

---

# Diffusion based Zero-shot Medical Image-to-Image Translation for Cross Modality Segmentation

---

**Zihao Wang**

Athinoula A. Martinos Center for Biomedical Imaging  
Massachusetts General Hospital and Harvard University  
02129, Boston, US.

**Yingyu Yang**

Inria centre at Université Côte d’Azur  
06902 Valbonne, France.

**Yuzhou Chen**

Department of Computer and Information Sciences at Temple University  
19122, Philadelphia, US.

**Tinting Yuan**

Institute of Computer Science at Georg-August-University of Göttingen  
37073 Göttingen, Germany.

**Maxime Sermesant**

Inria centre at Université Côte d’Azur  
06902 Valbonne, France.

**Hervé Delingette**

Inria centre at Université Côte d’Azur  
06902 Valbonne, France.

**Ona Wu**

Athinoula A. Martinos Center for Biomedical Imaging  
Massachusetts General Hospital and Harvard University  
02129, Boston, US.

Cross-modality image segmentation aims to segment the target modalities using a method designed in the source modality. Deep generative models can translate the target modality images into the source modality, thus enabling cross-modality segmentation. However, a vast body of existing cross-modality image translation methods relies on supervised learning. In this work, we aim to address the challenge of zero-shot learning-based image translation tasks (extreme scenarios in the target modality is unseen in the training phase). To leverage generative learning for zero-shot cross-modality image segmentation, we propose a novel unsupervised image translation method. The framework learns to translate the unseen source image to the target modality for image segmentation by leveraging the inherent statistical consistency between different modalities for diffusion guidance. Our framework captures identical cross-modality features in the statistical domain, offering diffusion guidance without relying on direct mappings between the source and target domains. This advantage allows our method to adapt to changing source domains without the need for retraining, making it highly practical when sufficient labeled source domain data is not available. The proposed framework is validated in zero-shot cross-modality image segmentation tasks through empirical comparisons with influential generative models, including adversarial-based and diffusion-based models.

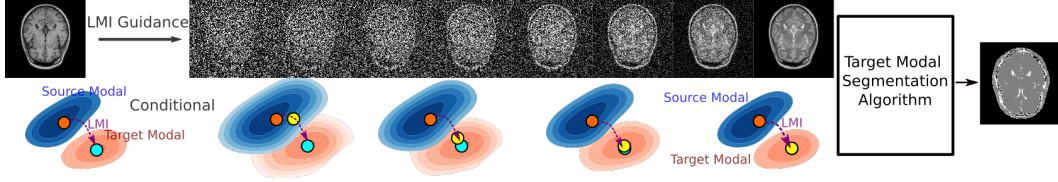


Figure 1: Schematic diagram shows the LMI-guided diffusion for zero-shot cross-modal segmentation. The blue and orange contours are source and target distributions. The blue dot in the orange contour represents the target datapoint of the source datapoint (orange dot in the blue contour) in the source distribution. LMIDiffusion uses explicit statistical features (LMI) to navigate the next step (yellow dot), providing continuous guidance (yellow dot) from start to finish. In the end, the translated image can be segmented using arbitrary segmentation methods that were trained only on the target modality.

## 1 Introduction

Leveraging existing tools to handle data across multiple modalities presents both practical benefits and inherent challenges. For example, in the medical imaging field, numerous resources are readily available for image segmentation within certain modalities, such as 3D T1-weighted MRI (referred to as T1w). However, for other modalities like Proton Density-weighted (PDw MRI), there are difficulties. It would be more resource-efficient to employ segmentation tools initially designed for T1w images on PDw images, instead of hunting for extra segmentation tools crafted specifically for the PDw modality. One promising avenue to navigate this issue is zero-shot cross-modality image translation.

Several methods grounded on generative adversarial networks (GAN) [25, 12, 3, 23, 20, 19, 2] have been put forth to tackle the image translation challenge. In these techniques, the generative model typically models the destination modality straightaway, thereby ensuring translation authenticity [1, 5]. Notably, designs based on GAN often incorporate intricate adversarial structures and demand the crafting of modality-specific loss functions tailored to diverse translation endeavors [13, 21]. While GAN-based translations can function without matching datasets during the training phase, they still hinge on data from the original domain. This dependence can pose difficulties, especially when there’s a scarcity of source domain data, complicating the equilibrium of cycle consistency training [24, 18]. This issue becomes even more pronounced when only a scant amount of samples is present for the target modality.

Recent studies [8, 7, 10, 4] have unveiled that score-based generative models outperform GAN-based models. Muzaffer *et al.* [14] introduced SynDiff, a framework with cycle consistency and dual diffusion for enhanced semantic alignment. Nonetheless, this method entails a doubled computational overhead, necessitates pre-training a generator to assess associated source images, and mandates source domain data, which deviates from the zero-shot learning paradigm. Meanwhile, Meng *et al.* [13] recommended harnessing a Diffusion Model (termed SDEdit) [17] to effectuate image translation via zero-shot learning. In contrast with GAN-based models, SDEdit exhibits superiority by its streamlined model configuration and a more straightforward loss function. However, a limitation of SDEdit emerges in the cross-modality translation context. It is predicted on perturbation-based guidance, which inherently assumes that both the original and destination modalities can be influenced uniformly by noise. Such an assumption often proves invalid in cross-modality image translation, for instance, transitioning from MRI PDw to T1 imaging.

To address the hurdles posed by zero-shot learning in cross-modality translation, our approach leverages statistical feature-wise homogeneity to condition a diffusion model tailored for zero-shot cross-modality image translation. This strategy leverages a connection between the origin and the target modalities, capitalizing on their localized statistical features for accomplishing cross-modality image transition via the diffusion paradigm. Fig. 1 illustrates our statistical feature-driven diffusion model (*LMI*: Locale-based Mutual Information), which performs zero-shot image translation for cross-modality image segmentation. Our proposition is an unsupervised Zero-shot Learning framework capable of navigating translations amidst previously unseen modalities.

## 2 Method

### 2.1 Diffusion Model for Cross-modality Image Translation

The cross-modality image translation can be tackled by score-matching frameworks for generating target data (represented as  $F$ ) based on source data  $G$ . This is followed by employing a perturbed source domain  $G$  to condition the step-by-step diffusion process [13, 14, 15, 26, 11, 4]. Yet, when viewed through the lens of zero-shot learning for image translation, the data from the source domain remains inaccessible during the learning phase. However, it is posited that the local statistical attributes between the source and target modalities bear a semblance. Maximizing Mutual Information (MI) has been validated as a potent strategy to equip neural constructs with the capability to model non-linear mappings [9]. To capture those shared representations for image translation, we propose using MI to model the local statistical features in the denoising process.

### 2.2 Local-wise Mutual Information

To distill semantic information from the dataset for conditioning, it is necessary to translate the raw data into statistical features as encapsulated by MI. Given an image  $X$ , for point,  $x_i \in X$  at position  $i$ , the local-wise statistical information at  $i$  can be captured through the probability density function (PDF)  $p_{\delta_{x_i}}(\cdot)$  of the neighborhood area  $\delta_{x_i}$  of  $x_i$ .

Concerning other data points, represented as  $x_j$ , that reside within the neighborhood,  $\delta_{x_i}$ , of  $x_i$ , the statistical features can be ascertained via the PDF:  $p_{\delta_{x_j}}(\cdot)$ , with  $j$  being an element of  $\delta_i$ . The local-wise MI ( $LMI$ ) from image  $X$  to image  $Y$  at point  $x_i$  is defined through:

$$LMI_{\delta}(x_i, y_j) = \sup \iint p_{\delta}(x, y) \log \frac{p_{\delta}(x, y)}{p_{\delta_{x_i}}(x)p_{\delta_{y_j}}(y)} dx dy, \forall y_j \in \delta_{x_i}. \quad (1)$$

We employ Eq. 1 as a guidance to condition every training step of diffusion. This is realized by evaluating the  $LMI(x_0, y_t), t \in [0, T]$  throughout the training steps of the score network.

**Property 1.** *The upper bound of the LMI from  $X$  to  $Y$  at location  $i$  is:  $LMI_{\delta}(x_i, y_i) \leq LMI_{\delta}(x_i, x_i)$ , which is the optimum informative match between  $X$  and  $Y$  at point  $x_i$ .*

**Property 2.** *The translation error for the LMI Diffusion generation is  $\lim_{\Delta LMI \rightarrow 0} \Delta \mathbb{E} \hat{F} = 0$ .*

In the supplemental section, we prove the properties 1 and 2. Property 1 underscores that  $LMI$  attains statistical congruence under the condition:  $p_{\delta}(X) = p_{\delta}(Y), \forall \delta \in X$ . Consequently, throughout the training iterations, the  $LMI$  consistently peaks at an identical locale between  $X_0$  and  $X_t$  in training steps. Yet, when  $p_{\delta}(X) \neq p_{\delta}(Y)$ , the  $LMI$  achieves the local maximum at  $j$ , which is located in the neighborhood  $\delta_{y_i}$  of  $y_i$ . The process of MI determination can be expedited by converting iterative mutual information computations into tensor-based operations. Such tensor operations can further benefit from speed enhancements through techniques like memory duplication and parallel reduction when executed on a GPU [6].

### 2.3 Conditioning the Diffusion through the LMI

During the training phase, we can guide the noise perturbation procedure by incorporating the  $LMI$  between a data point, denoted as  $F$ , and its perturbed counterpart,  $F_t$ , into the score network, symbolized as  $s_{\theta}$ . This is realized by adapting the training objective of score matching as:

$$\arg \min_{\theta} \mathbb{E}_{t \sim \mathcal{U}(0, T)} \{ \mathbb{E}_{x_0 \sim p_0(x)} \mathbb{E}_{x_t \sim q_{\sigma_t}(x_t, x_0)} [ \frac{1}{2} | s_{\theta}(\hat{F}_t, LMI(F; F, F_t), t) - \frac{\partial \log p_{\sigma_t}(x_t | x_0)}{\partial x_t} |^2 ] \} \quad (2)$$

During the sampling phase, we can integrate the suggested conditioning operator into the SDE:

$$d\hat{F}_t = -\frac{d\sigma_t^2}{dt} s_{\theta}(\hat{F}_t, LMI(G; G, \hat{F}_t), t) dt + \sqrt{\frac{d\sigma_t^2}{dt}} d\mathbf{w}_t, \quad t_{T \rightarrow 0} \in [0, T] \quad (3)$$

and use a naive Euler-Maruyama solver to integrate the SDE. The sampled images can be segmented directly by the tool designed solely for the target domain image segmentation.

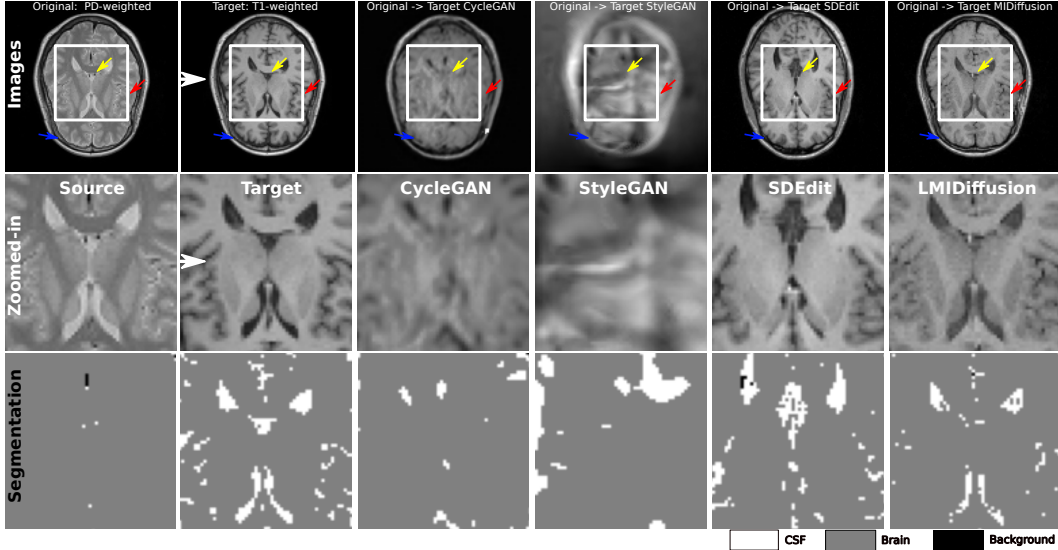


Figure 2: Qualitative evaluation of different models’ translation results. The first two rows show target and original modality images, with close-ups of ROIs, followed by transformations from CycleGAN, StyleGAN, SDEdit, and LMIDiffusion. The subsequent row displays binarized segmentation results in the ROIs using a 3 clusters K-Means method for segmentation, trained solely on the target modality.

### 3 Experiment and Result

**Dataset** we utilize the *IXI dataset* dataset [16] to demonstrate the efficacy of the proposed cross-modality method for image segmentation. This dataset comprises 600 pre-aligned multi-modality images from healthy subjects. We undertake translation tasks between PD and T1w modalities. From a subset of the *IXI dataset*, a training set consisting of 300 slices (drawn from 100 subjects) and a testing set consisting of 75 slices (drawn from 25 subjects) were formulated.

**Experiment** We compare our LMIDiffusion (**unsupervised**) with a few-shot learning-based CycleGAN (**supervised**) translation model [28], a GAN inversion-based approach [22] (**unsupervised**) and a diffusion-based method SDEdit [13] (zero-shot **supervised** learning with SDE-based perturbing guidance). The CycleGAN (**supervised**) in this experiment will be allowed to see both the full target domain dataset and a small group (about 11% of dataset *IXI dataset*) of the source domain dataset. The second baseline is a GAN inversion-based approach (**unsupervised**). A StyleGAN2-ADA [22] is allowed to see the target domain training data. The out-domain guided generation is performed through 5000 steps of optimization of inversion in the latent space of the trained StyleGAN2-ADA [27]. An extra network is introduced in the generation step for inversion. The SDEdit [13] (**unsupervised**) uses a distribution perturbation guidance for diffusion. Our score network structure is the same as the UNet-like score matching network[13], which was optimized through an Adam optimizer with a learning rate of  $3e - 4$ . The model was trained on an NVIDIA DGX Station with four Tesla V100 GPUs, over 300K iterations. For the segmentation backend, we employed the K-Means (5 clusters) algorithm, which was trained exclusively on the target modality. Subsequent segmentation was performed on the translation results obtained from the various methodologies mentioned above.

**Result** Fig. 2 shows both the translation and segmentation outcomes from various models. The bottom rows provide a clear illustration of segmentation results based on translated images. Notably, the segmentation derived from LMIDiffusion translations closely mirrors the segmentation ground truth. While direct segmentation using a method trained on the target image is not feasible for the source domain image, our translation facilitates this, yielding commendable segmentation results on the translated image. Our approach not only offers maximal resemblance to the translation target (PDw) but also retains the superior anatomical features of the original modality (T1w). Although the CycleGAN is trained with supervised (few-shot) data, it fails to produce high-quality images. This shortcoming can be attributed to GAN-based models’ difficulty in discerning relationships between source and target modalities when training data from both domains is limited. On the other hand,

Table 1: Quantitative evaluation of the translation performance for different methods.

PDw→T1	CycleGAN[28]	StyleGAN[22]	SDEdit [13]	LMIDiffusion
Dice Score	$0.85 \pm 0.07$	$0.66 \pm 0.10$	$0.82 \pm 0.06$	<b><math>0.88 \pm 0.05</math></b>
PSNR	$18.88 \pm 1.26$	$10.15 \pm 1.49$	$15.96 \pm 1.54$	<b><math>20.22 \pm 1.43</math></b>
SSIM	$0.27 \pm 0.04$	$0.06 \pm 0.03$	$0.50 \pm 0.06$	<b><math>0.69 \pm 0.06</math></b>

while SDEdit does produce images in the PDw domain, it compromises the anatomical features of the original domain, rendering it less effective for zero-shot image segmentation.

We present the Dice score, PSNR and SSIM values computed based on different models in Table 1. It is evident that the LMIDiffusion achieves a leading average Dice score of  $0.88 \pm 0.05$ . This is closely followed by the CycleGAN-based method with a score of  $0.85 \pm 0.07$ . The SDEdit technique demonstrates suboptimal performance with a Dice score of 0.82. In contrast, the GAN-inversion method (StyleGAN) yields a result of 0.66, emphasizing its challenges in cross-modality image translation. In terms of translation quality metrics, LMIDiffusion leads the pack, boasting PSNR and SSIM values of 20.22 and 0.69, respectively. CycleGAN follows suit with a PSNR of 18.88 and SSIM of 0.27, while SDEdit reports a PSNR of 15.96 and SSIM of 0.50. The StyleGAN method, consistent with its lower Dice score, presents the lowest PSNR of 10.15 and SSIM of 0.06, further evidences the limitation of the GAN-inversion approach for zero-shot image translation task.

## 4 Conclusion

We propose a novel method for zero-shot cross-modality image translation with application in cross-modality image segmentation. For challenging of zero-shot learning, our method outperforms existing GAN-based and diffusion-based methods regarding translation quality and zero-shot segmentation performance. The results show that the proposed LMI-guided diffusion is a promising approach for cross-modality image translation-based segmentation in a zero-shot learning way. The performance of the model can be potentially improved by introducing one-shot or few-shot learning to refine the LMI computing for diffusion guidance.

## A Appendix

*Property 1.*

$$LMI_{\delta}(x_i, y_j) = \sup \iint p_{\delta}(x, y) \log \frac{p_{\delta}(x, y)}{p_{\delta_{x_i}}(x)p_{\delta_{y_j}}(y)} dx dy \quad (4)$$

$$= \sup \left( \iint p_{\delta}(x, y) \log \frac{p_{\delta}(x, y)}{p_{\delta_{y_j}}(y)} dx dy - \iint p_{\delta}(x, y) \log p_{\delta_{x_i}}(x) dx dy \right), \forall y_j \in \delta_{x_i} \quad (5)$$

Applying Bayes' theorem we obtain:

$$= \sup \left( \int p_{\delta_{y_j}}(y) \int p_{\delta_{x_i|y_j}}(x|y) \log p_{\delta_{x_i|y_j}}(x|y) dx dy \right. \quad (6)$$

$$\left. - \iint p_{\delta}(x, y) \log p_{\delta_{x_i}}(x) dx dy \right), \forall y_j \in \delta_{x_i} \quad (7)$$

$$= \sup (h_{\delta_{x_i}}(x) - h_{\delta_{x_i|y_j}}(x|y)), \forall y_j \in \delta_{x_i} \quad (8)$$

$$\leq \sup (h_{\delta_{x_i}}(x)), \forall y_j \in \delta_{x_i} \quad (9)$$

where  $h$  is entropy that measures the informative of the random variables. We can have the following corollaries:

**LMI bound of Forward SDE:** If and only if  $x_i = y_j$ :  $LMI_{\delta}(x_i, x_i) = h_{\delta_{x_i}}(x) - h_{\delta_{x_i|x_i}}(x|x) = h_{\delta_{x_i}}(x)$ , which is the upper bound; therefore for the forward SDE:  $LMI_{\delta}(x_i, y_j) \leq LMI_{\delta}(x_i, x_i), \forall y_j \in \delta_{x_i}$ .

**LMI bound of Reverse SDE:** If  $x_i \neq y_j, x_i \neq z_j$ :  $LMI_\delta(x_i, y_j) = h_{\delta_{x_i}}(x) - h_{\delta_{x_i|y_j}}(x|y)$ , then the  $LMI_\delta(x_i, y_j)$  is the supremum in  $\delta$  neighborhood, bounded by  $LMI_\delta(x_i, x_i)$ :  $LMI_\delta(x_i, z_j) \leq LMI_\delta(x_i, y_j) < LMI_\delta(x_i, x_i), \forall y_j, z_j \in \delta_{x_i}$   $\square$

*Property 2.* For cross-modality data translation, the reverse SDE for  $t_{T \rightarrow 0} \in [0, T]$  is:

$$d\hat{F}_t = -\frac{d\sigma_t^2}{dt} s_\theta(\hat{F}_t, LMI(G; G, \hat{F}_t), t)dt + \sqrt{\frac{d\sigma_t^2}{dt}} d\mathbf{w}_t \quad (10)$$

Its solution is:

$$\hat{F}_t = \hat{F}_0 - \int_0^t \frac{d\sigma_s^2}{ds} s_\theta(\hat{F}_s, LMI(G; G, \hat{F}_s), s)ds + \int_0^t \sqrt{\frac{d\sigma_s^2}{ds}} d\mathbf{w}_s \quad (11)$$

Take the expectation on both side of Eq. 11:

$$\begin{aligned} \mathbb{E}\hat{F}_t &= \mathbb{E}\hat{F}_0 - \mathbb{E} \int_0^t \frac{d\sigma_s^2}{ds} s_\theta(\hat{F}_s, LMI(G; G, \hat{F}_s), s)ds + \mathbb{E} \int_0^t \sqrt{\frac{d\sigma_s^2}{ds}} d\mathbf{w}_s \\ &= \hat{F}_0 - \int_0^t \mathbb{E} \frac{d\sigma_s^2}{ds} s_\theta(\hat{F}_s, LMI(G; G, \hat{F}_s), s)ds \end{aligned} \quad (12)$$

Thus, the expectation of generation error is:

$$\Delta \mathbb{E}\hat{F}_t = - \int_0^t \frac{d\sigma_s^2}{ds} \mathbb{E} \left[ s_\theta(\hat{F}_s, LMI(G; G, \hat{F}_s), s) - s_\theta(\hat{F}_s, LMI(F; F, \hat{F}_s), s) \right] ds \quad (13)$$

Eq. 13 shows that the expectation of generation error is the accumulated expectation of the conditioning error of score model  $s_\theta$  between the training and testing steps.

Without loss of generality, assuming that the score model  $s_\theta$  satisfied *additivity* and *homogeneity* for the guidance  $LMI$ :

$$s_\theta(\hat{F}_t, LMI(\cdot, t), t) = \alpha(\hat{F}_t, t) LMI(\cdot, t) + \beta(\hat{F}_t, t) \quad (14)$$

Substituting Eq. 14 into Eq. 13:

$$\Delta \mathbb{E}\hat{F}_s = - \int_0^t \frac{d\sigma_s^2}{ds} \mathbb{E} \left[ \alpha(\hat{F}_s, s) LMI(G; G, \hat{F}_s) + \beta(\hat{F}_s, s) - \right. \quad (15)$$

$$\left. \alpha(\hat{F}_s, s) LMI(F; F, \hat{F}_s) - \beta(\hat{F}_s, s) \right] ds \quad (16)$$

$$= - \int_0^t \frac{d\sigma_s^2}{ds} \alpha(\hat{F}_s, s) \mathbb{E} \left[ LMI(G; G, \hat{F}_s) - LMI(F; F, \hat{F}_s) \right] ds \quad (17)$$

Denote:  $\Delta LMI = LMI(G; G, \hat{F}_s) - LMI(F; F, \hat{F}_s)$ , the error for the expectation of  $LMI$ -guided generation is:

$$\Delta \mathbb{E}\hat{F}_s = - \int_0^t \frac{d\sigma_s^2}{ds} \alpha(\hat{F}_s, s) \Delta LMI ds \quad (18)$$

Thus,  $\lim_{\Delta LMI \rightarrow 0} \Delta \mathbb{E}\hat{F}_s = 0$   $\square$

## References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to edit the embedded images? In *Proc. of the IEEE CVPR*, June 2020.
- [2] Moab Arar, Yiftach Ginger, Dov Danon, Amit H. Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proc. of the IEEE CVPR*, June 2020.

- [3] Karim Armanious and *et al.*. MedGAN: Medical image translation using GANs. *Comput Med Imaging Graph*, 79:101684, 2020.
- [4] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Non-uniform diffusion models, 2022.
- [5] Andrew Brock, Theodore Lim, J.M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In *ICLR*, 2017.
- [6] John Cheng, Max Grossman, and Ty McKercher. *Professional CUDA c programming*. John Wiley & Sons, 2014.
- [7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models, 2021.
- [8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.
- [9] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.
- [10] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models, 2022.
- [11] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.
- [12] Vasant Kearney and *et al.*. Attention-aware discrimination for mr-to-ct image translation using cycle-consistent generative adversarial networks. *Radiology. Artificial Intelligence*, 2(2), 2020.
- [13] Chenlin Meng and *et al.*. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.
- [14] Muzaffer Ozbey and *et al.*. Unsupervised medical image translation with adversarial diffusion models. *IEEE Trans. Med. Imag.*, pages 1–1, 2023.
- [15] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *arXiv preprint arXiv:2207.14626*, 2022.
- [16] Ayed Samy-Safwan Silva Santiago, Lorenzi Marco. IXI sample dataset, 2022.
- [17] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.
- [18] Hao Tang, Hong Liu, and Nicu Sebe. Unified generative adversarial networks for controllable image-to-image translation. *IEEE Trans. Image Process*, 29:8916–8929, 2020.
- [19] Hisatoshi Toriya, Ashraf Dewan, and Itaru Kitahara. Sar2opt: Image alignment between multi-modal images using generative adversarial networks. In *IGARSS 2019*, pages 923–926. IEEE, 2019.
- [20] Leichen Wang, Bastian Goldluecke, and Carsten Anklam. L2R GAN: Lidar-to-radar translation. In *Proc. of the ACCV*, 2020.
- [21] Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. Inner-ear augmented metal artifact reduction with simulation-based 3d generative adversarial networks. *Computerized Medical Imaging and Graphics*, 93:101990, 2021.
- [22] Tianyi Wei and *et al.*. E2Style: Improve the efficiency and effectiveness of stylegan inversion. *IEEE Trans. Image Process*, 31:3267–3280, 2022.
- [23] Jelmer M Wolterink and *et al.*. Deep MR to CT synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer, 2017.
- [24] Wenju Xu and Guanghui Wang. A domain gap aware generative adversarial network for multi-domain image translation. *IEEE Trans. Image Process*, 31:72–84, 2022.
- [25] Chao Yang, Taehwan Kim, Ruizhe Wang, Hao Peng, and C.-C. Jay Kuo. Show, attend, and translate: Unsupervised image translation with self-regularization and attention. *IEEE Trans. Image Process*, 28(10):4845–4856, 2019.

- [26] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations, 2022.
- [27] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV 2020*, pages 592–608, Cham, 2020. Springer.
- [28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE ICCV*, 2017.