ASCON: Anatomy-Aware Supervised Contrastive Learning Framework for Low-Dose CT Denoising

Zhihao Chen[1], Qi Gao[1], Yi Zhang[2], and Hongming Shan[1,3,4](✉)

[1] Institute of Science and Technology for Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China
hmshan@fudan.edu.cn

[2] School of Cyber Science and Engineering, Sichuan University, Chengdu, China

[3] Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China

[4] Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

Abstract. While various deep learning methods have been proposed for low-dose computed tomography (CT) denoising, most of them leverage the normal-dose CT images as the ground-truth to supervise the denoising process. These methods typically ignore the inherent correlation within a single CT image, especially the anatomical semantics of human tissues, and lack the interpretability on the denoising process. In this paper, we propose a novel **A**natomy-aware **S**upervised **CON**trastive learning framework, termed ASCON, which can explore the anatomical semantics for low-dose CT denoising while providing anatomical interpretability. The proposed ASCON consists of two novel designs: an efficient self-attention-based U-Net (ESAU-Net) and a multi-scale anatomical contrastive network (MAC-Net). First, to better capture global-local interactions and adapt to the high-resolution input, an efficient ESAU-Net is introduced by using a channel-wise self-attention mechanism. Second, MAC-Net incorporates a patch-wise non-contrastive module to capture inherent anatomical information and a pixel-wise contrastive module to maintain intrinsic anatomical consistency. Extensive experimental results on two public low-dose CT denoising datasets demonstrate superior performance of ASCON over state-of-the-art models. Remarkably, our ASCON provides anatomical interpretability for low-dose CT denoising for the first time. Source code is available at https://github.com/hao1635/ASCON.

Keywords: CT denoising · Deep learning · Self-attention · Contrastive learning · Anatomical semantics

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43999-5_34.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 355–365, 2023.
https://doi.org/10.1007/978-3-031-43999-5_34

1 Introduction

With the success of deep learning in the field of computer vision and image processing, many deep learning-based methods have been proposed and achieved promising results in low-dose CT (LDCT) denoising [1,4–6,9,12,23,24,26]. Typically, they employ a supervised learning setting, which involves a set of image pairs, LDCT images and their normal-dose CT (NDCT) counterparts. These methods typically use a pixel-level loss (*e.g.* mean squared error or MSE), which can cause over-smoothing problems.

To address this issue, a few studies [23,26] used a structural similarity (SSIM) loss or a perceptual loss [11]. However, they all perform in a sample-to-sample manner and ignore the inherent anatomical semantics, which could blur details in areas with low noise levels. Previous studies have shown that the level of noise in CT images varies depending on the type of tissues [17]; see an example in Fig. S1 in Supplementary Materials. Therefore, it is crucial to characterize the anatomical semantics for effectively denoising diverse tissues.

In this paper, we focus on taking advantage of the inherent anatomical semantics in LDCT denoising from a contrastive learning perspective [7,25,27]. To this end, we propose a novel **A**natomy-aware **S**upervised **CON**trastive learning framework (ASCON), which consists of two novel designs: an efficient self-attention-based U-Net (ESAU-Net) and a multi-scale anatomical contrastive network (MAC-Net). First, to ensure that MAC-Net can effectively extract anatomical information, diverse global-local contexts and a larger input size are necessary. However, operations on full-size CT images with self-attention are computationally unachievable due to potential GPU memory limitations [20]. To address this limitation, we propose an ESAU-Net that utilizes a channel-wise self-attention mechanism [2,22,28] which can efficiently capture both local and global contexts by computing cross-covariance across feature channels.

Second, to exploit inherent anatomical semantics, we present the MAC-Net that employs a disentangled U-shaped architecture [25] to produce global and local representations. Globally, a patch-wise non-contrastive module is designed to select neighboring patches with similar semantic context as positive samples and align the same patches selected in denoised CT and NDCT which share the same anatomical information, using an optimization method similar to the BYOL method [7]. This is motivated by the prior knowledge that adjacent patches often share common semantic contexts [27]. Locally, to further improve the anatomical consistency between denoised CT and NDCT, we introduce a pixel-wise contrastive module with a hard negative sampling strategy [21], which randomly selects negative samples from the pixels with high similarity around the positive sample within a certain distance. Then we use a local InfoNCE loss [18] to pull the positive pairs and push the negative pairs.

Our contributions are summarized as follows. 1) We propose a novel ASCON framework to explore inherent anatomical information in LDCT denoising, which is important to provide interpretability for LDCT denoising. 2) To better explore anatomical semantics in MAC-Net, we design an ESAU-Net, which utilizes a channel-wise self-attention mechanism to capture both local and global contexts.

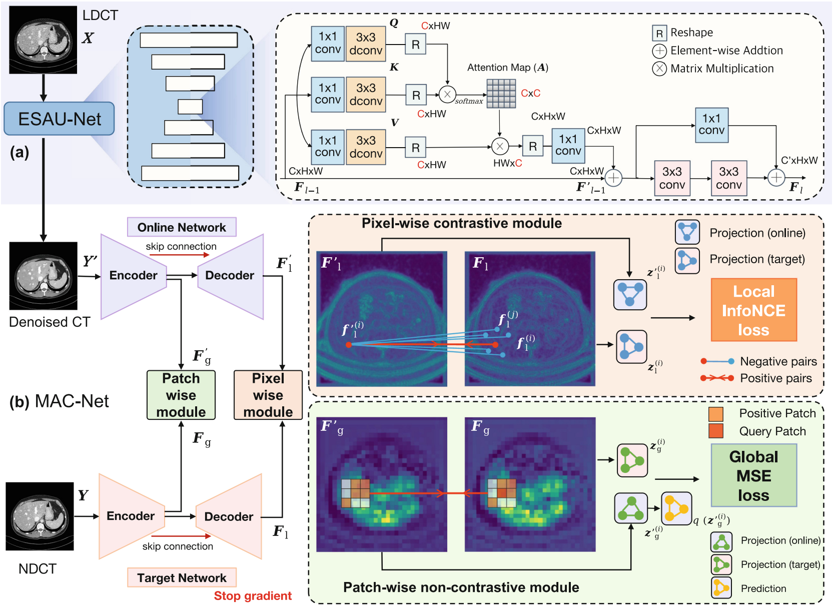

Fig. 1. Overview of our proposed ASCON. (a) Efficient self-attention-based U-Net (ESAU-Net); and (b) multi-scale anatomical contrastive network (MAC-Net).

3) We propose a MAC-Net that employs a disentangled U-shaped architecture and incorporates both global non-contrastive and local contrastive modules. This enables the exploitation of inherent anatomical semantics at the patch level, as well as improving anatomical consistency at the pixel level. 4) Extensive experimental results demonstrate that our ASCON outperforms other state-of-the-art methods, and provides anatomical interpretability for LDCT denoising.

2 Methodology

2.1 Overview of the Proposed ASCON

Figure 1 presents the overview of the proposed ASCON, which consists of two novel components: ESAU-Net and MAC-Net. First, given an LDCT image, $\boldsymbol{X} \in \mathbb{R}^{1\times H\times W}$, where $H \times W$ denotes the image size. $\boldsymbol{X}$ is passed through the ESAU-Net to capture both global and local contexts using a channel-wise self-attention mechanism and obtain a denoised CT image $\boldsymbol{Y}' \in \mathbb{R}^{1\times H\times W}$.

Then, to explore inherent anatomical semantics and remain inherent anatomical consistency, the denoised CT $\boldsymbol{Y}'$ and NDCT $\boldsymbol{Y}$ are passed to the MAC-Net to compute a global MSE loss $\mathcal{L}_{\text{global}}$ in a patch-wise non-contrastive module and a local infoNCE loss $\mathcal{L}_{\text{local}}$ in a pixel-wise contrastive module. During training, we use an alternate learning strategy to optimize ESAU-Net and MAC-Net separately, which is similar to GAN-based methods [10]. Please refer to Algorithm S1 in Supplementary Materials for a detailed optimization.

2.2 Efficient Self-attention-Based U-Net

To better leverage anatomical semantic information in MAC-Net and adapt to the high-resolution input, we design the ESAU-Net that can capture both local and global contexts during denoising. Different from previous works that only use self-attention in the coarsest level [20], we incorporate a channel-wise self-attention mechanism [2,28] at each up-sampling and down-sampling level in the U-Net [22] and add an identity mapping in each level, as shown in Fig. 1(a).

Specifically, in each level, given the feature map $\boldsymbol{F}_{l-1}$ as the input, we first apply a 1×1 convolution and a 3×3 depth-wise convolution to aggregate channel-wise contents and generate query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) followed by a reshape operation, where $\boldsymbol{Q} \in \mathbb{R}^{C\times HW}$, $\boldsymbol{K} \in \mathbb{R}^{C\times HW}$, and $\boldsymbol{V} \in \mathbb{R}^{C\times HW}$ (see Fig. 1(a)). Then, a channel-wise attention map $\boldsymbol{A} \in \mathbb{R}^{C\times C}$ is generated through a dot-product operation by the reshaped query and key, which is more efficient than the regular attention map of size $HW \times HW$ [3], especially for high-resolution input. Overall, the process is defined as

$$\text{Attention}(\boldsymbol{F}) = w(\boldsymbol{V}^T\boldsymbol{A}) = w(\boldsymbol{V}^T \cdot \text{Softmax}(\boldsymbol{K}\boldsymbol{Q}^T/\alpha)), \tag{1}$$

where $w(\cdot)$ first reshapes the matrix back to the original size $C\times H\times W$ and then performs 1×1 convolution; α is a learnable parameter to scale the magnitude of the dot product of $\boldsymbol{K}$ and $\boldsymbol{Q}$. We use multi-head attention similar to the standard multi-head self-attention mechanism [3]. The output of the channel-wise self-attention is represented as: $\boldsymbol{F}'_{l-1} = \text{Attention}(\boldsymbol{F}_{l-1}) + \boldsymbol{F}_{l-1}$. Finally, the output $\boldsymbol{F}_l$ of each level is defined as: $\boldsymbol{F}_l = \text{Conv}(\boldsymbol{F}'_{l-1}) + \text{Iden}(\boldsymbol{F}_{l-1})$, where Conv($\cdot$) is a two-layer convolution and Iden($\cdot$) is an identity mapping using a 1×1 convolution; refer to Fig. S2(a) for the details of ESAU-Net.

2.3 Multi-scale Anatomical Contrastive Network

Overview of MAC-Net. The goal of our MAC-Net is to exploit anatomical semantics and maintain anatomical embedding consistency, First, a disentangled U-shaped architecture [22] is utilized to learn global representation $\boldsymbol{F}_\text{g}$ $\in \mathbb{R}^{512\times\frac{H}{16}\times\frac{W}{16}}$ after four down-sampling layers, and learn local representation $\boldsymbol{F}_\text{l}\in \mathbb{R}^{64\times H\times W}$ by removing the last output layer. And we cut the connection between the coarsest feature and its upper level to make $\boldsymbol{F}_\text{g}$ and $\boldsymbol{F}_\text{l}$ more independent [25] (see Fig. S2(b)). The online network and the target network, both using the same architecture above, handle denoised CT $\boldsymbol{Y}'$ and NDCT $\boldsymbol{Y}$, respectively, with $\boldsymbol{F}'_\text{g}$ and $\boldsymbol{F}'_\text{l}$ generated by the online network, and $\boldsymbol{F}_\text{g}$ and $\boldsymbol{F}_\text{l}$ generated by the target network (see Fig. 1(b)). The parameters of the target network are an exponential moving average of the parameters in the online network, following the previous works [7,8]. Next, a patch-wise non-contrastive module uses $\boldsymbol{F}_\text{g}$ and $\boldsymbol{F}'_\text{g}$ to compute a global MSE loss $\mathcal{L}_\text{global}$, while a pixel-wise contrastive module uses $\boldsymbol{F}_\text{l}$ and $\boldsymbol{F}'_\text{l}$ to compute a local infoNCE loss $\mathcal{L}_\text{local}$. Let us describe these two loss functions specifically.

Patch-Wise Non-contrastive Module. To better learn anatomical representations, we introduce a patch-wise non-contrastive module, also shown in Fig. 1(b). Specifically, for each pixel $\boldsymbol{f}_{\mathrm{g}}^{(i)} \in \mathbb{R}^{512}$ in the $\boldsymbol{F}_{\mathrm{g}}$ where $i \in \{1, 2, \ldots, \frac{HW}{256}\}$ is the index of the pixel location, it can be considered as a patch due to the expanded receptive field achieved through a sequence of convolutions and down-sampling operations [19]. To identify positive patch indices, we adopt a neighboring positive matching strategy [27], assuming that a semantically similar patch $\boldsymbol{f}_{\mathrm{g}}^{(j)}$ exists in the vicinity of the query patch $\boldsymbol{f}_{\mathrm{g}}^{(i)}$, as neighboring patches often share a semantic context with the query. We empirically consider a set of 8 neighboring patches. To sample patches with similar semantics around the query patch $\boldsymbol{f}_{\mathrm{g}}^{(i)}$, we measure the semantic closeness between the query patch $\boldsymbol{f}_{\mathrm{g}}^{(i)}$ and its neighboring patches $\boldsymbol{f}_{\mathrm{g}}^{(j)}$ using the cosine similarity, which is formulated as

$$s(i,j) = (\boldsymbol{f}_{\mathrm{g}}^{(i)})^{\top}(\boldsymbol{f}_{\mathrm{g}}^{(j)}) / \|\boldsymbol{f}_{\mathrm{g}}^{(i)}\|_2 \|\boldsymbol{f}_{\mathrm{g}}^{(j)}\|_2. \tag{2}$$

We then select the top-4 positive patches $\{\boldsymbol{f}_{\mathrm{g}}^{(j)}\}_{j \in \mathcal{P}^{(i)}}$ based on $s(i,j)$, where $\mathcal{P}^{(i)}$ is a set of selected patches (*i.e.*, $|\mathcal{P}^{(i)}| = 4$). To obtain patch-level features $\boldsymbol{g}^{(i)} \in \mathbb{R}^{512}$ for each patch $\boldsymbol{f}_{\mathrm{g}}^{(i)}$ and its positive neighbors, we aggregate their features using global average pooling (GAP) in the patch dimension. For the local representation of $\boldsymbol{f'}_{\mathrm{g}}^{(i)}$, we select positive patches as same as $\mathcal{P}^{(i)}$, *i.e.*, $\{\boldsymbol{f'}_{\mathrm{g}}^{(j)}\}_{j \in \mathcal{P}^{(i)}}$. Formally,

$$\boldsymbol{g}^{(i)} := \mathrm{GAP}(\boldsymbol{f}_{\mathrm{g}}^{(i)}, \{\boldsymbol{f}_{\mathrm{g}}^{(j)}\}_{j \in \mathcal{P}^{(i)}}), \quad \boldsymbol{g'}^{(i)} := \mathrm{GAP}(\boldsymbol{f'}_{\mathrm{g}}^{(i)}, \{\boldsymbol{f'}_{\mathrm{g}}^{(j)}\}_{j \in \mathcal{P}^{(i)}}). \tag{3}$$

From the patch-level features, the online network outputs a projection $\boldsymbol{z}_{\mathrm{g}}'^{(i)} = p_{\mathrm{g}}'(\boldsymbol{g'}^{(i)})$ and a prediction $q'(\boldsymbol{z}_{\mathrm{g}}'^{(i)})$ while target network outputs the target projection $\boldsymbol{z}_{\mathrm{g}}{}^{(i)} = p_{\mathrm{g}}(\boldsymbol{g}^{(i)})$. The projection and prediction are both multilayer perceptron (MLP). Finally, we compute the global MSE loss between the normalized prediction and target projection [7],

$$\mathcal{L}_{\mathrm{global}} = \sum\nolimits_{i \in \mathcal{N}_{\mathrm{pos}}^{\mathrm{g}}} \left\| q'(\boldsymbol{z}_{\mathrm{g}}'^{(i)}) - \boldsymbol{z}_{\mathrm{g}}{}^{(i)} \right\|_2^2 = \sum\nolimits_{i \in \mathcal{N}_{\mathrm{pos}}^{\mathrm{g}}} 2 - 2 \cdot \frac{\langle q'(\boldsymbol{z}_{\mathrm{g}}'^{(i)}), \boldsymbol{z}_{\mathrm{g}}{}^{(i)} \rangle}{\|q'(\boldsymbol{z}_{\mathrm{g}}'^{(i)})\|_2 \cdot \|\boldsymbol{z}_{\mathrm{g}}{}^{(i)}\|_2}, \tag{4}$$

where $\mathcal{N}_{\mathrm{pos}}^{\mathrm{g}}$ is the indices set of positive samples in the patch-level embedding.

Pixel-Wise Contrastive Module. In this module, we aim to improve anatomical consistency between the denoised CT and NDCT using a local InfoNCE loss [18] (see Fig. 1(b)). First, for a query $\boldsymbol{f}_{\mathrm{l}}'^{(i)} \in \mathbb{R}^{64}$ in the $\boldsymbol{F}_{\mathrm{l}}'$ and its positive sample $\boldsymbol{f}_{\mathrm{l}}^{(i)} \in \mathbb{R}^{64}$ in the $\boldsymbol{F}_{\mathrm{l}}$ ($i \in \{1, 2, \ldots, HW\}$ is the location index), we use a hard negative sampling strategy [21] to select "diffcult" negative samples with high probability, which enforces the model to learn more from the fine-grained details. Specifically, candidate negative samples are randomly sampled from $\boldsymbol{F}_{\mathrm{l}}$ as long as their distance from $\boldsymbol{f}_{\mathrm{l}}^{(i)}$ is less than m pixels ($m = 7$). We also use cosine similarity in Eq. (2) to select a set of semantically closest pixels, *i.e.* $\{\boldsymbol{f}_{\mathrm{l}}^{(j)}\}_{j \in \mathcal{N}_{\mathrm{neg}}^{(i)}}$. Then we concatenate $\boldsymbol{f'}_{\mathrm{l}}^{(i)}$, $\boldsymbol{f}_{\mathrm{l}}^{(i)}$, and $\{\boldsymbol{f}_{\mathrm{l}}^{(j)}\}_{j \in \mathcal{N}_{\mathrm{neg}}^{(i)}}$ and map them to a K-dimensional vector (K=256) through a two-layer MLP, obtaining $\boldsymbol{z}^{(i)}_{\text{l}} \in \mathbb{R}^{(2+|\mathcal{N}^{(i)}_{\text{neg}}|)\times 256}$. The local InfoNCE loss in the pixel level is defined as

$$\mathcal{L}_{\text{local}} = -\sum\nolimits_{i\in\mathcal{N}^{\text{l}}_{\text{pos}}} \log \frac{\exp\left(\boldsymbol{v'}^{(i)}_{\text{l}}\cdot\boldsymbol{v}^{(i)}_{\text{l}}/\tau\right)}{\exp\left(\boldsymbol{v'}^{(i)}_{\text{l}}\cdot\boldsymbol{v}^{(i)}_{\text{l}}/\tau\right)+\sum_{j=1}^{|\mathcal{N}^{(i)}_{\text{neg}}|}\exp\left(\boldsymbol{v'}^{(i)}_{\text{l}}\cdot\boldsymbol{v}^{(j)}_{\text{l}}/\tau\right)}, \tag{5}$$

where $\mathcal{N}^{\text{l}}_{\text{pos}}$ is the indices set of positive samples in the pixel level. $\boldsymbol{v'}^{(i)}_{\text{l}}$, $\boldsymbol{v}_{\text{l}}{}^{(i)}$, and $\boldsymbol{v}^{(j)}_{\text{l}} \in \mathbb{R}^{256}$ are the query, positive, and negative sample in $\boldsymbol{z}^{(i)}_{\text{l}}$, respectively.

2.4 Total Loss Function

The final loss is defined as $\mathcal{L} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{local}} + \lambda\mathcal{L}_{\text{pixel}}$, where $\mathcal{L}_{\text{pixel}}$ consists of two common supervised losses: MSE and SSIM, defined as $\mathcal{L}_{\text{pixel}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SSIM}}$. λ is empirically set to 10.

3 Experiments

3.1 Dataset and Implementation Details

We use two publicly available low-dose CT datasets released by the NIH AAPM-Mayo Clinic Low-Dose CT Grand Challenge in 2016 [15] and lately released in 2020 [16], denoted as Mayo-2016 and Mayo-2020, respectively. There is no overlap between the two datasets. Mayo-2016 includes normal-dose abdominal CT images of 10 anonymous patients and corresponding simulated quarter-dose CT images. Mayo-2020 provides the abdominal CT image data of 100 patients with 25% of the normal dose, and we randomly choose 20 patients for our experiments.

For the Mayo-2016, we choose 4800 pairs of 512 × 512 images from 8 patients for the training and 1136 pairs from the rest 2 patients as the test set. For the Mayo-2020, we employ 9600 image pairs with a size of 256 × 256 from randomly selected 16 patients for training and 580 pairs of 512 × 512 images from rest 4 patients for testing. The use of large-size training is to adapt our MAC-Net to exploit inherent semantic information. The default sampling hyper-parameters for Mayo-2016 are $|\mathcal{N}^{\text{l}}_{\text{pos}}| = 32$, $|\mathcal{N}^{\text{g}}_{\text{pos}}| = 512$, $|\mathcal{N}^{(i)}_{\text{neg}}| = 24$, while $|\mathcal{N}^{\text{l}}_{\text{pos}}| = 16$, $|\mathcal{N}^{\text{g}}_{\text{pos}}| = 256$, $|\mathcal{N}^{(i)}_{\text{neg}}| = 24$ for Mayo-2020. We use a binary function to filter the background while selecting queries in MAC-Net. For the training strategy, we employ a window of [−1000, 2000] HU. We train our network for 100 epochs on 2 NVIDIA GeForce RTX 3090, and use the AdamW optimizer [14] with the momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$ and the weight decay of 1.0×10^{-9}. We initialize the learning rate as 1.0×10^{-4}, gradually reduced to 1.0×10^{-6} with the cosine annealing [13]. Since MAC-Net is only implemented during training, the testing time of ASCON is close to most of the compared methods.

3.2 Performance Comparisons

Quantitative Evaluations. We use three widely-used metrics including peak signal-to-noise ratio (PSNR), root-mean-square error (RMSE), and SSIM. Table 1 presents the testing results on Mayo-2016 and Mayo-2020 datasets. We compare our methods with 5 state-of-the-art methods, including RED-CNN [1], WGAN-VGG [26], EDCNN [12], DU-GAN [9], and CNCL [6]. Table 1 shows that our ESAU-Net with MAC-Net achieves the best performance on both the Mayo-2016 and the Mayo-2020 datasets. Compared to the ESAU-Net, ASCON further improves the PSNR by up to 0.54 dB on Mayo-2020, which demonstrates the effectiveness of the proposed MAC-Net and the importance of the inherent anatomical semantics during CT denoising. We also compute the contrast-to-noise ratio (CNR) to assess the detectability of a selected area of low-contrast lesion and our ASCON achieves the best CNR in Fig. S3.

Table 1. Performance comparison on the Mayo-2016 and Mayo-2020 datasets in terms of PSNR [dB], RMSE [$\times10^{-2}$], and SSIM [%]

Methods	Mayo-2016			Mayo-2020		
	PSNR↑	RMSE↓	SSIM↑	PSNR↑	RMSE↓	SSIM↑
U-Net [22]	$44.13_{\pm1.19}$	$0.64_{\pm0.12}$	$97.38_{\pm1.09}$	$47.67_{\pm1.64}$	$0.43_{\pm0.09}$	$99.19_{\pm0.23}$
RED-CNN [1]	$44.23_{\pm1.26}$	$0.62_{\pm0.09}$	$97.34_{\pm0.86}$	$48.05_{\pm2.14}$	$0.41_{\pm0.11}$	$99.28_{\pm0.18}$
WGAN-VGG [26]	$42.49_{\pm1.28}$	$0.76_{\pm0.12}$	$96.16_{\pm1.30}$	$46.88_{\pm1.81}$	$0.46_{\pm0.10}$	$98.15_{\pm0.20}$
EDCNN [12]	$43.14_{\pm1.27}$	$0.70_{\pm0.11}$	$96.45_{\pm1.36}$	$47.90_{\pm1.27}$	$0.41_{\pm0.08}$	$99.14_{\pm0.17}$
DU-GAN [9]	$43.06_{\pm1.22}$	$0.71_{\pm0.10}$	$96.34_{\pm1.12}$	$47.21_{\pm1.52}$	$0.44_{\pm0.10}$	$99.00_{\pm0.21}$
CNCL [6]	$43.06_{\pm1.07}$	$0.71_{\pm0.10}$	$96.68_{\pm1.11}$	$45.63_{\pm1.34}$	$0.53_{\pm0.11}$	$98.92_{\pm0.59}$
ESAU-Net (ours)	$\underline{44.38}_{\pm1.26}$	$\underline{0.61}_{\pm0.09}$	$\underline{97.47}_{\pm0.87}$	$\underline{48.31}_{\pm1.87}$	$\underline{0.40}_{\pm0.12}$	$\underline{99.30}_{\pm0.18}$
ASCON (ours)	$\mathbf{44.48}_{\pm1.32}$	$\mathbf{0.60}_{\pm0.10}$	$\mathbf{97.49}_{\pm0.86}$	$\mathbf{48.84}_{\pm1.68}$	$\mathbf{0.37}_{\pm0.11}$	$\mathbf{99.32}_{\pm0.18}$

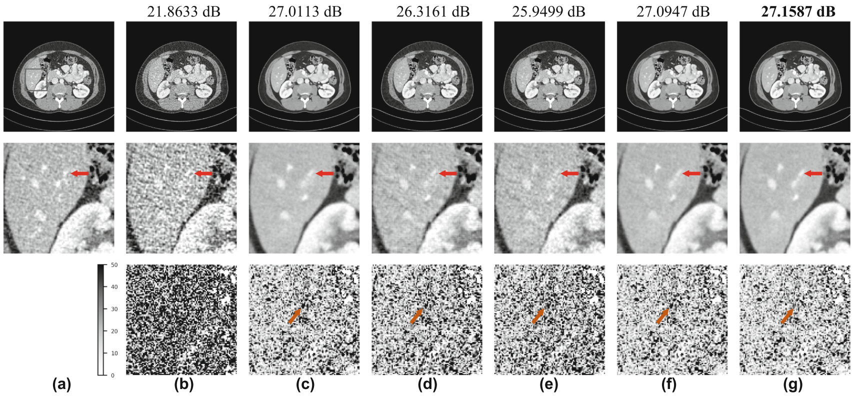

Fig. 2. Transverse CT images and corresponding difference images from the Mayo-2016 dataset: (a) NDCT; (b) LDCT; (c) RED-CNN [1]; (d) EDCNN [12]; (e) DU-GAN [9]; (f) ESAU-Net (ours); and (g) ASCON (ours). The display window is [−160, 240] HU.

Qualitative Evaluations. Figure 2 presents qualitative results of three representative methods and our ESAU-Net with MAC-Net on Mayo-2016. Although ASCON and RED-CNN produce visually similar results in low-contrast areas after denoising. However, RED-CNN results in blurred edges between different tissues, such as the liver and blood vessels, while ASCON smoothed the noise and maintained the sharp edges. They are marked by arrows in the regions-of-interest images. We further visualize the corresponding difference images between NDCT and the generated images by our method as well as other methods as shown in the third row of Fig. 2. Note that our ASCON removes more noise components than other methods; refer to Fig. S4 for extra qualitative results on Mayo-2020.

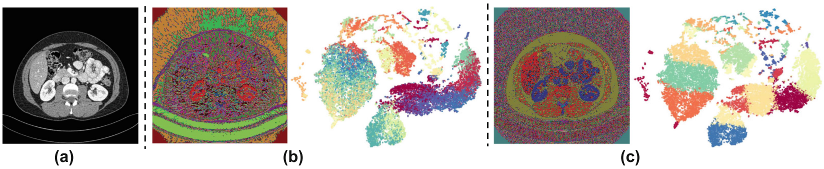

Fig. 3. Visualization of inherent semantics; (a) NDCT; (b) clustering and t-SNE results of ASCON w/o MAC-Net; and (c) clustering and t-SNE results of ASCON.

Table 2. Ablation results of Mayo-2020 on the different types of loss functions.

Loss	PSNR↑	RMSE↓	SSIM↑
$\mathcal{L}_{\text{MSE}}$	$48.34_{\pm 2.22}$	$0.40_{\pm 0.11}$	$99.27_{\pm 0.18}$
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{Perceptual}}$	$47.83_{\pm 1.99}$	$0.42_{\pm 0.10}$	$99.13_{\pm 0.19}$
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SSIM}}$	$48.31_{\pm 1.87}$	$0.40_{\pm 0.12}$	$99.30_{\pm 0.18}$
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{global}}$	$48.58_{\pm 2.12}$	$0.39_{\pm 0.10}$	$99.31_{\pm 0.17}$
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{local}}$	$48.48_{\pm 2.37}$	$0.38_{\pm 0.11}$	$99.31_{\pm 0.18}$
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{global}}$	$\mathbf{48.84}_{\pm 1.68}$	$\mathbf{0.37}_{\pm 0.11}$	$\mathbf{99.32}_{\pm 0.18}$

Visualization of Inherent Semantics. To demonstrate that our MAC-Net can exploit inherent anatomical semantics of CT images during denoising, we select the features before the last layer in ASCON without MAC-Net and ASCON from Mayo-2016. Then we cluster these two feature maps respectively using a K-means algorithm and visualize them in the original dimension, and finally visualize the clustering representations using t-SNE, as shown in Fig. 3. Note that ASCON produces a result similar to organ semantic segmentation after clustering and the intra-class distribution is more compact, as well as the inter-class separation is more obvious. To the best of our knowledge, this is the first time that anatomical semantic information has been demonstrated in a CT denoising task, providing interpretability to the field of medical image reconstruction.

Ablation Studies. We start with a ESAU-Net using MSE loss and gradually insert some loss functions and our MAC-Net. Table 2 presents the results of different loss functions. It shows that both the global non-contrastive module and local contrastive module are helpful in obtaining better metrics due to the capacity of exploiting inherent anatomical information and maintaining anatomical consistency. Then, we add our MAC-Net to two supervised models: RED-CNN [1] and U-Net [22] but it is less effective, which demonstrates the importance of our ESAU-Net that captures both local and global contexts during denoising in Table S1. In addition, we evaluate the effectiveness of the training strategies including alternate learning, neighboring positive matching and hard negative sampling in Table S2.

4 Conclusion

In this paper, we explore the anatomical semantics in LDCT denoising and take advantage of it to improve the denoising performance. To this end, we propose an **A**natomy-aware **S**upervised **CON**trastive learning framework (ASCON), consisting of an efficient self-attention-based U-Net (ESAU-Net) and a multi-scale anatomical contrastive network (MAC-Net), which can capture both local and global contexts during denoising and exploit inherent anatomical information. Extensive experimental results on Mayo-2016 and Mayo-2020 datasets demonstrate the superior performance of our method, and the effectiveness of our designs. We also validated that our method introduces interpretability to LDCT denoising.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (No. 62101136), Shanghai Sailing Program (No. 21YF1402800), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab, Shanghai Municipal of Science and Technology Project (No. 20JC1419500), and Shanghai Center for Brain Science and Brain-inspired Technology.

References

1. Chen, H., et al.: Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging **36**(12), 2524–2535 (2017)
2. Chen, Z., Niu, C., Wang, G., Shan, H.: LIT-Former: Linking in-plane and through-plane transformers for simultaneous CT image denoising and deblurring. arXiv preprint arXiv:2302.10630 (2023)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Gao, Q., Li, Z., Zhang, J., Zhang, Y., Shan, H.: CoreDiff: Contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization. arXiv preprint arXiv:2304.01814 (2023)
5. Gao, Q., Shan, H.: CoCoDiff: a contextual conditional diffusion model for low-dose CT image denoising. In: Developments in X-Ray Tomography XIV, vol. 12242. SPIE (2022)

6. Geng, M., et al.: Content-noise complementary learning for medical image denoising. IEEE Trans. Med. Imaging **41**(2), 407–419 (2021)
7. Grill, J.B., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Proc. Adv. Neural Inf. Process. Syst. **33**, 21271–21284 (2020)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
9. Huang, Z., Zhang, J., Zhang, Y., Shan, H.: DU-GAN: generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising. IEEE Trans. Instrum. Meas. **71**, 1–12 (2021)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
12. Liang, T., Jin, Y., Li, Y., Wang, T.: EDCNN: edge enhancement-based densely connected network with compound loss for low-dose CT denoising. In: 2020 15th IEEE International Conference on Signal Processing, vol. 1, pp. 193–198. IEEE (2020)
13. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
15. McCollough, C.H., et al.: Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. Med. Phys. **44**(10), e339–e352 (2017)
16. Moen, T.R., et al.: Low-dose CT image and projection dataset. Med. Phys. **48**(2), 902–911 (2021)
17. Mussmann, B.R., et al.: Organ-based tube current modulation in chest CT. A comparison of three vendors. Radiography **27**(1), 1–7 (2021)
18. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
19. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 319–345. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_19
20. Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L.: U-Net transformer: self and cross attention for medical image segmentation. In: Lian, C., Cao, X., Rekik, I., Xu, X., Yan, P. (eds.) MLMI 2021. LNCS, vol. 12966, pp. 267–276. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87589-3_28
21. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
23. Shan, H., et al.: Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. Nat. Mach. Intell. **1**(6), 269–276 (2019)

24. Shan, H., et al.: 3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network. IEEE Trans. Med. Imaging **37**(6), 1522–1534 (2018)
25. Yan, K., et al.: SAM: self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Trans. Med. Imaging **41**(10), 2658–2669 (2022)
26. Yang, Q., et al.: Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging **37**(6), 1348–1357 (2018)
27. Yun, S., Lee, H., Kim, J., Shin, J.: Patch-level representation learning for self-supervised vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8354–8363 (2022)
28. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728–5739 (2022)