Cross-Modal CXR-CTPA Knowledge Distillation using latent diffusion priors towards CXR Pulmonary Embolism Diagnosis

Noa Cahan[1,†[0000−0002−7264−9541]], Meshi Sizikov[2,†[0009−0000−5505−8611]], and Hayit Greenspan[1,3[0000−0001−6908−7552]]

[1] Faculty of Engineering, Tel Aviv University, Tel-Aviv, Israel. noa.cahan@gmail.com
[2] Blavatnik School of Computer Science and AI, Tel Aviv University
[3] Biomedical Engineering and Imaging Institute, Department of Radiology, Icahn School of Medicine at Mt Sinai, NY, USA

Abstract. Pulmonary Embolism (PE) is a life-threatening condition. Computed tomography pulmonary angiography (CTPA) is the gold standard for PE diagnosis, offering high-resolution soft tissue visualization and three-dimensional imaging. However, its high cost, increased radiation exposure, and limited accessibility restrict its widespread use. In this work, we aim to introduce faster diagnosis opportunities by using 2D chest X-ray (CXR) data. CXR provides only limited two-dimensional visualization and is not typically used for PE diagnosis due to its inability to capture soft tissue contrast effectively. Here, we develop a novel methodology that distills knowledge from a trained CTPA-based teacher classifier model embedding to a CXR-based student embedding, by feature alignment - leveraging paired CTPA and CXR features as supervision, which can be readily acquired. This enables us to train without requiring annotated data. Our approach utilizes a latent diffusion model to generate CTPA-based PE classifier embeddings from CXR embeddings. In addition, we show that incorporating cross-entropy loss together with the corresponding loss of the teacher-student embeddings increases performance, bringing it close to clinical-level performance. We show state-of-the-art AUC in a PE categorization task using only the initial CXR input. This approach broadens the diagnostic capabilities of CXRs by enabling their use in PE classification, thereby extending their applicability beyond traditional imaging roles. The code for this project is available: https://github.com/meshims/Cross-Modal_CXR-CTPA_Knowledge_Distillation

Keywords: Generative models · Cross-modal Knowledge Distillation · Pulmonary embolism diagnosis.

† These authors contributed equally to this work.

1 Introduction

Accurate and timely diagnosis of pulmonary embolism (PE) is critical to reducing mortality and morbidity [1]. While computed tomography pulmonary angiography (CTPA) offers detailed three-dimensional visualization and remains the golden standard for PE diagnosis, its widespread adoption is often limited by high costs, increased radiation exposure, and logistical constraints [8,17]. On the other hand, chest X-rays (CXR) are ubiquitously available and low-cost but traditionally fall short in detecting subtle features indicative of PE [22].

Cross Modality Knowledge Distillation (CMKD) techniques have been successful in various medical imaging tasks [23,33,25,28] and generative cross-modal translation has also effectively been applied in the field [27,7,10,32]. However, as of now, there are very few works specifically combining diffusion models with CMKD in medical imaging. Most diffusion models are primarily focused on generative image tasks, such as image synthesis or denoising, and knowledge distillation is typically applied using standard machine learning models (such as teacher-student networks). A closely related application is X-ray to CT generation [16,2,21,6,30], which similarly leverages cross-modality generative capabilities, though typically without incorporating explicit distillation frameworks. In contrast, our approach employs a diffusion prior that operates on 1D embeddings tailored for classification tasks, rather than on voxel-based image synthesis as done in these generation methods.

Recent advancements in generative AI and cross-modal data translation have opened new avenues for leveraging complementary information from different imaging modalities. To this end, we assembled a unique dataset of 900 patients suspected of having PE, each with a paired 3D CTPA and 2D CXR. We use the cross-modality paired scans to train a novel diffusion-based framework that capitalizes on the rich diagnostic insights captured in CTPA scans by transferring this knowledge to enhance the analysis of CXR images. By aligning and matching the latent representations derived from paired CTPA and CXR examinations, our approach aims to bridge the gap between these modalities and improve the performance of PE classification using only CXR data.

Our contributions in this work are as follows:

1. *Innovative CMKD Method.* We introduce a novel approach that leverages diffusion models to translate CXR latent representations into corresponding CTPA embeddings. This method preserves high fidelity and diagnostic relevance, as evidenced by both quantitative metrics and improved PE classification performance using the generated data. To the best of our knowledge, this is the first study to utilize real paired CXR and CTPA data for CXR-to-CTPA CMKD.
2. *State-of-the-art Classification Results of PE in CXR.* By generating synthetic CTPA embeddings from CXR embeddings, our approach significantly enhances the performance of PE identification in CXRs, leading to improved classification accuracy and increased specificity in PE diagnosis.

3. *Generalizability.* The proposed method demonstrates potential for broader applications in medical imaging, paving the way for more accessible and advanced diagnostic tools.

2 Methods

Fig. 1 depicts the suggested pipeline. Our methodology is constructed from three sequential phases. We begin with the development of two single-modality encoders. One encoder processes CTPA data, and the other handles CXR data. These encoders initially serve as baseline classification models. More crucially, they function as feature extractors, capturing the meaningful latent representations of each data modality (Fig. 1.(A)). The next phase of our research involves training the CXR encoder to align its latent representation with the CTPA's. To this end, we apply a novel approach that involves freezing both encoders and deploying an additional generative network to synthesize a CTPA-like latent representation from the CXR representation. This method is inspired by the initial phase of DALL-E 2 [24], where a diffusion prior is used to generate a CLIP image embedding given a text caption (Fig. 1.(B)). Finally, we apply the generated CTPA embeddings to solve the PE classification problem (Fig. 1.(C)).

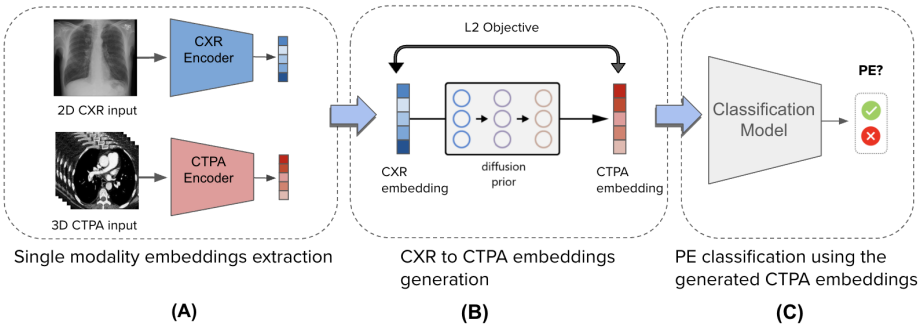

Fig. 1. Our three-phase method for PE classification from chest X-ray data. (A) Two single modality encoders are trained separately for the CTPA and X-ray dataset. (B) We align the X-ray encoding to their matching CTPA ones using a generative diffusion prior model. The embeddings alignment is achieved using an L_2 objective. (C) The generated CTPA embeddings are used to solve the PE classification problem.

2.1 Unimodal Encoders

CTPA Encoder Our 3D CTPA dataset is limited in sample size and includes a single binary label for the whole volume. These challenges guided us in selecting the CTPA classifier architecture. This architecture choice follows SOTA

methods for CTPA PE classification [13] and leverages powerful 2D pre-trained backbones. Fig. 2 illustrates the two-step process of our CTPA model. First, we employ a pretrained 2D image encoder to process axial slices of the 3D CT scan. This model acts as a slice feature extractor, processing each CT slice into a latent embedding, while capturing spatial relationships within each slice. We pass the whole volume in a sequential slice-by-slice manner and extract the embedding vector from each slice, resulting in an array of size: #2D slices x embedding size. For this model we chose PENet [14], a 2D model for PE classification. Given the absence of 2D labels in our dataset, this model cannot undergo fine-tuning and must therefore remain frozen. Subsequently, a 2D sequential model is applied to the sequence of processed slices for our final prediction, effectively capturing temporal information across the 3D volume. Specifically, for this model we use DINOv2 [20], which is one of the most powerful existing 2D foundation model. DINOv2 is a self-supervised distillation method applied on Vision Transformers (ViTs) [26]. The model's self-supervised ViT features contain explicit information about the semantic segmentation of an image and the extracted features can be used for many downstream tasks. We treat each embedding as a projected patch from an image, therefore we use DINOv2 but without the patch projection layer. *Preprocessing:* We resized the CTPA scans to a spacial resolution of 128×128 with varying number of axial slices. All scans were resampled to a standard uniform voxel spacing of 1 mm in all dimensions. The pixel values were converted to Hounsfield units (HU) and windowed to the range of $(-100HU, +900HU)$, to enhance the lower and upper boundaries of the area of interest. Finally, the scans were cropped to the region of the lung area using lung segmentation.

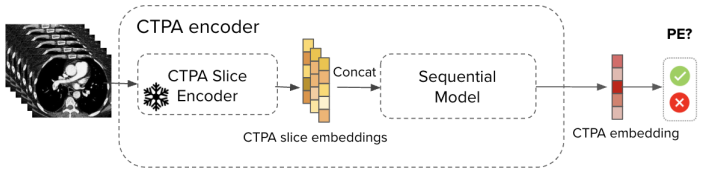

Fig. 2. CTPA encoder model. The pipeline is constructed from two networks: 1. The first one is a 2D pretrained "of-the-shelf" image encoder. The 3D volume is fed through this network slice by slice and concatenated to create a 2D representation of the 3D volume. 2. A 2D sequential model for the final output prediction.

CXR Encoder We evaluated several pre-trained models for the CXR encoder, using them both as fixed feature extractors without fine-tuning and by fine-tuning them for the specific task of PE classification: 1. *RadImageNet* [19], a ResNet-50 based model pretrained on millions of radiologic images. 2. *Biomed-CLIP Vision encoder* [31]. The BiomedCLIP visual encoder, part of the Biomed-CLIP foundation model, was pretrained on a biomedical dataset of 15 million

figure-caption pairs from PubMed Central. It achieves state-of-the-art performance across a variety of medical vision-language processing (VLP) tasks, including cross-modal retrieval, image classification, and visual question answering. 3. *DINOv2*, which was previously discussed in the CTPA encoder section. *Preprocessing:* The 2D CXR scans were normalized and resized to the size of 224 x 224 to match the expected size of the pretrained encoders.

2.2 Cross Modality Knowledge Distillation Using a Diffusion Prior

Our training dataset consists of paired exams – CTPA scans and corresponding CXR images, from the same patient. For each pair, we denote the CTPA and CXR embeddings as z_{CTPA} and z_{CXR}, respectively. Our generative model learns a conditional prior, $P(z_{\text{CTPA}} \mid z_{\text{CXR}})$, that generates CTPA embeddings based on the given CXR embeddings. Specifically, we model the continuous vector z_{CTPA} using a Gaussian diffusion model conditioned on the CXR embedding.

To achieve this, we train a decoder-only Transformer with a causal attention mask on a sequence comprising the CXR image embedding, an embedding for the diffusion timestep, the noised CTPA scan embedding, and a final embedding whose output is used to predict the denoised CTPA embedding. The block diagram for the diffusion prior is depicted in Fig. 3. Rather than predicting the noise residual as in Ho et al. [12], our model directly predicts the unnoised z_{CTPA} - $f_\theta\big(z^{(t)}_{CTPA}, t, z_{CXR}\big)$. A simple L_2 loss between the original CTPA embeddings and predicted prior embeddings is used as an objective for the prior:

$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{t\sim[1,T],\, z^{(t)}_{CTPA}\sim q_t}\left\| f_\theta\big(z^{(t)}_{CTPA}, t, z_{CXR}\big) - z_{CTPA}\right\|^2. \tag{1}$$

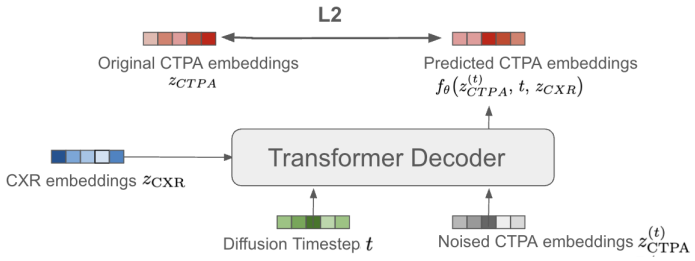

Fig. 3. Diffusion Prior model. During training, a diffusion model takes an CTPA scan embedding to which noise has been added, the matching CXR embedding, and an embedding of the current time step. The system learns to use the CXR embedding to remove the noise in successive time steps. At inference, it generates a CTPA embedding by starting with pure noise and a CXR embedding and removing noise iteratively according to that embedding.

2.3 Dataset

Our cohort consisted of 898 patients from a single center, each of whom underwent both a chest X-ray and a CTPA —within a 24-hour period. Board-certified radiologists annotated the CT volumes with a binary label indicating the presence or absence of pulmonary embolism (PE). The CTPA scans were de-identified and, based on the radiologists' reports, were categorized by experts to indicate whether PE was present, though no assessments of severity or detailed segmentations were provided. For model evaluation, we employed stratified sixfold cross-validation. Out of the total samples, 305 (34%) were identified as positive for PE. This work was approved by the institutional review board (IRB), with informed consent waived by the committee. All methods and experiments were conducted in accordance with the relevant guidelines, regulations, and the Declaration of Helsinki.

2.4 Training Strategy and Details

Throughout the diffusion-prior training, both the CTPA and CXR encoders were kept frozen (i.e., no finetuning was performed). To address the issue of limited data, we began by pretraining the diffusion model on RSPECT — the largest publicly accessible annotated dataset for pulmonary embolism (PE), which includes more than 7,000 3D CTPA scans [5]. Because RSPECT contains only CTPA volumes, we used synthetic CXR images or digitally reconstructed radiographs (DRRs) for conditioning, generated via the DiffDRR [9] Python package for differentiable X-ray rendering. All classification models were trained with a batch size of 2 using the AdamW optimizer, at a learning rate of 1×10^{-6}. For the diffusion-prior design, we adopted a Transformer architecture with a width of 2048 and 24 layers. During inference, we applied Analytic DPM [3] with 64 strided sampling steps. At this phase, the batch size was set to 64, and the learning rate was 1.1×10^{-4}. Training was conducted on a single NVIDIA RTX A5000 GPU and completed within a few hours. Our model implementation was adapted from DALLE-2 [24], with various custom modifications.

3 Experiments and Results

In the following, we detail the results of applying our proposed generation approach to enhance PE classification, including quantitative and qualitative evaluations of the generated samples. Additionally, we conduct an ablation study to investigate the effects of different model architecture choices.

3.1 Using Generated Embeddings to Enhance PE Classification

We present the PE classification results obtained using the generated embeddings. We evaluated performance by feeding the generated latent representations into the matching layer of the CTPA model. We compare the performance

against our baseline results which include using a single modality for classification. Additionally, we compare them against several CMKD teacher-student methods, in which the X-ray encoder weights are trained to mirror the CTPA embeddings. We compare to the classic CKMD method by Hinton et. al [11], as well as more recent CKMD methods [29,15,28]. In this model the CTPA encoder remains frozen. Since no prior studies have attempted PE classification from CXRs, we do not perform a comparative analysis with other existing methods. Table 1 presents the results for this classification task. First, we note the baseline results, where using the original CTPA data achieves an AUC of 0.858 [95% CI: 0.77-0.94], while relying solely on CXR data yields a considerably lower AUC of 0.691 [95% CI: 0.54-0.77]. When utilizing the CKMD methods, it is evident that all methods significantly improve results compared to the original CXR data. Among these approaches, our diffusion-prior model stands out. Using generated embeddings for classification, the model reaches an AUC of 0.824 [95% CI: 0.72-0.93], coming close to the benchmark set by true CTPA data. Moreover, compared to all other methods, our diffusion-prior model has significantly fewer trainable parameters, offering both greater computational efficiency and easier optimization.

Table 1. Comparing the average AUC scores (with a 95% confidence interval) of the PE classification from CXR, CTPA and CMKD methods on our test set.

Model	AUC	Accuracy	Specificity	Sensitivity	#Trainable Params
Baselines					
CTPA only classifier	0.858 [0.77-0.94]	80.77	80.00	82.61	459M
CXR only classifier	0.691 [0.54-0.77]	71.79	70.91	73.91	126M
Cross Modality Knowledge Distillation Methods					
Classic CKMD [11]	0.77 [0.65-0.84]	70.59	68.25	77.27	126M
Discom-kd [15]	0.75 [0.62-0.84]	72.1	73.02	69.56	350M
AFT-KD [29]	0.68 [0.55-0.81]	67.44	66.66	69.56	126M
ProtoKD [28]	0.8 [0.67-0.9]	74.42	80.95	56.52	126M
Our	0.824 [0.72-0.93]	73.27	73.02	73.91	49M

All AUC results are statistically significant with Delong-ROC test, with $p \leq 0.05$

3.2 Mapping the Latent Space of Generated Samples

In Fig. 4 we present the visualization results of the t-SNE [18] of the latents generated by the diffusion prior compared to their corresponding ground truth samples from the test set. We compare the embeddings at the beginning (Fig. 4.(A)), middle (Fig. 4.(B)) and end (Fig. 4.(C)) of training stages. As can be seen from the results, by the end of the training, the generated samples closely follow the ground-truth samples.

3.3 Ablation study on Model Architecture Choices

We conducted ablation experiments, in which the architecture components and training techniques are compared to measure their effect. The results are sum-

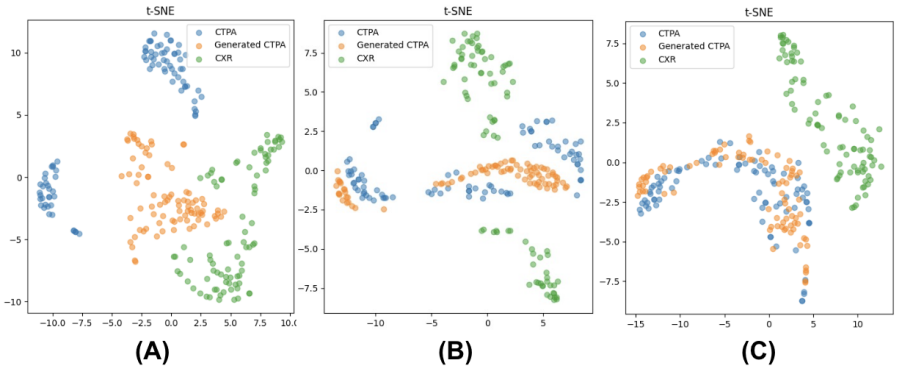

Fig. 4. t-SNE visualizations results of generated latents compared to the ground truth samples on the test set in the beginning (A), middle (B) and end of training (C). The ground truth CTPA latent samples are marked in blue, the generated CTPA latent samples are marked in orange and the original CXR latent samples in green.

marized in Table 2. *Different CXR Embeddings.* We compare the distillation performance using various CXR (or student) embeddings, derived from three distinct models. These models serve as the basis for our training process, providing the initial embeddings used as conditioning in our diffusion model. Detailed descriptions of these models can be found in Section 2.1. RadImageNet and DINOv2 achieved similar results and outperformed the BiomedCLIP Vision encoder. *Embeddings Size.* We investigated the model's sensitivity to the vector size of embeddings extracted from unimodal classifiers. In addition to the 512 baseline dimension we also tested 1024. Our results demonstrate a degradation in terms of AUC when enlarging the embedding size. *Loss.* Beyond the traditional correspondence loss of L_2, we examined the addition of a binary cross-entropy (BCE) classification loss and the combination of the two. As can be seen, including this loss boosts our performance. In addition, we test three other losses - attention-based feature matching loss [29], prototype knowledge distillation which can not only distillates the pixel-wise knowledge of multi-modality data to single-modality data, but can also transfer intra-class and inter-class feature variations and contrastive loss that leverages both positive and negative correspondence [4]. *Effect of pretraining.* Table 2 also demonstrates the effect of pretraining on the RSPECT dataset. Incorporating this additional pretraining step significantly improved our performance.

4 Conclusion and Discussion

The proposed method, which integrates CMKD with diffusion models, significantly advances PE classification from CXRs, delivering state-of-the-art performance. Extensive experiments confirm that our approach outperforms traditional baselines in PE classification. As part of future work, we plan to improve

Table 2. Classification results of the synthesized test set samples using different metrics for different model variants.

Model Variant	AUC	Accuracy	Specificity	Sensitivity
CXR model				
BiomedCLIP	0.7 [0.57-0.85]	70.93	74.60	60.87
RadImageNet	0.766 [0.63-0.9]	79.07	85.71	60.87
DINOv2	0.758 [0.63-0.88]	72.093	73.02	69.56
Embedding Size				
d=1024	0.645 [0.49 - 0.8]	67.45	69.84	60.87
Loss				
$L_2 + \lambda L_{BCE}$	0.795 [0.68-0.92]	77.91	82.54	65.22
$L_2 + \lambda AFTKD$ [29]	0.8 [0.68-0.91]	75.58	74.6	78.26
$L_2 + \lambda ProtoKD$ [28]	0.706 [0.58-0.81]	66.27	66.66	65.21
Contrastive loss [4]	0.66 [0.52-0.79]	67.44	69.84	60.87
Pretraining				
RSPECT Pretraining	0.8 [0.67-0.91]	76.744	77.77	73.91
Full				
Full	0.824 [0.72-0.93]	73.27	73.016	73.913

the clinical interpretability of our framework by incorporating Grad-CAM visualizations to highlight anatomically relevant features and better understand failure cases. Additionally, the method's potential to generalize to other cross-modality applications paves the way for faster diagnoses, earlier treatment response, and more accessible advanced diagnostic tools, all while reducing the need for large labeled datasets.

Acknowledgments. This study was funded by the Israel Science Foundation (ISF), grant no. 20/2629.

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

References

1. Agnelli, G., Becattini, C.: Acute pulmonary embolism. New England Journal of Medicine **363**(3), 266–274 (2010). https://doi.org/10.1056/NEJMra0907731, pMID: 20592294
2. Bai, Q., Liu, T., Liu, Z., Tong, Y., Torigian, D., Udupa, J.: Xctdiff: Reconstruction of ct images with consistent anatomical structures from a single radiographic projection image. arXiv preprint arXiv:2406.04679 (2024)
3. Bao, F., Li, C., Zhu, J., Zhang, B.: Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. arXiv preprint arXiv:2201.06503 (2022)
4. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546 vol. 1 (2005). https://doi.org/10.1109/CVPR.2005.202

5. Colak, E., Kitamura, F.C., Hobbs, S.B., Wu, C.C., Lungren, M.P., Prevedello, L.M., Kalpathy-Cramer, J., Ball, R.L., Shih, G., Stein, A., Halabi, S.S., Altinmakas, E., Law, M., Kumar, P., Manzalawi, K.A., Nelson Rubio, D.C., Sechrist, J.W., Germaine, P., Lopez, E.C., Amerio, T., Gupta, P., Jain, M., Kay, F.U., Lin, C.T., Sen, S., Revels, J.W., Brussaard, C.C., Mongan, J.: The rsna pulmonary embolism ct dataset. Radiol Artif Intell **3**(2), e200254 (Mar 2021)
6. Corona-Figueroa, A., Shum, H.P., Willcocks, C.G.: Repeat and concatenate: 2d to 3d image translation with 3d to 3d generative modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2315–2324 (2024)
7. Dorjsembe, Z., Pao, H.K., Odonchimed, S., Xiao, F.: Conditional diffusion models for semantic 3d medical image synthesis. arXiv preprint arXiv:2305.18453 (2023)
8. Ferrara, R., Mansi, L.: Paul suetens (ed): Fundamentals of medical imaging (2nd edition). European Journal of Nuclear Medicine and Molecular Imaging **38**(2), 409–409 (2011)
9. Gopalakrishnan, V., Golland, P.: Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In: Workshop on Clinical Image-Based Procedures. pp. 1–11. Springer (2022)
10. Hamamci, I.E., Er, S., Simsar, E., Tezcan, A., Simsek, A.G., Almas, F., Esirgun, S.N., Reynaud, H., Pati, S., Bluethgen, C., et al.: Generatect: Text-guided 3d chest ct generation. arXiv preprint arXiv:2305.16037 (2023)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
13. Huang, S.C., Huo, Z., Steinberg, E., Chiang, C.C., Lungren, M.P., Langlotz, C.P., Yeung, S., Shah, N.H., Fries, J.A.: Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. arXiv preprint arXiv:2311.10798 (2023)
14. Huang, S.C., Kothari, T., Banerjee, I., Chute, C., Ball, R.L., Borus, N., Huang, A., Patel, B.N., Rajpurkar, P., Irvin, J., Dunnmon, J., Bledsoe, J., Shpanskaya, K., Dhaliwal, A., Zamanian, R., Ng, A.Y., Lungren, M.P.: Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. npj Digital Medicine **3**(1), 61 (2020). https://doi.org/10.1038/s41746-020-0266-y, https://doi.org/10.1038/s41746-020-0266-y
15. Ienco, D., Dantas, C.F.: Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning. arXiv preprint arXiv:2408.07080 (2024)
16. Liu, J., Bai, X.: VolumeNeRF: CT Volume Reconstruction from a Single Projection View . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15007. Springer Nature Switzerland (October 2024)
17. Lo, P., van Ginneken, B., Reinhardt, J.M., Yavarna, T., de Jong, P.A., Irving, B., Fetita, C., Ortner, M., Pinho, R., Sijbers, J., Feuerstein, M., Fabijańska, A., Bauer, C., Beichel, R., Mendoza, C.S., Wiemker, R., Lee, J., Reeves, A.P., Born, S., Weinheimer, O., van Rikxoort, E.M., Tschirren, J., Mori, K., Odry, B., Naidich, D.P., Hartmann, I., Hoffman, E.A., Prokop, M., Pedersen, J.H., de Bruijne, M.: Extraction of airways from ct (exact'09). IEEE Trans Med Imaging **31**(11), 2093–2107 (Nov 2012)
18. van der Maaten, L., Hinton, G.E.: Visualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (2008)

19. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell **4**(5), e210315 (Sep 2022)
20. 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)
21. Paulson, B., Goldshteyn, J., Balboni, S., Cisler, J., Crisler, A., Bukowski, N., Kalish, J., Colwell, T.: Xprospect: Ct volume generation from paired x-rays. arXiv preprint arXiv:2403.00771 (2024)
22. Pfeiffer, D., Pfeiffer, F., Rummeny, E.: Advanced x-ray imaging technology. Recent Results Cancer Res **216**, 3–30 (2020)
23. Qi, X., Wu, Z., Zou, W., Ren, M., Gao, Y., Sun, M., Zhang, S., Shan, C., Sun, Z.: Exploring generalizable distillation for efficient medical image segmentation. IEEE Journal of Biomedical and Health Informatics (2024)
24. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
25. Schaudt, D., von Schwerin, R., Hafner, A., Riedel, P., Späte, C., Reichert, M., Hinteregger, A., Beer, M., Kloth, C.: Leveraging human expert image annotations to improve pneumonia differentiation through human knowledge distillation. Scientific Reports **13**(1), 9203 (2023)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
27. Waibel, D.J., Röell, E., Rieck, B., Giryes, R., Marr, C.: A diffusion model predicts 3d shapes from 2d microscopy images. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
28. Wang, S., Yan, Z., Zhang, D., Wei, H., Li, Z., Li, R.: Prototype knowledge distillation for medical segmentation with missing modality. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
29. Yang, G., Yu, S., Sheng, Y., Yang, H.: Attention and feature transfer based knowledge distillation. Scientific Reports **13**(1), 18369 (2023)
30. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10619–10628 (2019)
31. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
32. Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., Yu, L.: Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 592–601. Springer (2023)
33. Zou, W., Qi, X., Zhou, W., Sun, M., Sun, Z., Shan, C.: Graph flow: Cross-layer graph flow distillation for dual efficient medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1159–1171 (2022)