Dense Transformer based Enhanced Coding Network for Unsupervised Metal Artifact Reduction

Wangduo Xie[(✉)] and Matthew B. Blaschko

Center for Processing Speech and Images, Department of ESAT,
KU Leuven, Leuven, Belgium
{wangduo.xie,matthew.blaschko}@esat.kuleuven.be

Abstract. CT images corrupted by metal artifacts have serious negative effects on clinical diagnosis. Considering the difficulty of collecting paired data with ground truth in clinical settings, unsupervised methods for metal artifact reduction are of high interest. However, it is difficult for previous unsupervised methods to retain structural information from CT images while handling the non-local characteristics of metal artifacts. To address these challenges, we proposed a novel ***D**ense **T**ransformer based **E**nhanced **C**oding **Net**work* (**DTEC-Net**) for unsupervised metal artifact reduction. Specifically, we introduce a Hierarchical Disentangling Encoder, supported by the high-order dense process, and transformer to obtain densely encoded sequences with long-range correspondence. Then, we present a second-order disentanglement method to improve the dense sequence's decoding process. Extensive experiments and model discussions illustrate DTEC-Net's effectiveness, which outperforms the previous state-of-the-art methods on a benchmark dataset, and greatly reduces metal artifacts while restoring richer texture details.

Keywords: Metal artifact reduction · CT image restoration · Unsupervised learning · Enhanced coding

1 Introduction

CT technology can recover the internal details of the human body in a non-invasive way and has been widely used in clinical practice. However, if there is metal in the tissue, metal artifacts (MA) will appear in the reconstructed CT image, which will corrupt the image and affect the medical diagnosis [1,6].

In light of the clinical need for MA reduction, various traditional methods [5,6,12,19] have been proposed to solve the problem by using interpolation and iterative optimization. As machine learning research increasingly impacts medical imaging, deep learning based methods have been proposed for MA reduction. Specifically, these methods can be roughly divided into supervised and unsupervised categories according to the degree of supervision. In the supervised category, the methods [9,14,16,18] based on the dual domain (sinogram and image

domains) can achieve good performance for MA reduction. However, supervised learning methods are hindered by the lack of large-scale real-world data pairs consisting of "images with MA" and "images without MA" representing the same region. The lack of such data can lead algorithms trained on synthetic data to over-fit simulated data pairs, resulting in difficulties in generalizing to clinical settings [11]. Furthermore, although sinogram data can bring additional information, it is difficult to collect in realistic settings [8,15]. Therefore, *unsupervised methods based only on the image domain* are strongly needed in practice.

For unsupervised methods in the image domain, Liao *et al.* [8] used Generative Adversarial Networks (GANs) [2] to disentangle the MA from the underlying clean structure of the artifact-affected image in latent space by using unpaired data with and without MA. Although the method can separate the artifact component in the latent space, the features from the latent space can't represent rich low-level information of the original input. Further, it's also hard for the encoder to represent long-range correspondence across different regions. Accordingly, the restored image loses texture details and can't retain structure from the CT image. In the same unsupervised setting, Lyu *et al.* [11] directly separate the MA component and clean structure in image space using a CycleGAN-based method [25]. Although implementation in the image space makes it possible to construct dual constraints, directly operating in the image space affects the algorithm's performance upper limit, because it is difficult to encode in the image space as much low-level information as the feature space.

Considering the importance of low-level features in the latent space for generating the artifact-free component, we propose a novel *Dense Transformer based Enhanced Coding Network*(DTEC-Net) for unsupervised metal artifact reduction, which can obtain low-level features with hierarchical information and map them to a clean image space through adversarial training. DTEC-Net contains our developed Hierarchical Disentangling Encoder (HDE), which utilizes long-range correspondences obtained by a lightweight transformer and a high-order dense process to produce the enhanced coded sequence. To ease the burden of decoding the sequence, we also propose a second-order disentanglement method to finish the sequence decomposition. Extensive empirical results show that our method can not only reduce the MA greatly and generate high-quality images, but also surpasses the competing unsupervised approaches.

2   Methodology

We design a Hierarchical Disentangling Encoder(HDE) that can capture low-level sequences and enable high-performance restoration. Moreover, to reduce the burden of the decoder group brought by the complicated sequences, we propose a second-order disentanglement mechanism. The intuition is shown in Fig. 1.

2.1   Hierarchical Disentangling Encoder (HDE)

As shown in Fig. 2(a), the generator of DTEC-Net consists of three encoders and four decoders. We design the HDE to play the role of Encoder1 for enhanced
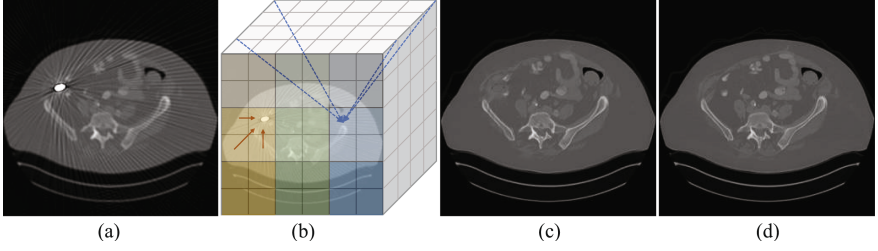
Fig. 1. (a) CT image with metal artifacts. (b) Blue/Orange arrows: Reuse of low-level features/Long-range correspondence. (c) Output of our DTEC-Net. (d) Ground truth. (Color figure online)
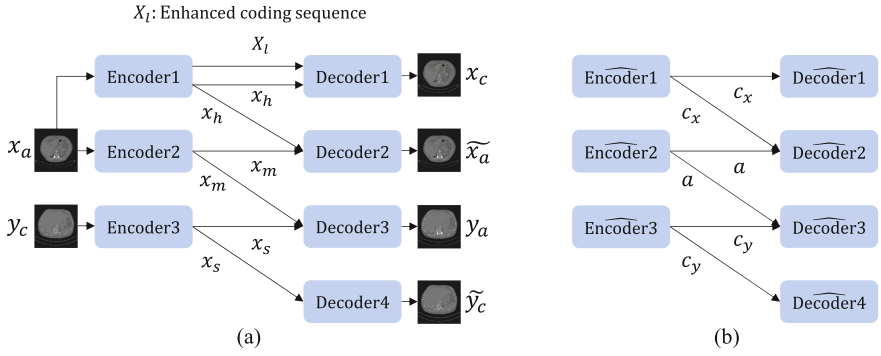

Fig. 2. (a) Generator of DTEC-Net. x_a: input with artifacts. y_c: unpaired input without artifacts. X_l and x_h are defined in the Sect. 2.1. x_m: the MA parts in latent space. x_s: the overall "structural part" in latent space. x_c (or y_a): the output after removing (or adding) the artifacts. $\widetilde{x_a}$ (or $\widetilde{y_c}$): the output of the identity map. (b) Generator of ADN [8]. The data relationship is shown in [8]. In addition to the difference in disentanglement, DTEC-Net and ADN also have different inner structures.

coding. Specifically, for the HDE's input image $x_a \in R^{1 \times H \times W}$ with MA, HDE first uses a convolution for the preliminary feature extraction and produces a high-dimensional tensor x_{l_0} with c channels. Then, x_{l_0} will be encoded by three *Dense Transformers for Disentanglement* (DTDs) in a first-order reuse manner [4,23]. Specifically, the output x_{l_i} of the ith DTD can be characterized as:

$$x_{l_i} = \begin{cases} f_{\text{DTD}_i}(f_{\text{s-hde}}(\text{cat}(x_{l_{i-1}}, x_{l_{i-2}}..., , x_{l_0}))), i = 2, ..., N. \\ f_{\text{DTD}_i}(x_{l_{i-1}}), i = 1. \end{cases} \quad (1)$$

In Eq. (1), $f_{\text{s-hde}}$ represents the channel compression of the concatenation of multiple DTDs' outputs, and N represents the total number of DTDs in the HDE. As shown in Fig. 3, HDE can obtain the hierarchical information sequence $X_l \triangleq \{x_{l_0}, ...x_{l_N}\}$ and high-level semantic features $x_h \triangleq x_{l_N}$.

As shown in Fig. 2(b), Encoder1 of ADN cannot characterize upstream low-level information, and results in limited performance. By using HDE, the
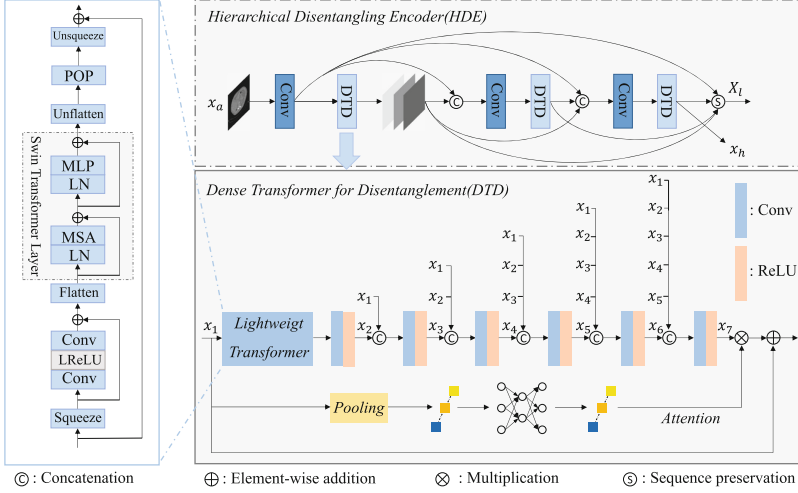
Fig. 3. The architecture of Hierarchical Disentangling Encoder (HDE).

upstream of the DTEC-Net's Encoder1 can represent rich low-level information, and be encoded in the efficient way described in Eq. (1). After generating the enhanced coding sequences X_l with long-range correspondence and densely reused information, DTEC-Net can decode it back to the clean image domain by using the proposed second-order disentanglement for MA reduction, which reduces the decoder group's burden to a large extent.

2.2 Dense Transformer for Disentanglement (DTD)

In addition to the first-order feature multiplexing given in Eq. (1), HDE also uses the DTD to enable second-order feature reuse. The relationship between HDE and DTD is shown in Fig. 3. Inspired by [7,20], DTD first uses a lightweight transformer based on the Swin transformer [7,10] to represent content-based information with long-range correspondence inside of every partition window. It then performs in-depth extraction and second-order reuse.

Specifically, the input $x_1 \in R^{C \times H \times W}$ of the DTD will be processed sequentially by the lightweight transformer and groups of convolutions in the form of second-order dense connections. The output x_{j+1} of the jth convolution with ReLU, which is connected in a second-order dense pattern, can be expressed as:

$$x_{j+1} = \begin{cases} f_{c_j}(\mathrm{cat}(x_1, x_2, ..., x_j)), j = 2, 3, ..., J. \\ f_{c_j}(f_{\text{transformer-light}}(x_j)), j = 1. \end{cases} \tag{2}$$

In Eq. (2), f_{c_j} indicates the jth convolution with ReLU after the lightweight transformer, and the J indicates the total number of convolutions after the lightweight transformer and is empirically set to six. The dense connection method can effectively reuse low-level features [22,23] so that the latent space

including these type of features will help the decoder to restore clean images without metal artifacts. Because the low-level information on different channels has different importance to the final restoration task, we use the channel attention mechanism [3] to filter the output of the final convolution layer:

$$x_{\text{out}} = x_{J+1} \odot f_{\text{MLP}}(f_{\text{pooling}}(x_1)) + x_1, \tag{3}$$

where $\odot$ represents the Hadamard product, f_{MLP} indicates a multi-layer perceptron with only one hidden layer, and f_{pooling} represents global pooling.

Because the transformer usually requires a large amount of data for training and CT image datasets are usually smaller than those for natural images, we do lightweight processing for the Swin transformer. Specifically, for an input tensor $x \in R^{C \times H \times W}$ of the lightweight transformer, the number of channels will be reduced from C to C_{in} to lighten the burden of the attention matrix. Then, a residual block is employed to extract information with low redundancy.

After completing lightweight handling, the tensor will first be partitioned into multiple local windows and flattened to $x_{\text{in}} \in R^{(\frac{HW}{P^2}) \times P^2 \times C_{in}}$ according the pre-operation [7,10] of the Swin transformer. $P \times P$ represents the window size for partitioning as shown in Fig. 1(b). Then, the attention matrix belonging to the ith window can be calculated by pairwise multiplication between converted vectors in $S_i \triangleq \{x_{\text{in}}(i,j,:)|j = 0,1,...,P^2 - 1\}$. Specifically, by using a linear map from $R^{C_{\text{in}}}$ to R^{C_a} for every vector in S_i, the query key and value: Q, K, $V \in R^{P^2 \times C_a}$ can be derived. Afterwards, the attention matrix for each window can be obtained by the following formula [10]:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{C_a})V. \tag{4}$$

In actual operation, we use window-based multi-head attention (MSA) [7,10,13] to replace the single-head attention because of the performance improvement [7]. The output of the Swin transformer layer will be unflattened and operated by post processing (POP) which consists of a classic convolution and layer norm (LN) with flatten and unflatten operations. After POP, the lightweight tensor with fewer channels will be unsqueezed to expand the channels, and finally added to the original input x in the form of residuals.

2.3   Second-Order Disentanglement for MA Reduction (SOD-MAR)

As mentioned in Sect. 2.1, X_l represents the hierarchical sequence and facilitates the generator's representation. However, X_l needs to be decoded by a high-capacity decoder to match the encoder. Considering that Decoder2 does not directly participate in the restoration branch and already loaded up the complicated artifact part x_m in traditional first-order disentanglement learning [8], to reduce the burden of the decoder group, we propose and analyze SOD-MAR.

Specifically, Decoder2 of DTEC-Net doesn't decode sequence X_l, it only decodes the combination of second-order disentangled information $x_h \in X_l$ and the latent feature x_m representing the artifact parts shown in Fig. 2(a). In order
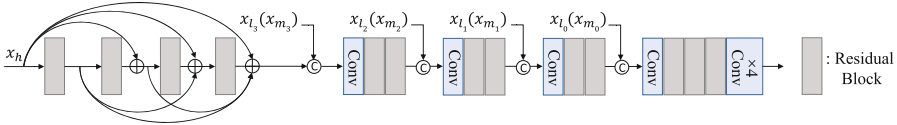
Fig. 4. The architecture of Decoder1 (Decoder2). (x_{m_*}) represents the Decoder2 case.

to complete the process, Decoder2 uses the structure shown in Fig. 4 to finish the decoding step, which is also used by Decoder1 to decode the sequence X_l.

Moreover, we don't only map the x_h into Decoder1 and Decoder2 while dropping the $X_l \backslash \{x_h\}$ to implement the burden reduction, because the low-level information in $X_l \backslash \{x_h\}$ is vital for restoring artifact-free images. Furthermore, x_h will be disturbed by noise from the approaching target x_a of Decoder2 while information $X_l \setminus \{x_h\}$ upstream from the HDE can counteract the noise disturbance to a certain extent. The reason behind the counteraction is that the update to upstream parameters is not as large as that of the downstream parameters.

2.4   Loss Function

Following [8], we use discriminators D_0, D_1 to constrain the output x_c and y_a:

$$L_{adv} = \mathbb{E}[\log(1 - D_0(x_c)) + \log(D_0(y_c))] + \mathbb{E}[\log(D_1(x_a)) + \log(1 - D_1(y_a))].$$
(5)

The above x_a, y_c represent the input as shown in Fig. 2(a). Following [8], we use the reconstruction loss L_{rec} to constrain the identity map, and also use the artifact consistency loss L_{art} and self-reduction loss L_{self} to control the optimization process. The coefficients for each of these losses are set as in [8].

3   Empirical Results

Synthesized DeepLesion Dataset. Following [8], we randomly select 4186 images from DeepLesion [17] and 100 metal templates [21] to build a dataset. The simulation is consistent with [8,21]. For training, we randomly select 3986 images from DeepLesion combined with 90 metal templates for simulation. The 3986 images will be divided to two disjoint image sets with and without MA after simulation. Then a random combination can form the physically unpaired data with and without MA in the training process. Besides, another 200 images combined with the remaining 10 metal templates are used for the testing process.

Real Clinic Dataset. We randomly combine 6165 artifacts-affected images and 20729 artifacts-free images from SpineWeb[1] [8] for training, and 105 artifacts-affected images from SpineWeb for testing.

---

[1] spineweb.digitalimaginggroup.ca.

Implementation Details. We use peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to measure performance. We use mean squared error (MSE) only for measuring ablation experiments. For Synthesized DeepLesion dataset (and Real Clinic dataset), we set the batch size to 2 (and 2) and trained the network for 77 (and 60) epochs using the Adam optimizer. Our DTEC-Net was implemented in Pytorch using an Nvidia Tesla P100.

Table 1. Ablation study on Synthesized DeepLesion under different settings. ↑: Higher value is better; ↓: Lower value is better. The best values are in **bold**.

| Model	PSNR↑	SSIM↑	MSE↓
HDE with one DTD	34.46	0.937	27.12
HDE with two DTD	34.71	0.938	24.96
HDE with three DTD (only Transformer)	34.31	0.936	27.40
HDE with three DTD (without SOD-MAR)	34.91	0.940	24.36
HDE with three DTD (with SOD-MAR, Ours)	**35.11**	**0.941**	**22.89**

3.1 Ablation Study

To verify the effectiveness of the proposed methods, ablation experiments were carried out on Synthesized DeepLesion. The results are shown in Table 1.

The Impact of DTD in HDE. In this experiment, we change the encoding ability of HDE by changing the number of DTDs. We first use only one DTD to build the HDE, then the PSNR is 0.65 dB lower than our DTEC-Net using three DTDs. Additionally, the average MSE in this case is much higher than DTEC-Net. When the number of DTDs increases to two, the performance improves by 0.25 dB and is already better than the SOTA method [11]. As we further increase the number of DTDs to three, the PSNR and SSIM increase 0.4 dB and 0.003, respectively. The number of DTDs is finally set to three in a trade-off between computation and performance. To match different encoders and decoders and facilitate training, we also adjust the accept headers of Decoder1 to adapt to the sequence length determined by the different numbers of DTDs.

Only Transformer in DTD. Although the transformer can obtain better long-range correspondence than convolutions, it lacks the multiplexing of low-level information. For every DTD in DTEC-Net, we delete the second-order feature reuse pattern and only keep the lightweight transformer, the degraded version's results are 0.8 dB lower than our DTEC-Net. At the same time, great instability appears in generative adversarial training. So, only using the transformer cannot achieve good results in reducing metal artifacts.

Removing SOD-MAR. Although SOD-MAR mainly helps by easing the burden of decoding as discussed in Sect. 2.3, it also has a performance gain compared to first-order disentanglement. We delete the SOD-MAR in DTEC-Net and let x_h be the unique feature decoded by Decoder1. The Performance is 0.2 dB lower than our DTEC-Net, while MSE increases by 1.47.

| Input image with artifact | ADN | Ours | GroundTruth |
| 28.79 / 0.733 | 33.52 / 0.935 | 35.19 / 0.951 | |
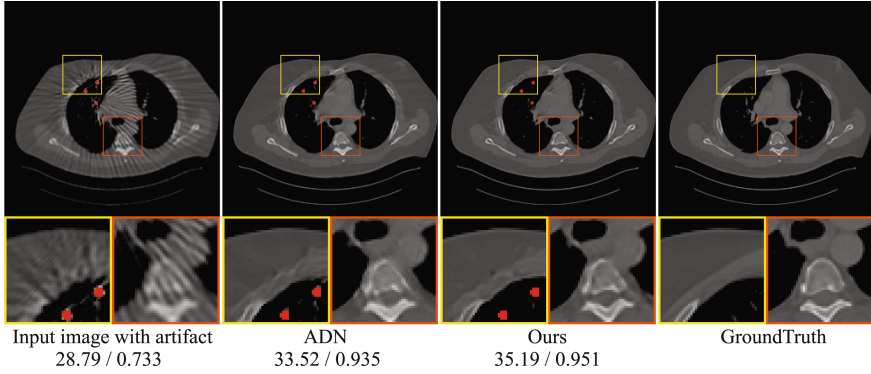
Fig. 5. Visual comparison(Metal implants are colored in red. The bottom values represent PSNR/SSIM). Our method has sharper edges and richer textures than ADN. (Color figure online)

Table 2. Quantitative comparison of different methods on Synthesized DeepLesion. The best results are in **bold**.

Method Classification	Method	PSNR↑	SSIM↑
Conventional	LI [5]	32.00 [8]	0.910 [8]
Supervised	CNNMAR [21]	32.50 [8]	0.914 [8]
Unsupervised	CycleGAN [25]	30.80 [8]	0.729 [8]
Unsupervised	RCN [24]	32.98 [11]	0.918 [11]
Unsupervised	ADN [8]	33.60 [8]	0.924 [8]
Unsupervised	U-DuDoNet [11]	34.54 [11]	0.934 [11]
Unsupervised	DTEC-Net(Ours)	**35.11**	**0.941**

3.2   Comparison to State-of-the-Art (SOTA)

For a fair comparison, we mainly compare with SOTA methods under unsupervised settings: ADN [8], U-DuDoNet [11], RCN [24], and CycleGAN [25]. We also compare with the traditional method LI [5] and classical supervised method CNNMAR [21]. The quantitative results of ADN, CycleGAN, CNNMAR and LI are taken from [8], the results of U-DuDoNet and RCN are taken from [11]. Because ADN has open-source code, we run their code for qualitative results.

Quantitative Results. As shown in Table 2. For the Synthesized DeepLesion Dataset, our method has the highest PSNR and SSIM value and outperforms the baseline ADN by 1.51 dB in PSNR and 0.017 in SSIM. At the same time, it also exceeds the SOTA method U-DuDoNet by 0.57 dB. For the Real Clinic Dataset, the numerical results can't be calculated because the ground truth does not exist. We will present the qualitative results in the appendix. Furthermore, as our work is single-domain based, it has the potential to be easily applied in clinical practice.

Qualitative Results. A visual comparison is shown in Fig. 5. Our method not only reduces artifacts to a large extent, but also has sharper edges and richer textures than the compared method. More results are shown in the appendix.

4 Conclusion

In this paper, we proposed a Dense Transformer based Enhanced Coding Network (DTEC-Net) for unsupervised metal-artifact reduction. In DTEC-Net, we developed a Hierarchical Disentangling Encoder (HDE) to represent long-range correspondence and produce an enhanced coding sequence. By using this sequence, the DTEC-Net can better recover low-level characteristics. In addition, to decrease the burden of decoding, we specifically design a Second-order Disentanglement for MA Reduction (SOD-MAR) to finish the sequence decomposition. The extensive quantitative and qualitative experiments demonstrate our DTEC-Net's effectiveness and show it outperforms other SOTA methods.

References

1. Barrett, J.F., Keat, N.: Artifacts in CT: recognition and avoidance. Radiographics **24**(6), 1679–1691 (2004)
2. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
5. Kalender, W.A., Hebel, R., Ebersberger, J.: Reduction of CT artifacts caused by metallic implants. Radiology **164**(2), 576–577 (1987)
6. Lemmens, C., Faul, D., Nuyts, J.: Suppression of metal artifacts in CT using a reconstruction procedure that combines map and projection completion. IEEE Trans. Med. Imaging **28**(2), 250–260 (2008)
7. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)
8. Liao, H., Lin, W.A., Zhou, S.K., Luo, J.: ADN: artifact disentanglement network for unsupervised metal artifact reduction. IEEE Trans. Med. Imaging **39**(3), 634–643 (2019)
9. Lin, W.A., et al.: Dudonet: dual domain network for CT metal artifact reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10512–10521 (2019)
10. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

11. Lyu, Y., Fu, J., Peng, C., Zhou, S.K.: U-DuDoNet: unpaired dual-domain network for CT metal artifact reduction. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 296–306. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_29

12. Meyer, E., Raupach, R., Lell, M., Schmidt, B., Kachelrieß, M.: Normalized metal artifact reduction (NMAR) in computed tomography. Med. Phys. **37**(10), 5482–5493 (2010)

13. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

14. Wang, H., Li, Y., Zhang, H., Meng, D., Zheng, Y.: Indudonet+: a deep unfolding dual domain network for metal artifact reduction in CT images. Med. Image Anal. **85**, 102729 (2023)

15. Wang, H., Xie, Q., Li, Y., Huang, Y., Meng, D., Zheng, Y.: Orientation-shared convolution representation for CT metal artifact learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 665–675. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_63

16. Wang, T., et al.: Dual-domain adaptive-scaling non-local network for CT metal artifact reduction. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 243–253. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_24

17. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging **5**(3), 036501–036501 (2018)

18. Yu, L., Zhang, Z., Li, X., Xing, L.: Deep sinogram completion with image prior for metal artifact reduction in CT images. IEEE Trans. Med. Imaging **40**(1), 228–238 (2020)

19. Zhang, H., Wang, L., Li, L., Cai, A., Hu, G., Yan, B.: Iterative metal artifact reduction for x-ray computed tomography using unmatched projector/backprojector pairs. Med. Phys. **43**(6Part1), 3019–3033 (2016)

20. Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Accurate image restoration with attention retractable transformer. arXiv preprint arXiv:2210.01427 (2022)

21. Zhang, Y., Yu, H.: Convolutional neural network based metal artifact reduction in x-ray computed tomography. IEEE Trans. Med. Imaging **37**(6), 1370–1381 (2018)

22. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 294–310. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_18

23. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)

24. Zhao, B., Li, J., Ren, Q., Zhong, Y.: Unsupervised reused convolutional network for metal artifact reduction. In: Yang, H., Pasupa, K., Leung, A.C.-S., Kwok, J.T., Chan, J.H., King, I. (eds.) ICONIP 2020. CCIS, vol. 1332, pp. 589–596. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63820-7_67

25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)