{(k)}$ updated at each stage using the l_1 loss and the l_2 loss. The total loss function of training PromptCT can be formulated as

$$\begin{aligned}\mathcal{L} = &\sum_{k=0}^{K}\left\{\omega_1||\boldsymbol{x}^{(k)}-\boldsymbol{x}_{gt}||_1+\omega_2||\boldsymbol{x}^{(k)}-\boldsymbol{x}_{gt}||_2^2\right\} \\ &+||\boldsymbol{x}-\boldsymbol{x}_{gt}||_2^2\end{aligned} \tag{24}$$

where $\boldsymbol{x}_{gt}$ is the ground truth CT image. Meanwhile, ω_1 and ω_2 are hyper-parameters to balance the weights of different loss items. Heuristically, we set $\omega_1 = 0.1$ and $\omega_2 = 0.1$.

C. Theoretical analysis

Classical fixed-point theory indicates that the iterations converge to a unique fixed-point if the iteration map $f_\theta(\cdot;\boldsymbol{y})$ is contractive [48], i.e., there exists a constant $0 \le c < 1$ such that $||f_\theta(\boldsymbol{x}_1;\boldsymbol{y}) - f_\theta(\boldsymbol{x}_2;\boldsymbol{y})||_2^2 \le c||\boldsymbol{x}_1-\boldsymbol{x}_2||_2^2$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. Based on this observation, we derive the following convergence result for PromptCT:

Theorem 3 (Convergence of PromptCT): According to **Theorem 2**, the network $\mathcal{D}_\theta(\cdot;\sigma)$ is υ-Lipschitz. Assume that $\epsilon = \delta_{min}(\mathcal{P}^{\mathrm{T}}\mathcal{P})$ and $\zeta = \delta_{max}(\mathcal{P}^{\mathrm{T}}\mathcal{P})$, where $\delta_{min}(\cdot)$ and $\delta_{max}(\cdot)$ denote the maximum and minimum eigenvalue, respectively. Then the iteration map of the proposed PromptCT, i.e., $f_\theta(\cdot;\boldsymbol{y})$, is contractive if the step size parameter η satisfies

$$\frac{1}{\epsilon}-\frac{1}{\upsilon\epsilon} < \eta < \frac{1}{\zeta}+\frac{1}{\upsilon\zeta} \tag{25}$$

Such an υ exists if $\upsilon < (\epsilon+\zeta)/(\zeta-\epsilon)$.

Proof: See the supplementary material[2].

V. Experiments

A. Experimental setup

1) Datasets: We perform experiments on two publicly available real-world CT image datasets: the 2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge dataset [49] and the DeepLesion dataset [50]. To evaluate the feasibility of our method in practical applications, experiments are conducted on real objects, including the equivalent water-bone phantom [51] and the pork with bone slices. The experimental setup is detailed below. The phantom is scanned using the industrial CT system in our laboratory. The X-ray source is HAMAMATSU-L12161-07, and the detector used is the EIGER2 1MW R-DECTRIS. The voltage and current settings for the X-ray source are 100 kVp and 500 μA, respectively. The detector integration time is set to 2 seconds. The distance from the X-ray source to the rotation center is 220.798 mm, and the distance from the rotation center to the detector is 197.247 mm. The detector has a unit size of 0.3 mm with 512 units in total. The sampling angular range spans 0 to 360°, with 60 angular sampling points. For simplicity, the experiments are restricted to the 2D fan-beam case. There should be no difficulty to extend it to the 3D cone-beam case. To validate that our method can be extended to 3D scenarios, we select the vertebrae localization and identification dataset from SpineWeb [52], which includes 14 testing volumes labeled

[2]https://github.com/shibaoshun/PromptCT

TABLE I: Quantitative evaluations for different SVCT methods on the AAPM dataset under various sparse-view conditions. We report the average PSNR (dB)↑/ SSIM↑/ RMSE↓ values of the testing dataset for each case. The best results are highlighted in red and the second-best results are highlighted in blue.

Method	60			90			120			180			average		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
single-view models															
FBP	13.06	0.3884	0.2242	14.47	0.5087	0.1906	16.16	0.6160	0.1570	20.99	0.7746	0.0900	16.17	0.5719	0.1655
RED-CNN [45]	36.09	0.9129	0.0159	38.28	0.9309	0.0123	39.78	0.9429	0.0103	41.63	0.9516	0.0083	38.95	0.9346	0.0117
FBPConvNet [22]	34.74	0.9036	0.0188	37.23	0.9249	0.0140	38.99	0.9374	0.0114	41.35	0.9540	0.0086	38.08	0.9300	0.0132
DD-Net [46]	35.50	0.8762	0.0170	37.28	0.8668	0.0138	38.46	0.8927	0.0120	40.29	0.9133	0.0097	37.88	0.8873	0.0131
CNNMAR [24]	36.11	0.9155	0.0159	38.42	0.9358	0.0122	39.77	0.9458	0.0104	41.97	0.9582	0.0080	39.07	0.9388	0.0116
MSANet [47]	36.20	0.9170	0.0157	38.36	0.9331	0.0122	39.91	0.9445	0.0102	41.93	0.9580	0.0081	39.10	0.9382	0.0116
DuDoTrans [23]	35.60	0.9049	0.0168	38.11	0.9274	0.0125	39.75	0.9405	0.0104	42.03	0.9571	0.0080	38.87	0.9325	0.0119
FreeSeed [7]	35.68	0.9005	0.0167	37.86	0.9257	0.0129	39.25	0.9376	0.0110	41.25	0.9524	0.0087	38.51	0.9291	0.0123
LipCT (ours)	36.38	0.9184	0.0154	38.84	0.9374	0.0115	40.10	0.9466	0.0100	42.16	0.9598	0.0078	39.37	0.9406	0.0112
multi-view models															
MLipCT (ours)	36.16	0.9167	0.0158	38.13	0.9309	0.0126	39.38	0.9399	0.0109	40.80	0.9528	0.0092	38.62	0.9351	0.0121
PromptCT (ours)	36.95	0.9224	0.0144	38.95	0.9375	0.0114	40.16	0.9464	0.0099	41.91	0.9578	0.0081	39.49	0.9410	0.0110

with multi-bone, i.e., sacrum, left hip, right hip, and lumbar spine. The clinical images are resized and processed using the same protocol to the synthesized data. We conduct experiments on the SpineWeb dataset and generate axial, coronal, and sagittal CT images for evaluation.

The normal-dose Mayo data are acquired using a 120 kVp and 235 effective mAs (500 mA/0.47 s) protocol, scanning from the chest to the abdomen. We select 1000 images for training and 500 images for testing, with each image resized to 512×512. The experiments are performed in a fan-beam X-ray source scanning setup with 800 detector elements. We generate a fully sampled sinogram by acquiring 360 projection views at regular intervals between 0 and 360 degrees. We obtain sparse-view projection data by uniformly sampling 60, 90, 120, and 180 views from the full-view sinogram. In the numerical experiments, the simulated sinogram is contaminated by Poisson noise with an intensity of 5×10^6 and electronic system noise (with standard deviation 5%). The electronic noise level is regarded to follow a zero-mean normal distribution and is assumed to be stable for a commercial CT scanner. For the DeepLesion dataset, we randomly choose 1000 images as the training set and another 200 images as the test set. Concretely, we use a 120 kVp polyenergetic X-ray source to simulate the equiangular fan-beam projection geometry and the incident X-ray contains 2×10^7 photons. We specify that the full-sampling sinograms are generated by uniformly spaced 360 projection views between 0 and 360 degrees, and the number of detector elements is 641. The corresponding artifact-affected CT images are reconstructed from the sparse-view sinograms by FBP, and all CT images are resized to 416×416. To simulate the photon noise numerically, we add mixed noise that is by default composed of 0.2% Gaussian noise and Poisson noise with an intensity of 2×10^7. To further validate the effectiveness of our approach, we reconstructed images with a resolution of 512×512 during testing to further evaluate its generalization capability in handling images with a larger number of pixel values.

2) Training Details: Our method is implemented using the PyTorch framework, and we use the differentiable operation $\mathcal{P}$ and $\mathcal{P}^{\mathrm{T}}$ operation in ODL[3] library on the NVIDIA RTX 3090Ti GPU. The ADAM algorithm is employed as the optimizer with $(\beta_1,\beta_2)=(0.5,0.999)$ whose initial learning rate is 1×10^{-4}. The number of training epochs is 100 with a batch size of 1. Heuristically, we manually set $\eta=2\times10^{-4}$, $\alpha=0.1$, $\beta=9$, and $\tau=0.8$ in our experiments. For training, we consider sparse view numbers of 60, 90, 120, and 180. Specifically, we train single-view models for each view separately. Moreover, the multi-view models are trained using a hybrid-view learning strategy, with the training data increasing from an initial training set of 1000 to 4000. It is worth noting that we did not add any additional original projection data to ensure fairness in experimental comparisons.

B. Compared with previous SVCT algorithms

To validate the effectiveness and generalization performance of our proposed method, we perform experiments under two different experimental setups: single-view models with separate training and multi-view models with mixed training. Specifically, we divide the proposed approaches into three versions for experiments: (*i*) separately trained single-view models, denoted as LipCT, where CGNet excludes the prompt module; (*ii*) mixed trained multi-view models, denoted as MLipCT, where CGNet excludes the prompt module; and (*iii*) mixed trained multi-view models, where CGNet includes the prompt module, denoted as PromptCT. It is worth noting that existing methods need to train the models separately for different sparse view configurations, whereas our proposed MLipCT and PromptCT only need to be trained once at the case of multiple views.

1) Quantitative results: We compare the proposed PromptCT with the current DL-based SVCT reconstruction methods including RED-CNN [45], FBPConvNet [22], DD-Net [46], CNNMAR [24], MSANet [47], DuDoTrans [23], and FreeSeed [7]. We employ the peak signal-to-noise ratio (PSNR), structured similarity index (SSIM), and root mean square error (RMSE) for perceived visual quality assessment. Table I reports the quantitative comparison of different SVCT reconstruction methods. Specifically, we

[3]https://github.com/odlgroup/odl

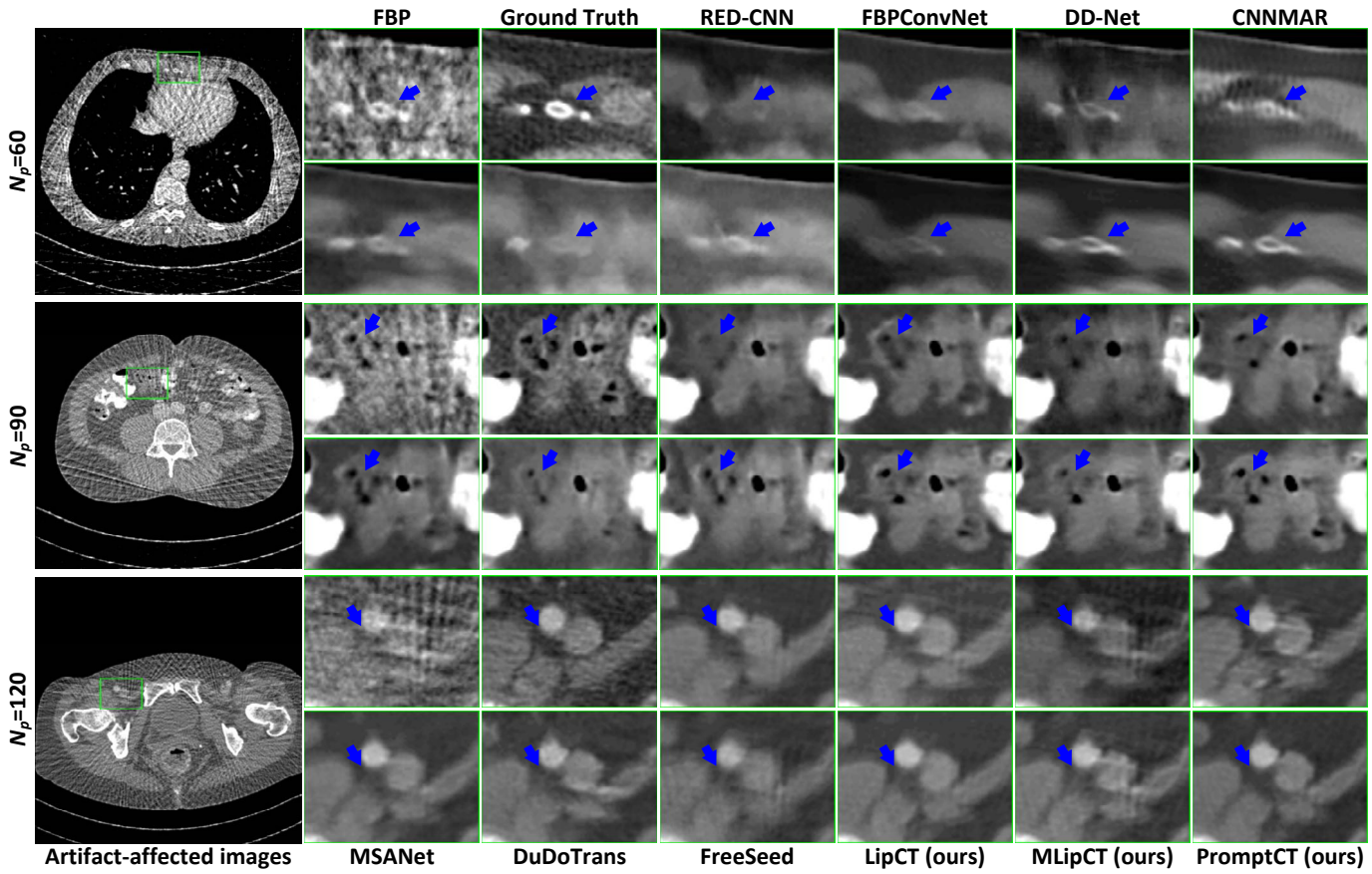


Fig. 4: Visual comparison of SVCT reconstruction methods on AAPM dataset under different sparse-view settings. The display window is [-175, 500] HU. Regions of interest are zoomed in for better viewing. Blue arrows indicate key areas, such as the sternum, intestinal tissues, and tissues near the femoral artery, which highlight differences in reconstruction quality among the methods. For single-view models, our LipCT outperforms other benchmark methods. For multi-view models, our PromptCT method, which incorporates explicit prompts, achieves better performance than that of MLipCT.

TABLE II: Storage requirement for different models. In comparison regarding the storage costs (MB) of models for four sparse views, the multiple-in-one model demonstrates better storage efficiency.

Method	RED-CNN	FBPConvNet	DD-Net	CNNMAR	MSANet	DuDoTrans	FreeSeed	LipCT (ours)	MLipCT (ours)	PromptCT (ours)
Storage	84.7	1582.8	26.8	125.1	396.1	1054.6	357.7	37.9	9.5	12.5

validated the effectiveness of the proposed method at multiple sampling views ($N_p = 60, 90, 120, 180$). From the results of the single-view models, it can be seen that our proposed deep unfolding approach (i.e., LipCT) outperforms the previous methods in most sampling settings and achieves the best average value, attributed to the superior representation ability of the proposed prior network. We observe that the DL-based dual-domain methods (i.e., CNNMAR) exhibit significant superiority over the image-domain methods due to the richness of information provided by the sinogram. However, as the number of sparse views decreases, the available information provided by the sinogram also decreases and the advantage of the dual-domain approach is no longer significant.

Single-view models require complicated training processes for each view as well as extensive storage costs, while the generalizability of single-view models is limited. Therefore, we employ the mixed training strategy to unlock multiple sparse-view settings using a single model. Consistent with our proposed single-view model architecture, we try to use mixed training instead of single-view training. The results indicate that without using prompts, training with multi-view data concurrently may adversely affect the learning process and lead to suboptimal performance. In contrast, our proposed PromptCT achieves higher PSNR/SSIM values and lower RMSE values compared to the single-view models, suggesting that the non-trivial transferability of PromptCT stems from the prompt-based model rather than the training strategy.

2) Qualitative results: Figure 4 shows representative sliced results for different methods in different sparse sampling views, where the region of interest is zoomed in to aid visualization. As shown by the blue arrows in Fig. 4, previous DL-based methods are successful in removing a large number of streak artifacts in some settings, but they introduce secondary artifacts, especially when the number of sparse views is small. These issues are especially pronounced in the regions near the

sternum, intestinal tissues, and tissues surrounding the femoral artery, where clear boundaries and fine textures are critical for diagnostic accuracy.

In contrast, our proposed LipCT effectively removes most streak artifacts while preserving the tissue structures of CT images. However, there is still room for improvement when the number of sampling angles is significantly reduced. For multi-view models, directly training LipCT with multi-view mixed data (i.e., MLipCT) results in suboptimal reconstruction performance, with noticeable degradation in reconstruction accuracy. However, our proposed PromptCT, by introducing explicit prompts, further improves the reconstruction quality. It not only eliminates artifacts more effectively but also preserves finer tissue details and structure integrity, even under highly sparse settings such as $N_p = 60$. For instance, in the blue-arrow-indicated regions, PromptCT demonstrates superior performance by reconstructing the contours of the sternum with high precision and preserving the subtle textures of intestinal tissues, which are often blurred by other methods. These regions clearly showcase the ability of PromptCT to deliver sharper and more accurate reconstructions compared to existing approaches, highlighting its effectiveness in addressing SVCT challenges.

TABLE III: Quantitative evaluation [PSNR (dB)↑/ SSIM↑/ RMSE↓] of different SVCT methods on the DeepLesion dataset at the sparse view number of 60. The best results are highlighted in red and the second-best results are highlighted in blue.

Method	PSNR	SSIM	RMSE
single-view models			
FBP	13.33	0.4466	0.2202
RED-CNN [45]	37.63	0.9426	0.0134
FBPConvNet [22]	37.63	0.9518	0.0134
DD-Net [46]	37.90	0.9463	0.0130
CNNMAR [24]	37.69	0.9566	0.0133
MSANet [47]	36.94	0.9405	0.0145
DuDoTrans [23]	37.87	0.9383	0.0130
FreeSeed [7]	37.71	0.9397	0.0132
LipCT (ours)	39.11	0.9581	0.0113
multi-view models			
MLipCT (ours)	35.38	0.9157	0.0173
PromptCT (ours)	39.17	0.9576	0.0112

3) Storage costs of models: In addition, we compute the storage costs of the models to comprehensively evaluate the performance of the algorithm. We provide the comparison of storage requirements in Tab. II. The single-view models create a large storage requirement since it requires saving a model for each view in Tab. I. In contrast, only one model of the multi-view models needs to be stored. Moreover, the proposed strategy achieves a good trade-off between performance improvement and saving storage costs.

4) Comparison on DeepLesion dataset: To further validate the generalizability of our method across different datasets, we conducted training and evaluation on the DeepLesion dataset. Notably, we trained these models using commonly used 416×416 images and reconstructed larger 512×512 images during testing to further evaluate the ability of these methods to handle high-resolution images. Table III presents a performance comparison of various SVCT reconstruction methods under the sparse-view condition with 60 sampling views. Among the single-view models, DL-based methods demonstrate significant improvements in reconstruction quality compared to traditional methods under this condition. However, their performance metrics are still suboptimal. Our proposed LipCT outperforms other single-view methods due to its ability to effectively integrate structural and contextual information through its tailored architecture. Furthermore, the multi-view model PromptCT, with explicit prompting, demonstrates superior performance compared to MLipCT, which does not incorporate prompts, under the sparse-view condition. This suggests that the remarkable transferability of PromptCT stems from its prompt-based design rather than the training strategy. As shown in Fig. 5, PromptCT achieves the best overall performance, owing to the incorporation of explicit prompts, which enhance the adaptability of the model to different sparse-view scenarios and enable it to preserve intricate details even in challenging regions. This underscores the effectiveness of our approach in improving generalization and reconstruction quality under sparse-view conditions.

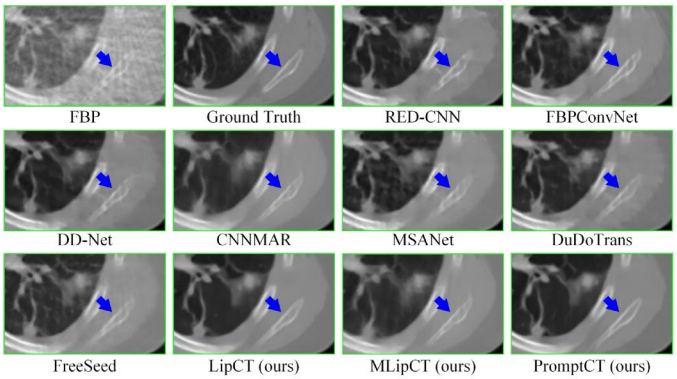


Fig. 5: Visual comparison of different SVCT methods on the DeepLesion dataset at the sparse view number of 60. Regions of interest are zoomed in for better viewing. Blue arrows indicate the bone structure and highlight the differences in reconstruction quality among the methods. For single-view models, our LipCT outperforms other benchmark methods. For multi-view models, our PromptCT method, which incorporates explicit prompts, achieves better performance than that of MLipCT.

5) Comparison on real experimental data: To validate the generalizability of our proposed methods, we conduct a comparison of SVCT reconstruction methods on real experimental data, analyzing both the equivalent water-bone phantom experiment and the pork with bone slices experiment. For the equivalent water-bone phantom experiment, we use an artificial synthetic material with attenuation coefficients that approximate those of water and bone, where the cylinder represents the equivalent bone, and the cube represents the equivalent water. Figure 6 shows the reconstruction results for the equivalent water and bone geometry phantom experiment. Data-driven methods, such as RED-CNN and FBPConvNet, reduce some artifacts but tend to excessively smooth the

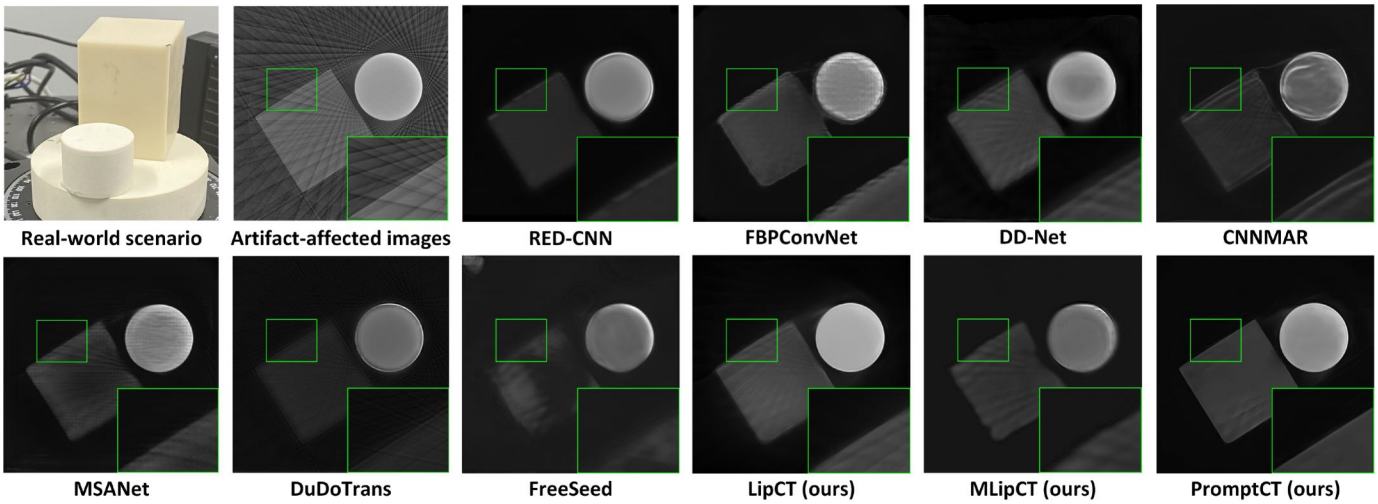


Fig. 6: Visual comparison of SVCT reconstruction methods on the equivalent water-bone phantom experiment at the sparse view number of 60. Regions of interest are zoomed in for better viewing. For single-view models, our LipCT outperforms other benchmark methods. For multi-view models, our PromptCT method, which incorporates explicit prompts, achieves better performance than that of MLipCT.

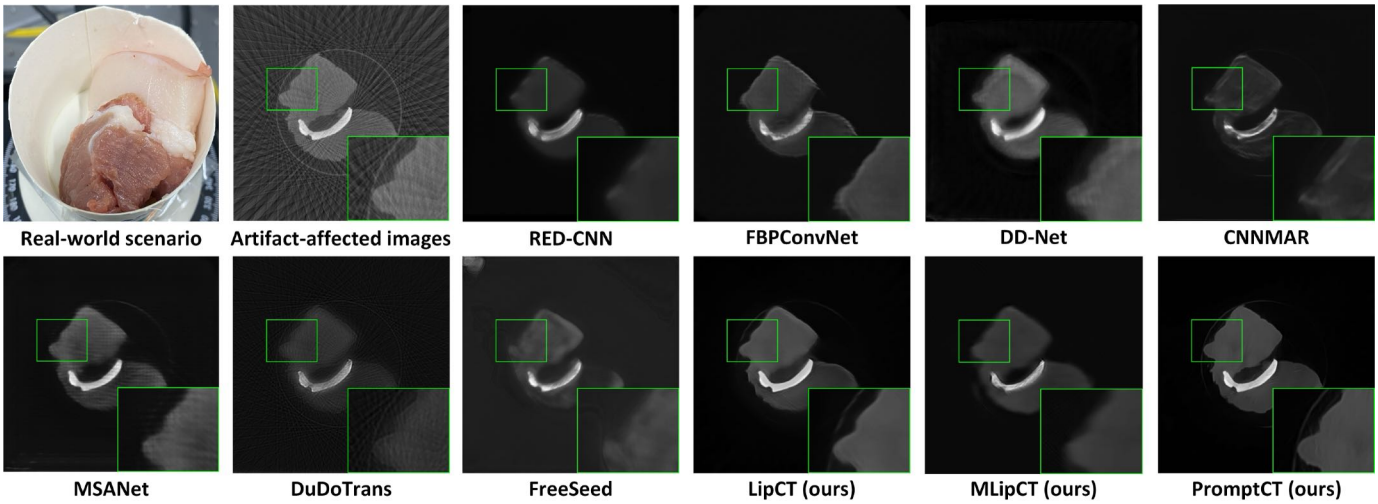


Fig. 7: Visual comparison of SVCT reconstruction methods on the pork with bone slices experiment at the sparse view number of 60. Regions of interest are zoomed in for better viewing. For single-view models, our LipCT outperforms other benchmark methods. For multi-view models, our PromptCT method, which incorporates explicit prompts, achieves better performance than that of MLipCT.

images, resulting in blurred details and loss of structures. Other DL-based methods, such as FreeSeed, still face challenges in preserving fine details, particularly at the edges of water-like tissue structures. In contrast, LipCT outperforms other single-view methods by effectively removing artifacts while preserving structural integrity, especially around water-tissue regions. For multi-view models, MLipCT struggles with artifact removal due to the lack of prompt information. In comparison, PromptCT demonstrates superior performance in handling equivalent water and bone structures, excelling in recovering edge information with great precision. For the pork with bone slice experiment, we scan real pork and insert a thin slice of bone within the pork. As shown in Fig. 7, comparisons of the reconstruction results on pork and bone slices further validate the effectiveness of our approach. LipCT performs better than other single-view methods, managing to preserve the complex boundaries between soft tissue and bone. PromptCT again leads in performance, delivering the most accurate reconstruction, particularly in the detailed regions of both bone and pork slices, demonstrating its effectiveness in preserving subtle tissue structures in challenging conditions.

C. Ablation studies

1) The effect of different network architectures in LipNet: We conduct a series of ablation experiments to analyze the effectiveness of network architectures in the proposed backbone of LipNet at $N_p = 120$. In the ablation experiments conducted for proposed LipNet, all other hyperparameters were kept constant, while maintaining the network structure as determined in its final form. The configurations are in five

groups: (*i*) Net-1: replacing LipNet with CNNs; (*ii*) Net-2: replacing our designed CGNet with a learnable parameter; (*iii*) Net-3: replacing LipNet with a tight frame where $\boldsymbol{W}$ is fixed; (*iv*) Net-4: replacing LipNet with a tight frame network containing tight constraint; (*v*) Net-5: our proposed LipNet, where $\boldsymbol{W}$ is adaptively learned without any tight constraints.

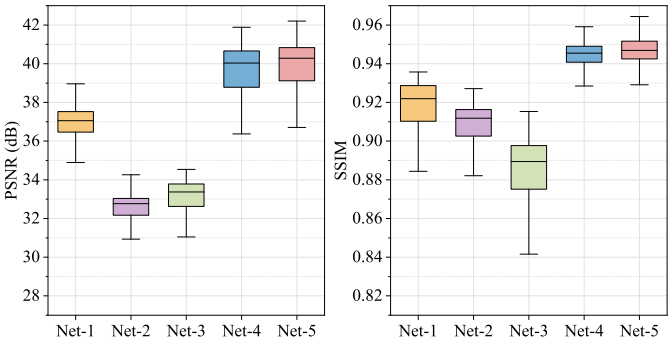


Fig. 8: Boxplots for PSNR(dB)/SSIM of different configurations at the sparse view number of 120. The left boxplot illustrates the PSNR values on the AAPM test dataset. The right boxplot displays the distribution of SSIM values.

Figure 8 presents the corresponding boxplots for PSNR/SSIM, where Net-5 is our baseline. When LipNet is replaced with simple CNNs in Net-1, the performance significantly degrades due to the insufficient representation ability provided by CNNs. Similarly, substituting CGNet with a simple learnable parameter in Net-2 results in inferior performance, highlighting the importance of carefully designing CGNet to enhance the representation ability of the sparse model. In Net-3, where $\boldsymbol{W}$ is fixed and unlearnable, the reconstruction quality suffers due to the limited adaptability of pre-defined sparsifying frames, which fail to capture data-specific structural nuances. In Net-4, introducing tight constraints on $\boldsymbol{W}$ restricts its flexibility, leading to suboptimal performance compared to the unconstrained $\boldsymbol{W}$ in Net-5. By contrast, Net-5 leverages the deep unfolding technique, allowing $\boldsymbol{W}$ to be learned end-to-end in a relaxed and adaptive manner. This enables LipNet to achieve superior results by maximizing the representation ability of sparsifying frames $\boldsymbol{W}$ while avoiding the restrictive nature of tight frame constraints.

2) The effect of explicit prompts: We conduct ablation experiments on explicit view prompts based on the PromptCT architecture. The configurations are as follows: (*i*) Net-6: removing view prompts from the default; (*ii*) Net-7: replacing explicit view prompts with adaptive prompts [33]; (*iii*) Net-8: replacing explicit view prompts with feature-learning prompts [25]; (*iv*) Net-9: default version with explicit prompts. The quantitative comparisons of different prompt learning methods are shown in Fig. 9. We observe that both adaptive prompts and feature-learning prompts are capable of perceiving different sparse view settings. However, due to the limited size of the training dataset, these pieces of information are not fully exploited for effectively distinguishing fine-grained settings, resulting in suboptimal performance. In contrast, the proposed explicit view prompts benefit our model by injecting downsampling sinogram masks to distinguish sparse sampling information.

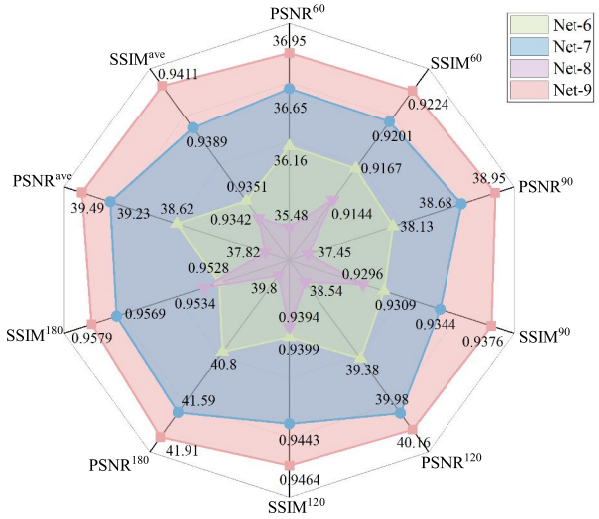


Fig. 9: Comparison of PSNR (dB) and SSIM for different prompt learning on the AAPM dataset under various SVCT scanning settings. 60, 90, 120, and 180 are the sparse view numbers we choose and their average scores are calculated. In radar charts, the farther away from the center, the higher the value, indicating better SVCT reconstruction results.

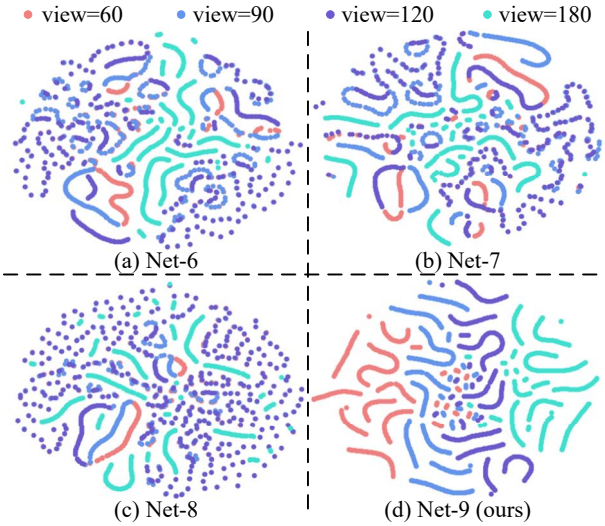


Fig. 10: Visualization of the features of prompt learning using t-SNE. The same object is coded by the same color. This indicates that most of features under explicit prompts are distinguishable, but the features become indistinct under adaptive prompts or feature-learning prompts conditions.

As illustrated in Fig. 10, we can see that the explicit view prompts are more capable of distinguishing specific sparse-view information than the adaptive prompts and the feature-learning prompts to encode the key discriminative information of artifact-affected CT images. The explicit prompt module can incorporate view prompts into feature information to distinguish different sparse sampling configurations, thereby providing decoupled properties for the multiple-in-one model and simultaneously enhancing reconstruction performance. Due to the inherent variability and adaptability of the learning process, explicit prompt-based learning methods often exhibit subtle feature overlaps, especially when sparse view angles

are numerically close. It is important to note that in prompt learning frameworks, minor feature overlaps or mixing are expected and acceptable, as they reflect the inherent flexibility of the model in capturing different patterns.

TABLE IV: Ablation study on the effect of STB/SFB in CGNet on AAPM dataset at the sparse view number of 120.

Network	STB	SFB	PSNR \ SSIM \ RMSE
Net-10	×	×	37.53 \ 0.9298 \ 0.0134
Net-11	✓	×	39.34 \ 0.9409 \ 0.0109
Net-12	×	✓	39.77 \ 0.9441 \ 0.0104
Net-13	✓	✓	40.10 \ 0.9466 \ 0.0100

3) The effect of STB/SFB in CGNet: To evaluate the individual contributions of STB and SFB within CGNet, we conduct ablation experiments, as summarized in Tab. IV. The results indicate that removing either STB or SFB significantly reduces the reconstruction quality, as these components play critical roles in the network. STB enhances global contextual modeling by capturing long-range dependencies, while SFB focuses on extracting fine-grained spatial frequency features essential for preserving structural details. Without these modules, the network struggles to suppress artifacts and maintain structural integrity, leading to poorer reconstruction performance. Combining both modules achieves the best results, showcasing their complementary strengths in addressing SVCT challenges.

D. Convergence and stability

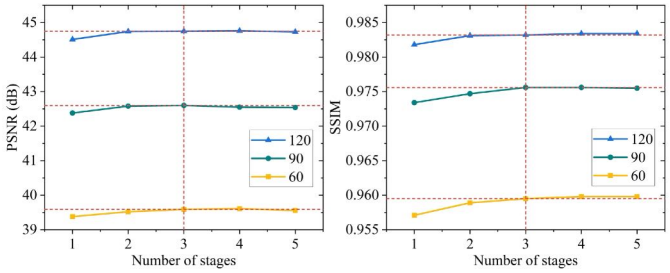


Fig. 11: Average PSNR (dB)↑/ SSIM↑ values for increasing network stages $K = 1, 2, 3, 4, 5$ in the inference process on the DeepLesion dataset under different sparse-view settings.

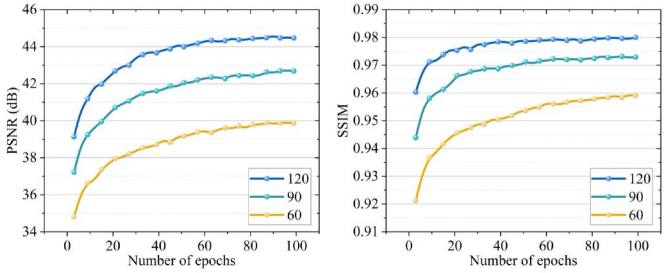


Fig. 12: Average PSNR (dB)↑/ SSIM↑ values for increasing epochs in the training process on the DeepLesion dataset under different sparse-view settings.

To experimentally validate the convergence property of our proposed iterative algorithm, we plot the reconstruction quality metrics (i.e., PSNR and SSIM) of reconstructed CT images at $N_p = 60, 90, 120$ for various numbers of stages (i.e., $K = 1, 2, 3, 4, 5$). As shown in Fig. 11, the proposed algorithm demonstrates stable and rapid convergence, with reconstruction accuracy steadily improving as the number of stages increases. Specifically, the average PSNR, SSIM, and RMSE values converge around $K = 3$, and the improvement in reconstruction performance gradually levels off as the number of stages increases. Theoretically, adding more stages enhances reconstruction quality by providing deeper iterative refinements. However, this improvement comes at the cost of increased computational time and complexity, resulting in lower computational efficiency. Considering the trade-offs between reconstruction quality and computational cost, we ultimately determined $K = 3$ to be the optimal number of stages for training. Furthermore, we provide a practical visualization of the training process. Figure 12 illustrates the reconstruction performance in terms of PSNR and SSIM as functions of the number of epochs under three different sparse-view conditions (i.e., $N_p = 60, 90, 120$). The results indicate that higher sparse sampling rates achieve better reconstruction quality, with PSNR and SSIM values steadily increasing and stabilizing as the number of epochs progresses, demonstrating the convergence of our method during the training phase.

E. Feasibility of 3D reconstruction

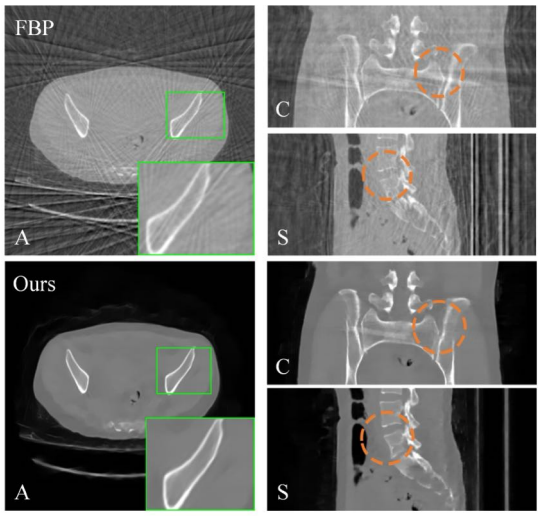


Fig. 13: The reconstruction results on the SpineWeb dataset at the sparse view number of 60. The 3D image volume is obtained by stacking all the axial slices together. A: Axial, C: Coronal, S: Sagittal. The green rectangles highlight regions of interest, which are zoomed in for better viewing. Orange circles indicate the hip and lumbar spine in the CT images.

To validate the extension of our method to the reconstruction of 3D image volumes, we utilize CT slices from a patient in the SpineWeb dataset at the sparse view number of 60, specifically focusing on vertebrae localization and identification. As shown in Fig. 13, we generate axial, coronal, and sagittal CT slices from the SpineWeb dataset and apply our proposed method to reconstruct clear and detailed images across these orientations.

The 3D image volume is constructed by stacking all the axial slices together. The analytical reconstruction algorithm, FBP, struggles with incomplete projection data, resulting in severe streak artifacts in the reconstructed slices. In contrast, our method demonstrates significant improvements, effectively removing artifacts and restoring subtle structures. The results indicate that our approach achieves stable reconstruction accuracy across all orientations, preserving clear details of the vertebrae and surrounding bones. This demonstrates the robustness of our method in producing high-quality reconstructions in multiple dimensions.

VI. Conclusion and Future work

In this work, we proposed a storage-saving deep unfolding framework for SVCT reconstruction, called PromptCT. Compared to existing models, we employed a single model to address SVCT reconstruction setups across various sparse sampling view numbers, thus significantly alleviating the storage cost of relevant medical institutions. On the theoretical side, we explicitly dropped the requirement for the closed-form solution and proposed a prompt-based sparse representation model-driven network satisfying Lipschitz constraint. On the experimental side, PromptCT achieved competitive or even better reconstruction performance at lower storage costs in multiple-in-one SVCT reconstruction compared with the benchmark methods, demonstrating the effectiveness and applicability of our proposed model.

In future work, we aim to address the challenges of GPU memory consumption and extend the proposed network to 3D reconstruction tasks by optimizing its computational efficiency and exploring lightweight architectures. Additionally, we will incorporate non-linear measurement models and validate the network on diverse real-world datasets, further bridging the gap between simulation and practical CT reconstruction applications. These efforts will enhance the scalability and adaptability of our network, making it suitable for broader clinical and industrial scenarios. Furthermore, we will use larger datasets and incorporate more clinically relevant evaluation metrics. These improvements will allow for a more robust and clinically applicable method for CT reconstruction.

References

[1] M. Bosch de Basea Gomez, I. Thierry-Chef, R. Harbron *et al.*, "Risk of hematological malignancies from CT radiation exposure in children, adolescents and young adults." *Nature Medicine*, vol. 29, no. 12, pp. 3111–3119, Dec 2023.

[2] B. S. Shi, K. Jiang, S. L. Zhang, Q. S. Lian, Y. W. Qin, and Y. S. Zhao, "Mud-Net: multi-domain deep unrolling network for simultaneous sparse-view and metal artifact reduction in computed tomography," *Machine Learning: Science and Technology*, vol. 5, p. 15010, Jan 2024.

[3] K. Xu, S. Y. Lu, B. Huang, W. W. Wu, and Q. G. Liu, "Stage-by-stage wavelet optimization refinement diffusion model for sparse-view CT reconstruction," *IEEE Transactions on Medical Imaging*, Jan 2024.

[4] T. Wang, W. J. Xia, J. F. Lu, and Y. Zhang, "A review of deep learning CT reconstruction from incomplete projection data," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 8, no. 2, pp. 138–152, Feb 2024.

[5] W. W. Wu, J. Y. Pan, Y. Y. Wang, S. Y. Wang, and J. J. Zhang, "Multi-channel optimization generative model for stable ultra-sparse-view CT reconstruction," *IEEE Transactions on Medical Imaging*, Mar 2024.

[6] Y. Han and J. C. Ye, "Framing U-Net via deep convolutional framelets: Application to sparse-view CT," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1418–1429, Apr 2018.

[7] C. L. Ma, Z. L. Li, J. P. Zhang, Y. Zhang, and H. M. Shan, "FreeSeed: Frequency-band-aware and self-guided network for sparse-view CT reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Oct 2023, pp. 250–259.

[8] Z. L. Li, C. L. Ma, J. Chen, J. P. Zhang, and H. M. Shan, "Learning to distill global representation for sparse-view CT," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2023, pp. 21 139–21 150.

[9] Y. K. Zhang, T. L. Lv, R. J. Ge, Q. L. Zhao *et al.*, "CD-Net: Comprehensive domain network with spectral complementary for DECT sparse-view reconstruction," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 436–447, Apr 2021.

[10] B. S. Shi, S. L. Zhang, and Z. R. Fu, "Artifact region-aware transformer: Global context helps CT metal artifact reduction," *IEEE Signal Processing Letters*, vol. 31, pp. 1249–1253, Apr 2024.

[11] B. S. Shi, Y. X. Wang, and D. Li, "Provable general bounded denoisers for snapshot compressive imaging with convergence guarantee," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 55–69, Feb 2023.

[12] B. S. Shi and K. X. Liu, "Regularization by multiple dual frames for compressed sensing magnetic resonance imaging with convergence analysis," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 11, pp. 2136–2153, Nov 2023.

[13] Y. B. Cheng, Q. Li, R. R. Li, T. Wang, J. J. Zhao *et al.*, "LIR-Net: Learnable iterative reconstruction network for fan beam CT sparse-view reconstruction," *IEEE Transactions on Computational Imaging*, vol. 10, pp. 181–195, Jan 2024.

[14] Y. Yang, Y. Z. Wang, J. Z. Wang, J. Sun, and Z. B. Xu, "An unrolled implicit regularization network for joint image and sensitivity estimation in parallel MR imaging with convergence guarantee," *SIAM Journal on Imaging Sciences*, vol. 16, pp. 1791–1824, Sep 2023.

[15] Y. A. Zhao, Y. L. Li, H. C. Zhang, V. Monga, and Y. C. Eldar, "Deep, convergent, unrolled half-quadratic splitting for image deconvolution," *IEEE Transactions on Computational Imaging*, vol. 10, pp. 574–588, Mar 2024.

[16] S. Mukherjee, A. Hauptmann, O. Oktem, M. Pereyra, and C.-B. Schonlieb, "Learned reconstruction methods with convergence guarantees: A survey of concepts and applications," *IEEE Signal Processing Magazine*, vol. 40, no. 1, pp. 164–182, Jan 2023.

[17] E. Ryu, J. L. Liu, S. C. Wang, X. H. Chen, Z. Y. Wang, and W. T. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, Jun 2019, pp. 5546–5557.

[18] A. M. Teodoro, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Scene-adapted plug-and-play algorithm with convergence guarantees," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep 2017, pp. 1–6.

[19] B. S. Shi, Y. X. Wang, and Q. S. Lian, "A trainable bounded denoiser using double tight frame network for snapshot compressive imaging," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 1516–1520.

[20] S. H. Chan, X. R. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 84–98, Mar 2017.

[21] W. W. Wu, Y. Y. Wang, Q. G. Liu, G. Wang, and J. J. Zhang, "Wavelet-improved score-based generative model for medical imaging," *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 966–979, Oct 2024.

[22] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, Sep 2017.

[23] C. Wang, K. Shang, H. M. Zhang, Q. Li, and S. K. Zhou, "DuDoTrans: Dual-domain transformer for sparse-view CT reconstruction," in *Machine Learning for Medical Image Reconstruction*, Sep 2022, pp. 84–94.

[24] M. D. Ketcha, M. Marrama, A. Souza, A. Uneri, P. W. Wu, X. X. Zhang, P. A. Helm, and J. H. Siewerdsen, "Sinogram + image domain neural network approach for metal artifact reduction in low-dose cone-beam computed tomography," *Journal of Medical Imaging*, vol. 8, no. 5, p. 052103, Mar 2021.

[25] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan, "PromptIR: Prompting for all-in-one image restoration," in *Advances in Neural Information Processing Systems*, vol. 36, Jun 2023, pp. 71 275–71 293.

[26] B. S. Shi, S. L. Zhang, K. Jiang, and Q. S. Lian, "Coupling model- and data-driven networks for CT metal artifact reduction," *IEEE Transactions on Computational Imaging*, vol. 10, pp. 415–428, Feb 2024.

[27] R. R. Li, Q. Li, H. X. Wang, S. Z. Li, J. J. Zhao, Q. Yan, and L. Wang, "DDPTransformer: Dual-domain with parallel transformer

network for sparse view CT image reconstruction," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 1101–1116, Sep 2022.

[28] W. W. Wu, D. L. Hu, C. Niu, H. Y. Yu, V. Vardhanabhuti, and G. Wang, "DRONE: Dual-domain residual-based optimization network for sparse-view CT reconstruction," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3002–3014, Nov 2021.

[29] C. Ding, Q. C. Zhang, G. Wang, X. J. Ye, and Y. M. Chen, "Learned alternating minimization algorithm for dual-domain sparse-view CT reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Oct 2023, pp. 173–183.

[30] W. J. Xia, Z. Y. Yang, Z. X. Lu, Z. X. Wang, and Y. Zhang, "RegFormer: A local–nonlocal regularization-based model for sparse-view CT reconstruction," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 8, no. 2, pp. 184–194, Feb 2024.

[31] H. Wang, M. H. Zhou, D. Wei, Y. X. Li, and Y. F. Zheng, "MEPNet: A model-driven equivariant proximal network for joint sparse-view reconstruction and metal artifact reduction in CT images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Oct 2023, pp. 181–195.

[32] H. M. Zhang, B. D. Liu, H. Y. Yu, and B. Dong, "MetaInv-Net: Meta inversion network for sparse view CT image reconstruction," *Journal of Medical Imaging*, vol. 40, no. 2, pp. 621–634, Feb 2021.

[33] B. Y. Li, X. Liu, P. Hu, Z. Q. Wu, J. C. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022, pp. 17 431–17 441.

[34] H. Gao, J. Yang, Y. Zhang, N. Wang, J. Yang, and D. Dang, "Prompt-based ingredient-oriented all-in-one image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9458–9471, May 2024.

[35] J. Zhang, J. Huang, M. Yao, Z. Yang, H. Yu, M. Zhou, and F. Zhao, "Ingredient-oriented multi-degradation learning for image restoration," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug 2023, pp. 5825–5835.

[36] D. Park, B. H. Lee, and S. Y. Chun, "All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug 2023, pp. 5815–5824.

[37] Y. Zhu, T. Wang, X. Fu, X. Yang, X. Guo, J. Dai, Y. Qiao, and X. Hu, "Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug 2023, pp. 21 747–21 758.

[38] B. S. Shi, D. Li, Y. X. Wang, Y. M. Su, and Q. S. Lian, "Provable deep video denoiser using spatial-temporal information for video snapshot compressive imaging: Algorithm and convergence analysis," *Signal Processing*, vol. 214, p. 109236, Jan 2024.

[39] B. S. Shi and Q. S. Lian, "DualPRNet: Deep shrinkage dual frame network for deep unrolled phase retrieval," *IEEE Signal Processing Letters*, vol. 29, pp. 1177–1181, Apr 2022.

[40] J. Portilla, "Image restoration through l_0 analysis-based sparse optimization in tight frames," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Feb 2009, pp. 3909–3912.

[41] J.-F. Cai, H. Ji, Z. Shen, and G.-B. Ye, "Data-driven tight frame construction and image denoising," *Applied and Computational Harmonic Analysis*, vol. 37, no. 1, pp. 89–105, Jul 2014.

[42] J.-F. Cai, J. K. Choi, and K. Wei, "Data driven tight frame for compressed sensing MRI reconstruction via off-the-grid regularization," *SIAM Journal on Imaging Sciences*, vol. 13, no. 3, pp. 1272–1301, Nov 2020.

[43] X. Chen, W. J. Xia, Z. Y. Yang, H. Chen, Y. Liu, J. L. Zhou, Z. Wang, Y. Chen, B. H. Wen, and Y. Zhang, "SOUL-Net: A sparse and low-rank unrolling network for spectral CT image reconstruction," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, Oct 2023.

[44] Z. Y. Zha, B. H. Wen, X. Yuan, J. T. Zhou, C. Zhu, and A. C. Kot, "Low-rankness guided group sparse representation for image restoration," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7593–7607, Feb 2023.

[45] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. X. Liao, J. L. Zhou, and G. Wang, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2524–2535, Dec 2017.

[46] Z. C. Zhang, X. K. Liang, X. Dong, Y. Q. Xie, and G. H. Cao, "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, Jun 2018.

[47] Y. B. Gou, P. Hu, J. C. Lv, J. T. Y. Zhou, and X. Peng, "Multi-scale adaptive network for single image denoising," in *Neural Information Processing Systems*, vol. 35, Dec 2022, pp. 14 099–14 112.

[48] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1123–1133, Oct 2021.

[49] C. McCollough, A. Bartley, R. Carter, B. Chen, T. Drees, P. Edwards, D. Holmes, A. Huang, F. Khan, S. Leng, K. McMillan, G. Michalak, K. Nunez, L. Yu, and J. Fletcher, "Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 low dose CT grand challenge," *Medical physics*, vol. 44, no. 10, pp. e339–e352, Oct 2017.

[50] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging*, vol. 5, no. 3, p. 036501, July 2018.

[51] X. Xue, S. Zhao, Y. Zhao, and P. Zhang, "Image reconstruction for limited-angle computed tomography with curvature constraint," *Measurement Science and Technology*, vol. 30, p. 125401, Sep 2019.

[52] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine ct via dense classification from sparse annotations," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, 2013, pp. 262–270.