Anatomy-Aware Low-Dose CT Denoising via Pretrained Vision Models and Semantic-Guided Contrastive Learning

Runze Wang[1,2,3], Zeli Chen[1,2], Zhiyun Song[1,4], Wei Fang[1], Jiajin Zhang[1,2,4], Danyang Tu[1,2], Yuxing Tang[1], Minfeng Xu[1], Xianghua Ye[5], Le Lu[1], and Dakai Jin[1]

[1] DAMO Academy, Alibaba Group
[2] Hupan Lab, 310023, Hangzhou, China
[3] Fudan University, Shanghai, China
[4] Shanghai Jiao Tong University, Shanghai, China
[5] The First Affiliated Hospital Zhejiang University, Hangzhou, China
{fashe.wrz, dakai.jin}@alibaba-inc.com

Abstract. To reduce radiation exposure and improve the diagnostic efficacy of low-dose computed tomography (LDCT), numerous deep learning-based denoising methods have been developed to mitigate noise and artifacts. However, most of these approaches ignore the anatomical semantics of human tissues, which may potentially result in suboptimal denoising outcomes. To address this problem, we propose ALDEN, an anatomy-aware LDCT denoising method that integrates semantic features of pre-trained vision models (PVMs) with adversarial and contrastive learning. Specifically, we introduce an anatomy-aware discriminator that dynamically fuses hierarchical semantic features from reference normal-dose CT (NDCT) via cross-attention mechanisms, enabling tissue-specific realism evaluation in the discriminator. In addition, we propose a semantic-guided contrastive learning module that enforces anatomical consistency by contrasting PVM-derived features from LDCT, denoised CT and NDCT, preserving tissue-specific patterns through positive pairs and suppressing artifacts via dual negative pairs. Extensive experiments conducted on two LDCT denoising datasets reveal that ALDEN achieves the state-of-the-art performance, offering superior anatomy preservation and substantially reducing over-smoothing issue of previous work. Further validation on a downstream multi-organ segmentation task (encompassing 117 anatomical structures) affirms the model's ability to maintain anatomical awareness.

Keywords: Anatomy-aware low-dose CT denoising · Pre-trained vision models · Semantic-guided contrastive learning.

1 Introduction

Low-dose computed tomography (LDCT) has become an important and popular diagnostic tool for reducing radiation exposure risks; however, its clinical utility

is hindered by the amplified noise and artifacts that degrade anatomical fidelity. Deep learning advances of convolutional neural networks [6,31], transformer [21], and diffusion models [8], have improved LDCT denoising, these methods share a common limitation: pixel-level constraints (e.g., L1/MSE losses) prioritize global error reduction at the expense of local anatomical plausibility, often resulting in oversmooth textures that obscure small tissues and subtle pathologies [29].

Generative adversarial networks (GANs) offer an alternative by learning data distributions rather than pixel-wise mappings [9,23,22]. However, conventional GAN-based LDCT denoising methods [11,25] often do not capture the important relationship between noise characteristics and anatomical semantics, as noise levels in CT images differ depending on tissue type [18,6]. Recent work also highlights the need for image understanding in restoration to improve the explainability and clinical application of medical imaging [20,14,6,5]. This calls for a shift towards fine-grained anatomy-aware denoising, where semantic consistency is crucial for effective texture restoration.

Integrating anatomical semantics into denoising models presents challenges since conventional task-specific segmentation networks [27,10] require costly and precise anatomical annotations, limiting generalizability across a large number of diverse anatomies. Progresses in foundation models [19,28] demonstrate that pretrained vision models (PVMs) pretrained on large-scale datasets possess exceptional transfer learning capability for semantic understanding. PVMs offer two key advantages: (1) their exposure to millions of natural images allows the development of rich hierarchical feature representations that capture universal texture and structure patterns, which can be adapted to the medical image domain [2,32,1,17]; and (2) unlike segmentation networks requiring predefined anatomical labels, PVMs generate semantic features without explicit supervision, enabling the discovery of latent anatomical relationships essential for fine-grained denoising.

Inspired by this, we present **ALDEN** (**A**natomy-aware **L**DCT **DEN**oising framework), which integrates PVMs within a GAN architecture for enhanced anatomy-aware restoration. ALDEN features an anatomy-aware discriminator that utilizes hierarchical semantic features extracted from reference NDCT via PVMs, guiding adversarial learning to concentrate on tissue-specific semantics. This differentiates it from previous GAN-based LDCT denoising methods that evaluate all anatomical structures uniformly. Additionally, we propose a semantically guided contrastive paradigm that uses PVM-extracted features to enforce anatomy-aware consistency. Positive pairs align features from corresponding anatomical regions in denoised CT and NDCT, while negative pairs include features from denoised CT and LDCT at the same location to emphasize noise, and from denoised CT and NDCT at mismatched locations to penalize anatomical misalignment. The InfoNCE loss [4] is utilized to minimize the distances between positive pairs and maximize the separation from negative pairs.

Our contributions are summarized as follows. 1) We first propose the integration of PVMs into LDCT denoising, uniquely combining PVMs with adversarial and contrastive learning approaches. 2) We introduce an anatomy-aware

discriminator that dynamically incorporates hierarchical semantic features from NDCT to enable a fine-grained semantic-aware LDCT denoising. 3) We present a semantically guided contrastive learning module to maintain anatomical consistency through positive pairs while reducing noise and artifacts with dual negative pairs. 4) Extensive experiments on two LDCT denoising datasets demonstrate that ALDEN achieves the state-of-the-art denoising performance, delivering enhanced texture preservation and avoiding over-smoothing. Further validation on a downstream multi-organ segmentation task with 117 anatomical structures demonstrates the effectiveness of our anatomy-aware denoising approach.

2 Methodology

2.1 Overview of ALDEN

Conventional GAN-based LDCT denoising frameworks [11,25] employ a generator-discriminator architecture to enhance perceptual quality. Let $X \sim P_X$ denote LDCT inputs and $Y \sim P_Y$ their NDCT counterparts. The generator (i.e., denoising network) G produces denoised outputs $\hat{Y} = G(X)$, optimized through two objectives: 1) *Pixel fidelity* via pixl-wise supervised loss. Here, L_1 loss is employed, which is expressed as $\mathcal{L}_1 = \|\hat{Y} - Y\|_1$. 2) *Distribution alignment* via adversarial learning. The discriminator D differentiates between real NDCT images Y and generated/denoised outputs $\hat{Y}$. Simultaneously, the generator G is encouraged to produce realistic denoised CT $\hat{Y}$ that can effectively compete with the discriminator. The G and D form a two-player minimax game, optimized through adversarial loss defined as $\mathcal{L}_{adv} = \mathbb{E}_{Y\sim P_Y}[\log D(Y)] + \mathbb{E}_{X\sim P_X}[\log(1 - D(G(X)))]$.

As illustrated in Fig. 1 (left), our ALDEN framework extends the traditional GAN-based LDCT denoising model by introducing two key components: the Anatomy-Aware Discriminator (AAD) and Semantic-guided Contrastive Learning (SCL). The AAD leverages NDCT-derived semantic features extracted from PVMs as conditional input. Let Ψ represent the PVMs, the goal is to align conditional distributions: $P(\hat{Y}|\Psi(Y)) \approx P(Y|\Psi(Y))$, facilitating detailed anatomy-aware texture restoration. In contrast, conventional discriminators operate only on marginal distributions: $P(\hat{Y}) \approx P(Y)$, potentially overlooking essential semantic content. The SCL component utilizes features extracted by PVMs to enforce anatomy-aware consistency through contrastive learning. It achieves this by forming positive pairs to preserve tissue-specific patterns and dual negative pairs to suppress noise and artifacts. More details on AAD can be found in Section 2.2, while the formalization of SCL is presented in Section 2.3.

2.2 Anatomy-Aware Discriminator

The AAD aims to achieve fine-grained semantic-aware denoising through adversarial learning. Inspired by [14], we propose an Attention-based Feature Fusion (AFF) module that integrates hierarchical semantic priors from the reference NDCT Y using PVMs. As shown in Fig. 1 (left), our multi-level AFF operates

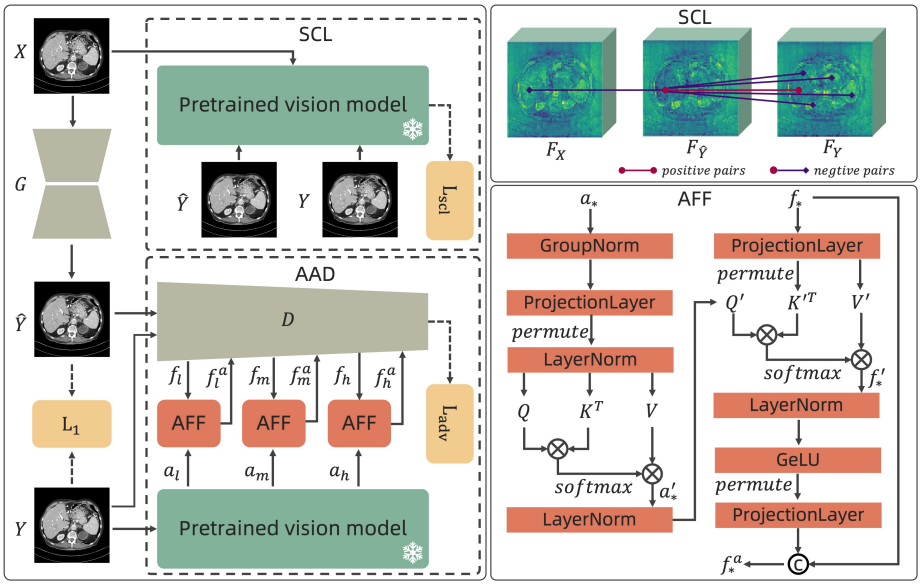

Fig. 1. Overview of the proposed ALDEN (**A**natomy-aware **L**DCT **DEN**oising) framework, which integrates the Anatomy-Aware Discriminator (AAD) for fine-grained texture restoration and Semantic-guided Contrastive Learning (SCL) for enhanced noise suppression.

across the feature hierarchies of both the PVM and the discriminator, providing progressive semantic guidance. To enhance the quality of semantic priors, we also explore advanced PVMs such as DINOv2 [19] and MedSAM [15], which are the state-of-the-art pretrained vision models in the fields of natural and medical images, offering robust semantic feature extraction capabilities.

Given the NDCT Y, we extract hierarchical semantic features from a fixed PVM at three levels: low (a_l), middle (a_m), and high (a_h). Both MedSAM and DINOv2 use the ViT-base architecture with 12 transformer blocks. Although ViT doesn't explicitly change scales, features from different layers are analogous to CNNs: early layers focus on low-level features, while later ones concentrate on high-level semantics [7]. We use outputs from the 4th, 8th, and 12th transformer blocks as low-, mid-, and high-level features, respectively. At the same time, when the denoised CT $\hat{Y}$ or the NDCT Y is the input to the discriminator D, it generates the corresponding hierarchical discriminative features, i.e., f_l, f_m and f_h at these levels. Then multiple AFF modules are used to align the features f_* (where $* \in l, m, h$) with the semantic features a_* before passing them to the discriminator. This process guides the discriminator to focus on semantically relevant texture distributions.

As shown in Fig. 1 (bottom right), the AFF module begins by standardizing the semantic features a_* through group normalization, followed by a projection layer with a $1 * 1$ convolution. After permutation and layer normalization, we

derive the query Q, key K, and value V. The self-attention outputs a'_* are computed as:

$$a'_* = Softmax(Q \cdot K^T / \sqrt{d_k}) \cdot V, \tag{1}$$

where d_k is the scale factor. This output a'_* is normalized to produce the query Q' for the cross-attention mechanism. The feature f_* undergoes a similar projection and permutation process to become the key K' and the value V'. Then we can obtain the cross-attention results f'_* using Equation 1. Finally, the fused anatomy-aware feature representation f^a_* is computed as follows:

$$f^a_* = Concat(PL(Permute(GELU(LN(f'_*)))), f_*). \tag{2}$$

Here, $GELU$, PL, and LN are the Gaussian Error Linear Units, projection layer and layer normalization, respectively.

2.3 Semantic-guided Contrastive Learning

The SCL component of the ALDEN framework utilizes a pretrained vision model to improve LDCT denoising by ensuring anatomical consistency between denoised CT and reference NDCT. As depicted in Fig. 1 (top right), the feature representation F_X, $F_{\hat{Y}}$ and F_Y are derived from the fixed PVM using input X, $\hat{Y}$ and Y, respectively. Here, F_X, $F_{\hat{Y}}$ and $F_Y \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H and W represent batch size, channel depth, height, and width of the features, respectively. Then, SCL operates contrastive learning on these features, explicitly aligning denoised CT features with NDCT references while contrasting against two types of negative samples: residual noise patterns from LDCT and anatomically discordant NDCT features.

Positive Pair Alignment. For each denoised CT image, we establish anatomical correspondence with its NDCT counterpart through spatially aligned feature pairs. Let $F_{\hat{Y}}(x, y) \in \mathbb{R}^C$ and $F_Y(x, y) \in \mathbb{R}^C$ denote the PVM-derived feature vectors at the spatial coordinate (x, y) in the i-th batch sample. We randomly sample K coordinates per image, constructing positive pairs from identical anatomical locations, i.e., $\mathcal{P} = \{(F^{(i)}_{\hat{Y}}(x_{ik}, y_{ik})), F^{(i)}_Y(x_{ik}, y_{ik})\}^{B,K}_{i=1,k=1}$, where (x_{ik}, y_{ik}) represents randomly selected positions. This alignment ensures that the denoised output preserves anatomical structures observed in the NDCT ground truth.

Dual Negative Sampling Strategy. We design two complementary negative sampling mechanisms: 1) same-location LDCT negatives capture residual noise by contrasting denoised features against their LDCT counterparts at identical coordinates, i.e., $\mathcal{N}_1 = \{(F^{(i)}_{\hat{Y}}(x_{ik}, y_{ik})), F^{(i)}_X(x_{ik}, y_{ik})\}^{B,K}_{i=1,k=1}$. 2) Cross-location NDCT negatives penalize anatomical misalignment by pairing denoised features with NDCT features from spatially discordant regions M, i.e., $\mathcal{N}_2 = \{(F^{(i)}_{\hat{Y}}(x_{ik}, y_{ik})), F^{(i)}_Y(\tilde{x}^{(m)}_{ik}, \tilde{y}^{(m)}_{ik})\}^{B,K,M}_{i=1,k=1,m=1}$, where $\{(\tilde{x}^{(m)}_{ik}, \tilde{y}^{(m)}_{ik})\}$ are random-sampled coordinates excluding (x_{ik}, y_{ik}).

Loss Formulation. The SCL loss adapts the InfoNCE [4] to optimize feature similarities:

$$\mathcal{L}_{scl} = -\frac{1}{B \cdot K} \sum_{i,k} \log \frac{exp(s_{ik}^{pos}/\tau)}{exp(s_{ik}^{pos}/\tau) + exp(s_{ik}^{neg_1}/\tau) + \sum_{m=1}^{M} exp(s_{ikm}^{neg_2}/\tau)}, \quad (3)$$

where the temperature $\tau = 0.1$ and the similarity scores are computed as $s_{ik}^{pos} = \left\langle F_{\hat{Y}}^{(i)}(x_{ik}, y_{ik}), F_{Y}^{(i)}(x_{ik}, y_{ik}) \right\rangle$, $s_{ik}^{neg_1} = \left\langle F_{\hat{Y}}^{(i)}(x_{ik}, y_{ik}), F_{X}^{(i)}(x_{ik}, y_{ik}) \right\rangle$ and $s_{ikm}^{neg_2} = \left\langle F_{\hat{Y}}^{(i)}(x_{ik}, y_{ik}), F_{Y}^{(i)}(\tilde{x}_{ik}^{(m)}, \tilde{y}_{ik}^{(m)}) \right\rangle$, respectively. Here, $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and the loss simultaneously maximizes positive pair alignment while repelling both negative types. The overall objective function of the proposed ALDEN is as follows:

$$\mathcal{L} = \mathcal{L}_1 + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{scl}, \quad (4)$$

where $\{\lambda_1, \lambda_2\}$ are parameters controlling the relative weights of different losses, which are empirically set as $\{0.01, 0.5\}$.

3 Experiments and Results

3.1 Experimental Setup

Denoising Datasets. We evaluate our method using two datasets. Mayo2016 dataset [16] consists of 2,378 CT slices from ten anonymized patients, each of which has paired low-dose (quarter-dose) and normal-dose scan. Consistent with the data split protocol of previous studies [8,21], we select slices from nine patients for training and slices from one patient for testing. The second dataset is a collection of CT scans from multiple centers, referred to as the Multi-center CT Denoising (MCTD) dataset. Normal-dose CT scans were acquired from diverse clinical sites, while the corresponding low-dose CT scans were generated using a simulation algorithm [26]. The dataset consists of 1,276 paired scans for training and 88 paired scans for validation. For both datasets, 2D slices were extracted from the axial plane with a resolution of 512×512 pixels for model training and validation.

We employ four metrics for denoising evaluation: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), root mean square error (RMSE) and a perceptual metric, i.e., learned perceptual image patch similarity (LPIPS) [30], to assess image quality.

Segmentation Dataset. We further evaluate the denoising performance on a downstream multi-organ segmentation task. Specifically, the TotalSegmentator [24] test set is used, which includes 89 CT scans with 117 organ types. To simulate LDCT data with varying noise levels, we follow the methodology in [26] to generate LDCT with low and high noise conditions. Then different denoising algorithms are applied to restore the simulated LDCT, after which we use the nnUNet [12] trained on the original CT in the TotalSegmentator training set to predict organ masks and calculate the mean Dice similarity coefficient (DSC) of all organs.

Table 1. Quantitative comparison of different methods on the Mayo2016 and MCTD datasets.

Methods	Mayo2016 dataset				MCTD dataset			
	PSNR↑	SSIM↑	RMSE↓	LPIPS↓	PSNR↑	SSIM↑	RMSE↓	LPIPS↓
RED-CNN [3]	32.32	0.9227	9.87	0.0235	37.81	0.9069	50.11	0.0535
DU-GAN [11]	32.64	0.9278	9.56	0.0198	38.52	0.9075	43.33	0.0291
SeD [14]	32.87	0.9283	9.32	0.0201	39.43	0.9157	40.78	0.0304
CTformer [21]	33.16	0.9287	8.98	0.0235	38.97	0.9134	40.54	0.0458
CoreDiff [8]	33.61	0.9342	8.58	0.0227	40.46	0.9290	34.40	0.0474
ASCON [6]	33.60	0.9318	8.57	0.0242	**40.58**	0.9278	34.43	0.0657
ALDEN-MedSAM	33.59	**0.9343**	8.56	0.0192	40.51	0.9281	34.86	0.0273
ALDEN-DINOv2	**33.71**	0.9341	**8.44**	**0.0176**	40.57	**0.9296**	**34.37**	**0.0265**

Implementation Details. The generator utilized in our study is the ESAU-Net as introduced in [6]. The basic discriminator we used is a popular patch-wise discriminator [13] that consists of five convolutional layers. The proposed AAD incorporates the AFF within the middle three convolutional layers. The sampling hyperparameters K and M in the SCL module are empirically set to 256 and 32, respectively. Our experiments are conducted on an NVIDIA H20 GPU with 96 GB of memory using the PyTorch framework. We optimize our network with the Adam optimizer, utilizing a batch size of 8 and a learning rate of 1e-4. The optimization process is carried out for a total of 300,000 iterations.

3.2 Experimental Results

Comparison with Previous State-of-the-Art Methods. We implement two ALDEN variants based on different PVMs: ALDEN-MedSAM and ALDEN-DINOv2. These variants are compared against six state-of-the-art denoising methods with diverse network architectures, including GAN-based (DU-GAN [11] and SeD [14]), CNN-based (RED-CNN [3] and ASCON [6]), Transformer-based (CTformer [21]) and Diffusion-based (CoreDiff [8]).

Table 1 reveals an inherent trade-off between fidelity metrics (PSNR, SSIM, RMSE) and perceptual quality (LPIPS) in previous work. For example, on the Mayo2016 dataset, DU-GAN achieves a competitive LPIPS score of 0.0198 but underperforms in fidelity with an RMSE of 9.56, compared to 8.57-8.58 for ASCON and CoreDiff. Conversely, CoreDiff and ASCON exhibit superior fidelity (PSNR: 33.60-33.61 dB) but have higher LPIPS values (0.0227-0.0242), indicating a decline in perceptual quality. In contrast, the ALDEN variants, especially ALDEN-DINOv2, effectively balance these objectives. From the Mayo2016 dataset, ALDEN-DINOv2 achieves a new state-of-the-art result, reporting a PSNR of 33.71 dB, an RMSE of 8.44, and an LPIPS of 0.0176, along with a competitive SSIM of 0.9341. In MCTD dataset, it maintains its superiority, leading in SSIM (0.9296), RMSE (34.37) and LPIPS of 0.0265, indicating a 44.1-59.7% reduction compared to CoreDiff and ASCON. These results confirm ALDEN's ability to harmonize fidelity and perceptual quality through PVM guidance. Fig.

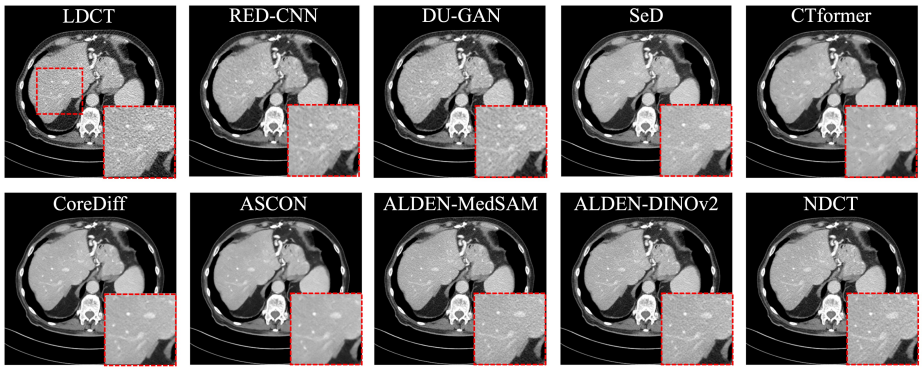

Fig. 2. Qualitative assessment of different methods on the Mayo2016 dataset. The display window is $[-160, 240]$ HU.

Table 2. Ablation study results on the Mayo2016 dataset.

Methods	AAD	SCL	PSNR↑	SSIM↑	RMSE↓	LPIPS↓
Baseline	-	-	32.70	0.9291	9.51	0.0206
AAD-DINOv2	✓	-	33.60	**0.9348**	8.57	0.0186
SCL-DINOv2	-	✓	33.41	0.9330	8.74	**0.0123**
ALDEN-DINOv2	✓	✓	**33.71**	0.9341	**8.44**	0.0176

2 presents the qualitative results of various methods applied to the Mayo2016 dataset. As shown, the proposed ALDEN-MedSAM and ALDEN-DINOv2 stand out by effectively preserving intricate textural details while maintaining exceptional noise reduction, leading to results that closely resemble the NDCT image.

Ablation Studies. Table 2 showcases the ablation results of ALDEN-DINOv2 in the Mayo2016 dataset, highlighting the roles of AAD and SCL in improving the performance of LDCT denoising. It is observed that incorporating AAD notably enhances PSNR, SSIM, RMSE and LPIPS as compared to the baseline model. Applying SCL independently produces improvements in these metrics, particularly in LPIPS, likely due to PVMs' efficacy as perceptual feature extractors. The combination of AAD and SCL delivers the best performance, achieving a PSNR of 33.71, RMSE of 8.44, and maintaining balanced results in SSIM (0.9341) and LPIPS (0.0176), highlighting the effectiveness of integrating these components to enhance denoising fidelity and maintain perceptual quality.

Table 3. Quantitative comparison of different denoising methods on DSC (%) for downstream multi-organ segmentation tasks.

Noise level	LDCT	RED-CNN	DU-GAN	SeD	CTformer	CoreDiff	ASCON	ALDEN-DINOv2
Low	87.54	88.22	88.74	88.68	88.53	89.06	88.94	89.20
High	75.74	78.14	78.01	79.37	78.96	79.99	79.60	81.06

Downstream Task Evaluation. We evaluate the performance of the proposed ALDEN-DINOv2 for multi-organ segmentation as downstream task. As shown in Table 3, our method consistently achieves the best segmentation performance in both low- and high-noise scenarios, with DSC values of 89.20% and 81.06%, respectively. Especially in the high-noise scenario, our method substantially outperforms the second-best CoreDiff by a mean DSC of 1.07% across 117 anatomical structures. These results demonstrate that our approach effectively improves anatomical perception and enhances segmentation performance.

4 Conclusion

In conclusion, we present ALDEN that combines pretrained vision models with adversarial and contrastive learning techniques for anatomy-aware low-dose CT denoising. The framework features an anatomy-aware discriminator for fine-grained denoising and a semantic-guided contrastive learning module to enhance anatomical consistency. Extensive experiments demonstrate that ALDEN achieves state-of-the-art performance, improving texture preservation while reducing over-smoothing. Validation on a multi-organ segmentation task with 117 anatomical structures underscores the model's robust anatomical awareness.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ayzenberg, L., Giryes, R., Greenspan, H.: Dinov2 based self supervised learning for few shot medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
2. Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al.: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering **7**(6), 756–779 (2023)
3. Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J., Wang, G.: Low-dose ct with a residual encoder-decoder convolutional neural network. IEEE transactions on medical imaging **36**(12), 2524–2535 (2017)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Chen, Z., Chen, T., Wang, C., Gao, Q., Niu, C., Wang, G., Shan, H.: Low-dose ct denoising with language-engaged dual-space alignment. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3088–3091. IEEE (2024)
6. Chen, Z., Gao, Q., Zhang, Y., Shan, H.: Ascon: Anatomy-aware supervised contrastive learning framework for low-dose ct denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 355–365. Springer (2023)

7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gao, Q., Li, Z., Zhang, J., Zhang, Y., Shan, H.: Corediff: Contextual error-modulated generalized diffusion model for low-dose ct denoising and generalization. IEEE Transactions on Medical Imaging (2023)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
10. Huang, W., Liu, W., Zhang, X., Yin, X., Han, X., Li, C., Gao, Y., Shi, Y., Lu, L., Zhang, L., et al.: Lidia: Precise liver tumor diagnosis on multi-phase contrast-enhanced ct via iterative fusion and asymmetric contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 394–404. Springer (2024)
11. 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 Transactions on Instrumentation and Measurement **71**, 1–12 (2021)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
13. 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)
14. Li, B., Li, X., Zhu, H., Jin, Y., Feng, R., Zhang, Z., Chen, Z.: Sed: Semantic-aware discriminator for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25784–25795 (2024)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. McCollough, C.H., Bartley, A.C., Carter, R.E., Chen, B., Drees, T.A., Edwards, P., Holmes III, D.R., Huang, A.E., Khan, F., Leng, S., et al.: Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. Medical physics **44**(10), e339–e352 (2017)
17. Müller-Franzes, G., Khader, F., Siepmann, R., Han, T., Kather, J.N., Nebelung, S., Truhn, D.: Medical slice transformer: Improved diagnosis and explainability on 3d medical images with dinov2. arXiv preprint arXiv:2411.15802 (2024)
18. Mussmann, B.R., Mørup, S.D., Skov, P.M., Foley, S., Brenøe, A.S., Eldahl, F., Jørgensen, G.M., Precht, H.: Organ-based tube current modulation in chest ct. a comparison of three vendors. Radiography **27**(1), 1–7 (2021)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y.: Coser: Bridging image and language for cognitive super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25868–25878 (2024)
21. Wang, D., Fan, F., Wu, Z., Liu, R., Wang, F., Yu, H.: Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising. Physics in Medicine & Biology **68**(6), 065012 (2023)

22. Wang, R., Heimann, A.F., Tannast, M., Zheng, G.: Cyclesgan: A cycle-consistent and semantics-preserving generative adversarial network for unpaired mr-to-ct image synthesis. Computerized Medical Imaging and Graphics **117**, 102431 (2024)
23. Wang, R., Zheng, G.: Cycmis: Cycle-consistent cross-domain medical image segmentation via diverse image augmentation. Medical Image Analysis **76**, 102328 (2022)
24. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
25. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Generative adversarial networks for noise reduction in low-dose ct. IEEE transactions on medical imaging **36**(12), 2536–2545 (2017)
26. Yu, L., Shiung, M., Jondal, D., McCollough, C.H.: Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols. Journal of computer assisted tomography **36**(4), 477–487 (2012)
27. Zhang, J., Chao, H., Xu, X., Niu, C., Wang, G., Yan, P.: Task-oriented low-dose ct image denoising. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 441–450. Springer (2021)
28. Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Disease-informed adaptation of vision-language models. IEEE Transactions on Medical Imaging (2024)
29. Zhang, J., Gong, W., Ye, L., Wang, F., Shangguan, Z., Cheng, Y.: A review of deep learning methods for denoising of medical low-dose ct images. Computers in Biology and Medicine p. 108112 (2024)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
31. Zhang, X., Li, T.P., Zhao, X.: Boosting single image super-resolution via partial channel shifting. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 13177–13186 (2023)
32. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)