# RaD: A Metric for Medical Image Distribution Comparison in Out-of-Domain Detection and Other Applications

Nicholas Konz<sup>1</sup> Yuwen Chen<sup>1\*</sup> Hanxue Gu<sup>1\*</sup>

Haoyu Dong<sup>1</sup> Yaqian Chen<sup>1</sup> Maciej A. Mazurowski<sup>1,2,3,4</sup>

<sup>1</sup> Department of Electrical and Computer Engineering <sup>2</sup> Department of Computer Science

<sup>3</sup> Department of Radiology <sup>4</sup> Department of Biostatistics & Bioinformatics

Duke University, NC, USA

{nicholas.konz, yuwen.chen, hanxue.gu, haoyu.dong151, yaqian.chen, maciej.mazurowski}@duke.edu Code: https://github.com/mazurowski-lab/RaD

## Abstract

Determining whether two sets of images belong to the same or different domain is a crucial task in modern medical image analysis and deep learning, where domain shift is a common problem that commonly results in decreased model performance. This determination is also important to evaluate the output quality of generative models, e.g., image-to-image translation models used to mitigate domain shift. Current metrics for this either rely on the (potentially biased) choice of some downstream task such as segmentation, or adopt task-independent perceptual metrics (e.g., FID) from natural imaging which insufficiently capture anatomical consistency and realism in medical images. We introduce a new perceptual metric tailored for medical images: Radiomic Feature Distance (RaD), which utilizes standardized, clinically meaningful and interpretable image features. We show that RaD is superior to other metrics for out-of-domain (OOD) detection in a variety of experiments. Furthermore, RaD outperforms previous perceptual metrics (FID, KID, etc.) for image-to-image translation by correlating more strongly with downstream task performance as well as anatomical consistency and realism, and shows similar utility for evaluating unconditional image generation. RaD also offers additional benefits such as interpretability, as well as stability and computational efficiency at low sample sizes. Our results are supported by broad experiments spanning four multi-domain medical image datasets, nine downstream tasks, six image translation models, and other factors, highlighting the broad potential of RaD for medical image analysis.

# 1. Introduction



Figure 1. Summary of our main contribution: RaD, a metric designed from the ground up for comparing unpaired distributions of real and/or generated medical images.

Comparing image distributions is crucial in deep learning-driven medical image analysis. Example applications include out-of-domain (OOD) detection [83]; evaluation of image-to-image translation models which convert

<sup>\*</sup>Equal contribution.

images between domains (*e.g.*, modalities or sequences) [4]; quality assessment of images generated to supplement real training data [13, 68]; and others [11].

However, image distribution metrics from general computer vision (*e.g.*, FID [32]) often miss key requirements for medical image analysis. For example, in medical image OOD detection and image-to-image translation, the focus extends beyond just general image quality to *image-level domain adaptation*: ensuring that source-domain images (*e.g.*, from one hospital/site) are compatible with diagnostic models trained on target-domain images from another site, addressing ubiquitous domain shift issues common in medical imaging [6, 19, 27, 54, 58, 62, 87, 90, 91, 95]. Additionally, medical imaging needs metrics that specifically capture anatomical consistency and realism, as well as clinical interpretability [12, 74, 78]. We argue that these specialized needs are overlooked by commonly used metrics for comparing sets of real and/or synthetic medical images.

Unfortunately, the common approach of comparing medical image distributions in terms of some downstream task performance (e.g., segmentation) is biased by the choice of task, and requires costly training and labeling efforts. A task-independent metric that captures general image quality and aligns with expected downstream task performance is therefore preferable. In computer vision, perceptual metrics like FID are commonly used to evaluate image quality relative to target images [7, 32, 75], yet these metrics are based on natural image features. Despite this, many medical image translation [51, 53, 77, 87] and generation [28, 30, 68, 79] works rely on FID (or KID [7]), even though recent findings suggest these metrics may indeed poorly reflect medical image quality [15, 47, 48]. Our experiments further support this issue. Moreover, to date, no studies have proposed interpretable metrics specifically tailored for comparing unpaired medical image distributions, despite the importance of explainability in this subfield.

In this paper, we showcase and address limitations in current metrics for comparing unpaired medical image distributions, proposing a new perceptual/task-independent metric designed for medical imaging. We begin by evaluating "RadiologyFID" (RadFID), a natural extension from FID which uses RadImageNet [60] features instead of ImageNet features, that has surprisingly seen little use. We find that RadFID improves upon prior metrics in some areas, yet lacks in interpretability, stability on small datasets, and other essential qualities. To address these gaps, we introduce Radiomic Distance (**RaD**), a metric leveraging predefined, interpretable radiomic features which are widely used in medical image analysis [24, 49, 84, 94]. RaD offers numerous advantages over learned feature metrics like RadFID and FID, including stronger alignment with downstream task metrics and anatomical consistency, low computational cost, stability for small sample sizes, sensitivity to performance-affecting image distortions, and notably, enhanced interpretability for clinical use due to the inherent meaning of individual radiomic features [5, 17, 64, 70, 92].

We demonstrate these results in key application areas for unpaired medical image distribution comparison, including out-of-domain (OOD) detection and analysis, and the evaluation of image-to-image translation and image generation. Our experiments cover a wide range of medical image datasets and downstream tasks, image translation and generation models, and perceptual metrics. The datasets cover broad medical imaging scenarios including different domains of images, including inter-scanner data such as Siemens and GE T1 breast MRI, inter-sequence data such as T1 and T2 brain MRI, and inter-modality data such as lumbar spine and abdominal MRI and CT, all of which present unique challenges for the explored tasks. **We summarize our contributions as follows:** 

- 1. We highlight the shortcomings of common metrics for medical image distribution comparison (*e.g.*, FID) in meeting the unique requirements of medical imaging.
- We introduce Radiomic Distance (RaD), a taskindependent perceptual metric based on radiomic features, which offers various improvements over prior metrics: (1) alignment with downstream tasks, (2) stability and computational efficiency for small datasets, and (3) clinical interpretability.
- 3. We validate RaD across diverse medical imaging datasets and applications, including practical out-of-domain detection (including proposing a novel, stan-dardized *dataset-level* OOD metric), as well as image-to-image translation and image generation, demonstrating RaD's effectiveness in unpaired medical image distribution comparison.

We have released RaD's code for easy usage at https: //github.com/mazurowski-lab/RaD.

## 2. Related Work

Metrics for Comparing Image Distributions. The standard approach for comparing two unpaired sets/distributions of images  $D_1, D_2 \subset \mathbb{R}^n$  involves defining a distance metric between them that satisfies basic properties (reflexivity, non-negativity, symmetry, and the triangle inequality). In image-to-image translation for example,  $D_2$  represents images translated from a source domain to a target domain, and  $D_1$  is a set of real target domain images. For unconditional generation,  $D_2$  contains generated images, and  $D_1$  serves as a real reference set.

Typically, images are first encoded into a lowerdimensional feature space  $F_1, F_2 \subset \mathbb{R}^m$  via an encoder  $f(x) : \mathbb{R}^n \to \mathbb{R}^m$ . Then, a distance such as the Fréchet distance [21] is computed between these feature distributions. Assuming  $F_1$  and  $F_2$  are Gaussian, this distance becomes

$$d_F(F_1, F_2) = \left( ||\mu_1 - \mu_2||_2^2 + \operatorname{tr} \left[ \Sigma_1 + \Sigma_2 - 2(\Sigma_1 \Sigma_2)^{\frac{1}{2}} \right] \right)^{\frac{1}{2}}.$$
 (1)

The popular **Fréchet Inception Distance (FID)** metric [32] is this distance applied with an Inception v3 network [80] as the encoder. Other metrics which we evaluate include **KID (Kernel Inception Distance)** [7], which uses Maximum Mean Discrepancy (MMD) and is suited for smaller datasets, and **CMMD (CLIP-MMD)** [35], which employs CLIP features with MMD as an alternative to FID.

**Radiology FID (RadFID).** Recent studies suggest that standard perceptual metrics like FID, which are pretrained on natural images, may be unsuitable for medical images [15, 47, 48]. A natural solution may be to use features from a model trained on a large "universal" medical image dataset, such as Radiology ImageNet (RadImageNet) [60]; a RadImageNet-pretrained Inception model can then be instead used to compute FID, as a "**RadFID**". While RadFID has very recently been applied to unconditional generative models for the first time [89], it has not seen widespread adoption, and this work is the first to explore its use for out-of-domain detection and image translation.

**Evaluation of Deep Generative Medical Image Models.** We study two types of generative models in medical imaging: image-to-image translation models and unconditional generative models. Image-to-image translation models, primarily used in radiology to mitigate domain shift between datasets [4, 59], have applications such as inter-scanner translation (e.g., across different manufacturers) [6, 8], inter-sequence translation (e.g., T1 to T2 MRI) [19, 53], and inter-modality translation (e.g., MRI to CT) [67, 87, 90, 95]. Our study includes datasets covering each of these scenarios.

Unconditional generative models, which generate synthetic images from unlabeled real images, are commonly used to supplement medical image datasets, *e.g.*, for the training of diagnostic models [14, 40, 93], including the generation of rare cases [13]. To our knowledge, no previous work has developed task-independent metrics specifically for medical image generation or translation models. The vast majority of works utilize FID [32] (rather than *e.g.*, RadFID), despite its aforementioned limitations.

**Radiomic Features for Medical Image Analysis.** Radiomic features have long been used in diverse medical image diagnostic tasks [24, 49, 84, 94], providing a meaningful, interpretable feature space for analyzing medical images. Applications include cancer screening [36], outcome prediction [1, 16], treatment response assessment [10, 18, 52], and many others. A number of radiomicbased clinical tests have even received FDA clearance [33]. While previous works mainly use learned network features over pre-defined radiomic features for diagnostics [86], few studies have applied radiomics for out-of-domain detection or the evaluation of image translation/generation models, which we show has strong potential.

### 3. Methods

## 3.1. Towards a Metric Designed for Medical Images: Radiomic Feature Distance (RaD)

While RadFID improves on typically-used perceptual metrics for medical images, it lacks interpretability (Sec. 5) and performs less effectively with small samples (Sec. 6). For medical imaging, especially in translation tasks, it's often critical to answer specific questions about how an image's features change—a need less relevant in natural image tasks like style transfer. However, the learned features used in RadFID/FID are difficult to interpret reliably (Sec. 5).

As a more interpretable alternative, we propose the **Ra**diomic Feature **D**istance (**RaD**), which utilizes a space of m = 464 real-valued *radiomic features* of images. The taxonomy and extraction process of these features are illustrated in Fig. 2, and they include image-level features such as basic first-order statistics, and textural statistics such as the gray level co-occurrence matrix [29], gray level run length matrix [23], and gray level size zone matrix [82], combined with a possible pre-filtering step using filters such as wavelet/frequency space conversion, Laplacian of Gaussian (LoG) filters, and others (all computed via the PyRadiomics library [84]). Note that we evaluate the importance of different types of features for RaD in Appendix C.2.1. Importantly, these features are in compliance with definitions provided by the Imaging Biomarker Standardization Initiative [97, 98], mitigating past issues of poorlystandardized radiomics in the field [94].

Each image  $x \in \mathbb{R}^n$  is mapped to its radiomic feature representation  $f_{\text{radio}}(x) \in \mathbb{R}^m$ , and we compute RaD as the Fréchet distance between radiomic feature distributions  $D_1$ and  $D_2$ , applying a logarithmic transformation for stability:

$$\operatorname{RaD}(D_1, D_2) := \log d_F(f_{\operatorname{radio}}(D_1), f_{\operatorname{radio}}(D_2)). \quad (2)$$

We also *z*-score normalize each feature with respect to  $D_1$ . We tested MMD distance as an alternative to Fréchet, but found it less effective (Appendix C.2.2).

#### **3.2.** Downstream Task-Based Image Metrics

The metrics discussed so far compare image distributions in a task-independent way. However, in medical image analysis, task-dependent metrics are often more common, such as



Figure 2. Our extraction process and taxonomy of radiomic image features.

assessing how closely translated/generated images resemble real target images by evaluating downstream tasks like segmentation. For example, in image translation, this often involves training a model on real target domain images and testing it on translated images [37, 85], or vice versa [15, 90]. If the task is segmentation, this approach also measures *anatomical consistency* between source and translated images. Such metrics will therefore serve as important targets for the task-independent metrics which we will evaluate.

While segmentation is the primary task of interest, we also assess other downstream tasks including object detection and classification. We denote downstream performance on a dataset D (e.g., a translated image test set  $D_{s \to t}^{\text{test}}$ ) as  $\operatorname{Perf}(D) \in \mathbb{R}$ , with higher values indicating better performance. Specifically, we use the Dice coefficient for segmentation, mIoU and mAP@[0.5, 0.95] for object detection, and AUC for classification—the latter computed on predicted logits to account for test set class imbalance [57].

#### **3.3. Datasets and Downstream Tasks**

We evaluate a range of multi-domain medical (radiology) 2D image datasets for out-of-domain detection, translation, and generation, covering inter-scanner, inter-sequence, and inter-modality cases (from least to most severe domain differences). The datasets include: (1) breast MRI from Siemens and GE scanners (DBC [50, 73]); (2) brain MRI (T1 and T2 sequences) from BraTS [61]; (3) lumbar spine MRIs and CTs (from TotalSegmentator [88] and in-house MRIs); and (4) abdominal CT and T1 in-phase MRI from CHAOS [39]. Each dataset is split by patient into training, validation, and test sets (details in Table 1), resized to  $256 \times 256$  and normalized to [0, 255]. Example images are in Fig. 3.

The lumbar spine and CHAOS datasets pose especially challenging scenarios due to their relatively small size and



Figure 3. Example images from each dataset, ordered left-to-right with respect to Table 1.

significant differences in visible features and anatomical structures between their pairs of domains.

**Downstream Task Evaluation.** In addition to prior taskindependent metrics (Sec. 2) and RaD (Sec. 3.1), we assess images using auxiliary models trained on downstream tasks (Table 1), as described in Sec. 3.2. These models, trained on target domain data, are tested on various domains, such as target  $(D_t^{\text{test}})$ , source  $(D_s^{\text{test}})$ , source-to-target translations  $(D_{s \to t}^{\text{test}})$ , or others, depending on the experiment. Full details on model training, architecture, dataset creation, and task labels are provided in Appendices B.2 and A.

### 4. Experiments and Results

#### 4.1. RaD for Out-of-Domain Detection

As discussed in the introduction, a common problem in deep learning for medical image analysis is *domain shift*: where when some diagnostic downstream task model is presented with images that were acquired from a site, sequence, or modality different from the one where its training data originated, there may be a performance drop due to the data being *out-of-distribution/out-of-domain* (OOD)

| Abbrev. dataset name | Full name/citation                         | Domains                 | Intra-      | Inter-            | Train/val/test sizes | Downstream tasks                           |
|----------------------|--|-------------------------|-------------|-------------------|----------------------|--|
| Breast MRI           | Duke Breast Cancer<br>MRI (DBC) [50, 73]   | Siemens→GE<br>(T1 MRI)  | Sequence    | Scanner<br>Manuf. | 12K/2.4K/2.6K        | FGT seg., breast seg., cancer classif.     |
| Brain MRI            | BraTS [61]                                 | T1→T2                   | Modality    | Sequence          | 28K/6K/6K            | Tumor seg., tumor detect., cancer classif. |
| Lumbar spine         | TotalSegmentator [88]<br>and in-house MRIs | T1 MRI→CT               | Body region | Modality          | 2K/0.6K/0.6K         | Bone seg.                                  |
| CHAOS                | CHAOS [39]<br>(Abdom. MRI & CT)            | CT→T1 MRI<br>(in-phase) | Body region | Modality          | 1.8K/1.1K/0.6K       | Liver seg.,<br>liver classif.              |

Table 1. Main datasets evaluated in this paper. "Domains" are the source  $\rightarrow$  target domain pairs used, e.g., for image translation.

from the training data [3, 58, 81]. In this section, we will show how RaD is overall superior to prior perceptual metrics for detecting when medical images are OOD.

Perceptual metrics like RaD and FID can help detect whether a new image  $x_{\text{test}}$  is in-distribution (ID) or out-ofdistribution (OOD) relative to a reference ID dataset  $D_{\text{ID}}$ (*e.g.*, some model's training set) without labels. The OOD score  $s(x_{\text{test}}; D_{\text{ID}})$  can be defined as the distance of  $x_{\text{test}}$ 's features from the mean features of  $D_{\text{ID}}$ :

$$s(x_{\text{test}}; D_{\text{ID}}) = ||f(x_{\text{test}}) - \mathbb{E}_{x_{\text{ID}} \sim D_{\text{ID}}} f(x_{\text{ID}})||_2,$$
 (3)

where f is an image feature encoder [69, 76]. OOD performance thus depends on the choice of feature space, so we compare radiomic (*i.e.*, RaD), RadImageNet (*i.e.*, RadFID) and ImageNet (*i.e.*, FID) features for OOD detection.

For each dataset, we use the target domain training set (Table 1) as  $D_{\rm ID}$  and compute the OOD score on the ID and OOD images of the test set, aggregating scores via AUC [20]. Results in the top block of Table 2 and Fig. 4 show that radiomic features outperform learned feature spaces (FID, RadFID) on average, more clearly separating ID and OOD distributions, particularly for breast MRI.

| Metric       | Feature<br>Space                     | Breast<br>MRI                      | Brain<br>MRI                              | Lumbar                             | CHAOS                              | Avg.   |
|--------------|--------------------------------------|------------------------------------|---|------------------------------------|------------------------------------|--|
| AUC:         | ImageNet<br>RadImageNet<br>Radiomics | 0.43<br>0.35<br><b>1.00</b>        | <b>0.91</b><br>0.64<br><u>0.76</u>        | 0.89<br>0.99<br><b>1.00</b>        | 0.94<br><u>0.99</u><br><b>1.00</b> | 0.79<br>0.74<br>0.94   |
| Accuracy:    | ImageNet<br>RadImageNet<br>Radiomics | 0.65<br><u>0.68</u><br><b>0.96</b> | <b>0.73</b><br>0.48<br><u>0.57</u>        | 0.81<br><b>0.98</b><br><u>0.92</u> | 0.84<br><u>0.92</u><br><b>0.95</b> | 0.76   |
| Sensitivity: | ImageNet<br>RadImageNet<br>Radiomics | 0.03<br>0.02<br><b>1.00</b>        | 0.51<br>0.03<br><b>0.95</b>               | 0.37<br><b>1.00</b><br><u>0.71</u> | 0.83<br><u>0.88</u><br><b>1.00</b> | $\left  \begin{array}{c} 0.44 \\ \underline{0.48} \\ 0.92 \end{array} \right $ |
| Specificity: | ImageNet<br>RadImageNet<br>Radiomics | 0.88<br><u>0.92</u><br><b>0.95</b> | <b>0.95</b><br><u>0.94</u><br><u>0.92</u> | 0.96<br><u>0.97</u><br><b>1.00</b> | <u>0.87</u><br><b>1.00</b><br>0.85 | 0.92<br>0.96<br>0.93   |

Table 2. Using different feature spaces for OOD detection.

However, for true practical use, a score threshold  $\hat{s}$  would need to be set to binarily classify ID vs. OOD images *without* a validation set of OOD examples, as AUC simply integrates over all possible thresholds. This is doable if we assume that scores for ID points form a Gaussian distribu-



Figure 4. OOD detection score distributions for in-domain (**blue**) and OOD (**red**) test images for each dataset (columns), using different feature spaces (rows). Computed detection thresholds (Eq. 4) shown as dashed **green** lines.

tion and set  $\hat{s}$  as its 95th percentile:

$$\hat{s} = \sigma_{\rm ID} \Phi^{-1}(0.95) + \mu_{\rm ID},$$
 (4)

where  $\mu_{\text{ID}}$  and  $\sigma_{\text{ID}}$  are the mean and standard deviation of a reference distribution of ID scores  $S_{\text{ID}}$  defined by  $S_{\text{ID}} := \{s(x; D_{\text{ID}} \setminus x) : x \in D_{\text{ID}}\}$ , and  $\Phi^{-1}$  is the normal inverse CDF. We illustrate these computed thresholds for each dataset using different feature spaces in Fig. 4, and show quantitative detection results using them in Table 2. We see that using RaD features over ImageNet or RadImageNet results in noticeably improved average accuracy and sensitivity, and on-par specificity, especially for the challenging subtle domain shift case of breast MRI.

**Out-of-Domain Performance Drop Prediction.** Another closely related question is "*does RaD detect when performance will drop on new data?*" for some downstream task model. We evaluated this for each of the downstream tasks of Table 1, and we see that in almost all cases, there is a drop in average performance on test data that was detected as OOD using the binary threshold approach of Eq. 4, compared to ID performance (full table in Appendix C.4). Additionally, in Appendix C.5 we show that RaD outperforms other metrics in ranking which of different OOD datasets will result in worse downstream task performance.

**Towards Practical Dataset-Level OOD Detection.** We also propose a RaD-based metric for *dataset-level* OOD detection, nRaD<sub>group</sub>, which is formulated to estimate the probability that some new test set  $D_{\text{test}}$  is OOD as a whole, relative to  $D_{\text{ID}}$ . This is particularly designed for the realistic scenario of receiving a new dataset from some outside hospital/site, and wanting an interpretable indication of if the dataset is suitable for some in-domain trained model. While preliminary, we found that this metric scores OOD datasets more consistently than other prior metrics, providing an estimated OOD probability of nRaD<sub>group</sub>  $\simeq 1$  for 3 out of 4 datasets which span a variety of modalities and body regions; see Appendix C.6 for the full details.

**Summary: Practical Medical Image OOD Detection using RaD.** In the interest of practical usage, we provide a step-by-step guide for OOD detection of medical images using RaD in Algorithm 1.

| Algorith        | m 1 Medical Ir   | nage OOD                   | Detection us | ing RaD.                  |
|-----------------|------------------|----------------------------|--------------|---------------------------|
| <b>Require:</b> | Test image set i | $D_{\text{test}}$ , refere | nce ID image | set D <sub>ID</sub> , ra- |
| diami           | a faatura anaada | r f f                      |              |                           |

diomic feature encoder  $f := f_{radio}$ . 1:  $S_{ID} := \{s(x_{ID}; D_{ID} \setminus x_{ID}) : x_{ID} \in D_{ID}\}$ 2:  $\mu_{ID} := \mathbb{E}[S_{ID}]; \sigma_{ID} := \sqrt{Var[S_{ID}]}$ 3:  $\hat{s} = \sigma_{ID} \Phi^{-1}(0.95) + \mu_{ID}$ 4:  $\ell_{test} := \{\mathbf{1}[s(x_{test}; D_{ID}) \ge \hat{s}] : x_{test} \in D_{test}\}$ 5: **return** Binary OOD labels  $\ell_{test}$ 6: **return** (Optional) dataset-level OOD score, nRaD<sub>group</sub>( $D_{test}; D_{ID}$ ) (Appendix C.6)

# 4.2. RaD for Image Translation Evaluation

#### 4.2.1. Translation Models

Unpaired image-to-image translation for medical images, lacking paired data, is challenging and typically relies on adversarial learning. We evaluate a variety of unpaired models: CycleGAN [96], MUNIT [34], CUT [66], Gc-GAN [22], MaskGAN [67], and UNSB [43], each representing diverse techniques such as contrastive learning and style/content disentanglement. All models are trained on source and target domain images from each dataset, with detailed training specifics in Appendix B.1.

#### 4.2.2. Evaluation with Perceptual Metrics

We first evaluate each translation model using perceptual metrics to measure the distance between translated test set source domain images and real test set target domain images. We compare RaD to RadFID, FID, KID, and CMMD for this task, with results in Table 3. While no single model performs best across all datasets due to varying domain shifts, we observe that FID often fails to capture visual quality and anatomical consistency, particularly when there's a high semantic shift between source and target domains, as shown in Fig. 5. For example, FID, KID and CMMD rate MUNIT as best for lumbar spine despite a clear loss of bone structure—shown by MUNIT being the *worst* by segmentation performance in Table 4, which RaD and RadFID capture successfully. Similarly, RaD and RadFID are also more sensitive to performance-affecting image distortions (*i.e.*, simple versions of image translation) than the other metrics (Appendix C.3). This highlights certain limitations of using prior perceptual metrics for medical images.



Figure 5. Translations  $x_{s \to t}^{\text{test}}$  from each translation model (non-top rows) given example inputs  $x_s^{\text{test}}$  (top row).

In Appendix C.1, we additionally show that similar to other perceptual metrics, RaD has the ability to identify the quality of the images generated by unconditional generative models, on all datasets<sup>1</sup>.

### 4.2.3. Correlation with Downstream Task Performance and Anatomical Consistency

Since the key goal of medical image translation is maintaining downstream task performance (*e.g.*, segmentation) and mitigating domain shift, we will now examine whether perceptual metrics can serve as proxies for task performance by correlating perceptual distances with downstream task metrics. We calculate the Pearson correlation r between each perceptual metric and downstream performance across all translation models (Tables 3 and 4).

As shown in Fig. 6, RaD has the strongest (most negative) average correlation with downstream task performance

<sup>&</sup>lt;sup>1</sup>We note that we focused our experimental effort on image translation over generation due to the direct relationship of it with the key problem of domain shift in medical imaging.

|  | Breast MRI   |  |   |  |  | Brain MRI  |  |   |  | I  | Lumbar  |  |  | CHAOS  |  |   |                                      |  |   |  |
|--|--|--|---|--|--|--|--|---|--|--|---|--|--|--|--|---|--------------------------------------|--|---|--|
| Method   | RaD  | RadFID   | FID   | KID  | CMMD   | RaD  | RadFID                                       | FID   | KID  | CMMD   | RaD   | RadFID   | FID  | KID  | CMMD   | RaD   | RadFID                               | FID  | KID   | CMMD   |
| CycleGAN<br>MUNIT<br>CUT<br>GcGAN<br>MaskGAN<br>UNSB | 38.1<br>43.7<br>24.8<br><b>24.6</b><br>48.7<br><b>24.6</b> | 0.26<br>0.29<br><b>0.17</b><br>0.35<br><u>0.19</u> | 107<br>144<br>106<br><u>104</u><br>118<br><b>91</b> | 0.049<br>0.089<br>0.053<br><u>0.040</u><br>0.089<br><b>0.033</b> | <b>0.308</b><br>1.480<br>0.362<br><u>0.322</u><br>0.642<br>0.388 | 33.4<br><u>25.7</u><br><u>33.8</u><br><b>12.1</b><br>27.8<br><b>12.1</b> | 0.06<br>0.05<br>0.13<br>0.04<br>0.06<br>0.08 | 21.7<br><u>21.6</u><br><u>29.4</u><br><b>19.0</b><br>23.5<br>26.0 | 0.004<br>0.006<br>0.012<br>0.003<br>0.008<br>0.010 | 0.378<br>0.388<br><u>0.259</u><br><b>0.239</b><br>0.392<br>0.563 | 6.71<br>9.31<br><b>6.48</b><br><u>6.52</u><br><u>6.64</u><br>6.59 | 0.25<br>0.30<br><b>0.21</b><br>0.25<br>0.27<br><u>0.23</u> | 210<br><b>197</b><br>245<br>226<br>248<br><u>208</u> | 0.161<br>0.151<br>0.206<br>0.161<br>0.217<br>0.172 | 2.950<br>2.317<br>3.373<br>3.300<br>3.237<br>2.579 | 42.8<br>5.41<br>6.84<br><u>6.38</u><br>58.8<br>51.8 | 0.11<br>0.10<br>0.12<br>0.22<br>0.11 | <b>122</b><br>136<br>145<br>141<br>212<br><u>135</u> | 0.051<br>0.073<br>0.083<br><u>0.064</u><br>0.130<br>0.078 | 0.379<br>0.904<br>0.444<br>0.507<br>2.120<br>0.356 |

Table 3. Perceptual/task-independent metrics  $d(D_{s \rightarrow t}^{\text{test}}, D_t^{\text{test}})$  for image translation models.

|  | Breast MRI   | Brain MRI  | Lumbar  | CHAOS   |
|--|--|--|---|---|
|  | $  Dice(\uparrow)   AUC(\uparrow)    Di$   | ice $(\uparrow) \mid mIoU(\uparrow) mAP(\uparrow) \mid A$  | AUC (†)    Dice (†)    Dice   | $e(\uparrow) AUC(\uparrow)$   |
| Method   | Breast FGT   Cancer   T  | 'umor Tumor  | Cancer    Bone    Liv   | ver Liver   |
| CycleGAN<br>MUNIT<br>CUT<br>GcGAN<br>MaskGAN<br>UNSB | $ \begin{bmatrix} 0.871 & \textbf{0.494} & 0.530 \\ 0.832 & 0.201 & 0.511 \\ 0.843 & 0.373 & 0.544 \\ 0.876 & 0.389 & 0.492 \\ 0.809 & 0.164 & 0.441 \\ \textbf{0.881} & 0.308 & \textbf{0.594} \end{bmatrix} ( \  \  \  \  \  \  \  \  \  \  \  \  \$ | $\begin{array}{c ccccc} 0.348 & 0.164 & 0.126 \\ 0.337 & 0.168 & 0.125 \\ 0.303 & 0.159 & 0.133 \\ 0.360 & 0.165 & 0.137 \\ 0.375 & 0.170 & 0.126 \\ 0.353 & \underline{0.169} & \underline{0.135} \\ \end{array}$ | $ \begin{array}{c c c c c c c c c c c c c c c c c c c $                   | 284         0.591           182         0.323           144         0.744           167         0.702           317         0.375           381         0.405 |
| In-domain<br>Out-of-domain                           | 0.883         0.696         0.670         0           0.747         0.446         0.538         0  | 0.442 0.174 0.169<br>0.005 0.152 0.065   | 0.841         0.949         0.8           0.727         0.007         0.0 | 364         0.866           062         0.504   |

Table 4. Downstream task performance metrics  $Perf(D_{s \to t}^{test})$  for image translation models. In-domain and out-of-domain performance shown at the bottom for reference for how susceptible each task is to domain shift, and as *expected* upper and lower performance bounds.

(r = -0.43), followed by RadFID (r = -0.36), while FID, KID, and CMMD are less consistent (r = -0.17, r = -0.17, r = -0.17, r = -0.08), respectively), especially for datasets with larger domain shifts. This is particularly the case for segmentation tasks (which measure anatomical consistency), where, excluding CHAOS, which had low correlations likely due to it generally being a difficult dataset (see Sec. 3.3), RaD and RadFID achieved mean correlations of r = -0.58 and r = -0.70, while FID, KID and CMMD have r = -0.34, r = -0.42, and r = -0.08, respectively.



Figure 6. Pearson correlation of perceptual metrics (vertical axis) (Table 3) with downstream task-based metrics (horizontal axis) (Table 4) for evaluating image translation, taken across all translation models (lower r is better).

Our results also potentially point to the generally best translation models for medical images, which we discuss in Appendix D.2.

#### 5. RaD for Interpretability

In this section, we demonstrate how RaD aids in interpreting differences between large sets of medical images, *i.e.*, understanding the main features that differ between the two sets. The example we will study is interpreting the effects of image-to-image translation models, but this formalism could be applied to any two distributions of images.

At the single-image level, an input image  $x_s$  and output translated image  $x_{s \to t}$  can be converted to feature representations (radiomic or learned)  $h_s := f(x_s)$  and  $h_{s \to t} := f(x_{s \to t})$ , and we can attempt to interpret the feature change vector  $\Delta h := h_{s \to t} - h_s$  and it's *absolute change* counterpart  $|\Delta h|$  defined by  $|\Delta h|^i := |h_{s \to t}^i - h_s^i|$ . At the image distribution level, we can define  $\Delta h := \mu_{s \to t} - \mu_s$  (and similarly  $|\Delta h|$  via  $|\Delta h|^i := |\mu_{s \to t}^i - \mu_s^i|$ ), where  $\mu_s, \mu_{s \to t} \in \mathbb{R}^m$  are the mean vectors of the input and output feature distributions, respectively. In this case, we also define the *individual feature* distributions  $F_s^i := \{h_s^i : h_s \in F_s\}$  and  $F_{s \to t}^i := \{h_{s \to t}^i : h_{s \to t} \in F_{s \to t}\}$ .

In either case,  $\Delta h$  is simply the linear direction vector in feature space between the input and output distributions, analogous with other interpretability works that utilize the linear representation hypothesis [2, 42, 65]. We will next discuss the options and challenges for interpreting  $\Delta h$ , for either learned features or fixed (radiomic) features.

Attempting Interpretability with Learned vs. Radiomic Features. A common method for interpreting directions v in a deep encoder's feature space, such as  $\Delta h$ , is *feature inversion* [55, 63], which uses optimization to find an input image  $x_v$  that aligns with v in feature space, *i.e.*,  $x_v = \operatorname{argmax}_x \operatorname{cossim}(v, f(x))$ . However, we found that doing so using either ImageNet or RadImageNet features resulted in abstract visualizations that lack clear, quantitative insights useful for clinical interpretation (Appendix C.7).

Alternatively, the *individual* features of  $\Delta h$  could be examined statistically with questions like "Which features changed most?" or "Did only a few features account for most changes?" However, concretely interpreting individual *learned* features remains challenging due to the qualitative nature of feature inversion, so we face the same problem.

Thankfully, the clear definitions of radiomic features (Sec. 3.1) allow for clear, quantitative answers to feature interpretability questions, beyond what is possible for learned feature techniques like feature inversion. Here we will exemplify this by interpreting a CUT model trained for lumbar translation, with the following questions.

- 1. Which features changed the most? Sorting features by their values in the  $|\Delta h|$  between input and output image distributions (Fig. 7(a)) identifies those with the highest change, primarily textural/gray-level matrix features, reflecting appearance shifts from MRI to CT (Fig. 7(b)).
- 2. Did only a few features change significantly? Yes— 50% of cumulative feature changes (measured by  $|\Delta h|$ ) are covered by only 37 out of 500 features, indicating a heavy-tailed distribution (Fig. 7(a)).
- 3. Which images changed the most or least? Sorting input/output image pairs (x<sub>s</sub>, x<sub>s→t</sub>) by their absolute feature change ||h<sub>s→t</sub> − h<sub>s</sub>||<sub>2</sub> = ||Δh||<sub>2</sub> (Fig. 7 (c)) shows that the most-changed images have distinct anatomical differences, while the least-changed images mainly differ in texture and intensity (Fig. 7(d)).



Figure 7. Translation interpretability using radiomic features.

These interpretability questions could also help compare translation models on the same dataset and assess model effects on the images, or analyze the domain shift between two certain datasets.

# 6. Further Properties of RaD

Sample Efficiency and Stability. The stability of perceptual metrics at small sample sizes is key for medical image datasets due to them being typically smaller (*e.g.*,  $N \approx 10^2 - 10^3$ ) than natural image datasets (*e.g.*, ImageNet with  $N \approx 10^6$ ). FID generally requires  $N \approx 10^4$  samples for stability [32, 35], which can be prohibitive in this setting. We will next evaluate RaD, RadFID and FID across varying sample sizes N to test for this stability.

We test this under the case of image translation evaluation for the main datasets (Table 1): CycleGAN for breast MRI, GcGAN for brain MRI, CUT for lumbar spine, and MUNIT for CHAOS, respectively. Shown in Fig. 8 left, RaD remains stable even for very small N (down to N = 10), while RadFID and FID diverge as N grows small across all datasets, indicating that these metrics are not suitable for comparing small medical datasets of different sizes.



Figure 8. Left: Sensitivity of RaD, RadFID and FID to sample size N. Metric values (vert. axes) are relative to their highest-N result. **Right:** Computation time for the metrics w.r.t N.

RaD's stability with small N is due to relatively few features dominating its computation (see *e.g.* Fig. 7(a)), meaning the effective dimensionality  $\tilde{m}$  is much smaller than the full  $m \approx 500$ . Thus, RaD's Fréchet distance behaves as though it operates in a lower-dimensional space, enhancing stability even with limited samples.

**Computation Time.** We next compare RaD's computation time to FID/RadFID across sample sizes N, using data parallelism (num\_workers=8) on UNSB-translated BraTS test images. As shown in Fig. 8 right, RaD is faster than FID/RadFID for small-to-moderate sample sizes  $(N \leq 500)$ . For larger N, RaD's compute time grows slightly faster, but both metrics remain efficient across sample sizes, with RaD particularly advantageous for small N.

## Conclusions

Our work covered a range of diagnostic tasks and imaging domains, but is still limited to radiology. Other modalities like histopathology could be explored, though radiomic features may need adjustment. Our interpretability contributions are nascent, and further work is needed to extract more qualitative, yet concrete, insights. Additionally, RaD's absolute values are only meaningful in relative comparisons between models on the same dataset, similar to FID.

Overall, our results show the value of using interpretable, medical-specific feature spaces like radiomic features for comparing unpaired medical image distributions, aiding in tasks such as OOD detection, as well as evaluating image translation and generation models. Future work could explore RaD in other applications, such as weaklyconditioned generative models and non-generative model evaluation. Such features could potentially even assist model training, not just evaluation.

### Acknowledgements

Research reported in this publication was supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB031575. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

- [1] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014. 3
- [2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017. 7
- [3] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018. 5
- [4] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020. 2, 3
- [5] Minji Bang, Jihwan Eom, Chansik An, Sooyon Kim, Yae Won Park, Sung Soo Ahn, Jinna Kim, Seung-Koo Lee, and Sang-Hyuk Lee. An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum. *Translational psychiatry*, 11(1):462, 2021. 2
- [6] Farzad Beizaee, Christian Desrosiers, Gregory A Lodygensky, and Jose Dolz. Harmonizing flows: Unsupervised mr harmonization based on normalizing flows. In *International Conference on Information Processing in Medical Imaging*, pages 347–359. Springer, 2023. 2, 3
- [7] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 2, 3

- [8] Shixing Cao, Nicholas Konz, James Duncan, and Maciej A Mazurowski. Deep learning for breast mri style transfer with limited training data. *Journal of Digital imaging*, 36(2):666– 678, 2023. 3
- [9] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murray, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A.D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. MONAI: An open-source framework for deep learning in healthcare. 2022. 2
- [10] Kenny H Cha, Lubomir Hadjiiski, Heang-Ping Chan, Alon Z Weizer, Ajjai Alva, Richard H Cohan, Elaine M Caoili, Chintana Paramagul, and Ravi K Samala. Bladder cancer treatment response assessment in ct using radiomics with deeplearning. *Scientific reports*, 7(1):8738, 2017. 3
- [11] Heang-Ping Chan, Ravi K Samala, Lubomir M Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Deep learning in medical image analysis: challenges* and applications, pages 3–21, 2020. 2
- [12] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. *NPJ digital medicine*, 5(1):156, 2022. 2
- [13] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical En*gineering, 5(6):493–497, 2021. 2, 3
- [14] Yizhou Chen, Xu-Hua Yang, Zihan Wei, Ali Asghar Heidari, Nenggan Zheng, Zhicheng Li, Huiling Chen, Haigen Hu, Qianwei Zhou, and Qiu Guan. Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 144:105382, 2022. 3
- [15] Yuwen Chen, Nicholas Konz, Hanxue Gu, Haoyu Dong, Yaqian Chen, Lin Li, Jisoo Lee, and Maciej A Mazurowski. Contourdiff: Unpaired image translation with contourguided diffusion models. arXiv preprint arXiv:2403.10786, 2024. 2, 3, 4
- [16] Gary M Clark. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. *Molecular oncol*ogy, 1(4):406–412, 2008. 3
- [17] Sunan Cui, Alberto Traverso, Dipesh Niraula, Jiaren Zou, Yi Luo, Dawn Owen, Issam El Naqa, and Lise Wei. Interpretable artificial intelligence in radiology and radiation oncology. *The British Journal of Radiology*, 96(1150): 20230142, 2023. 2

- [18] Karen Drukker, Hui Li, Natalia Antropova, Alexandra Edwards, John Papaioannou, and Maryellen L Giger. Most-enhancing tumor volume by mri radiomics predicts recurrence-free survival "early on" in neoadjuvant treatment of breast cancer. *Cancer imaging*, 18:1–9, 2018. 3
- [19] Alicia Durrer, Julia Wolleb, Florentin Bieder, Tim Sinnecker, Matthias Weigel, Robin Sandkuehler, Cristina Granziera, Özgür Yaldizli, and Philippe C Cattin. Diffusion models for contrast harmonization of magnetic resonance images. In *Medical Imaging with Deep Learning*, pages 526–551. PMLR, 2024. 2, 3
- [20] Tom Fawcett. An introduction to roc analysis. Pattern recognition letters, 27(8):861–874, 2006. 5
- [21] Maurice Fréchet. Sur la distance de deux lois de probabilité. In *Annales de l'ISUP*, pages 183–198, 1957. 2
- [22] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometryconsistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2427–2436, 2019. 6, 2
- [23] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2): 172–179, 1975. 3
- [24] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radi*ology, 278(2):563–577, 2016. 2, 3
- [25] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015. 2
- [26] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 3
- [27] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 2
- [28] Colin Hansen, Simas Glinskis, Ashwin Raju, Micha Kornreich, JinHyeong Park, Jayashri Pawar, Richard Herzog, Li Zhang, and Benjamin Odry. Inpainting pathology in lumbar spine mri with latent diffusion. arXiv preprint arXiv:2406.02477, 2024. 2
- [29] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973. 3
- [30] Anees Ur Rehman Hashmi, Ibrahim Almakky, Mohammad Areeb Qazi, Santosh Sanjeev, Vijay Ram Papineni, Dwarikanath Mahapatra, and Mohammad Yaqub. Xreal: Realistic anatomy and pathology-aware x-ray generation via controllable diffusion model. *arXiv preprint arXiv:2403.09240*, 2024. 2
- [31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2
- [32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a

two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 3, 8

- [33] Erich P Huang, James PB O'Connor, Lisa M McShane, Maryellen L Giger, Philippe Lambin, Paul E Kinahan, Eliot L Siegel, and Lalitha K Shankar. Criteria for the translation of radiomics into clinically useful tests. *Nature reviews Clinical oncology*, 20(2):69–82, 2023. 3
- [34] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Proceedings of the European conference on computer vision (ECCV), pages 172–189, 2018. 6, 2
- [35] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3, 8, 2
- [36] Yulei Jiang, Alexandra V Edwards, and Gillian M Newstead. Artificial intelligence applied to breast mri for improved diagnosis. *Radiology*, 298(1):38–46, 2021. 3
- [37] Myeongkyun Kang, Philip Chikontwe, Dongkyu Won, Miguel Luna, and Sang Hyun Park. Structure-preserving image translation for multi-source medical image domain adaptation. *Pattern Recognition*, 144:109840, 2023. 4
- [38] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 2
- [39] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 4, 5, 1
- [40] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial intelligence in medicine*, 109:101938, 2020. 3
- [41] Lim Swee Kiat. greentfrapp/lucent: Lucid library adapted for PyTorch, 2021. 6
- [42] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 7
- [43] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *The Twelfth International Conference on Learning Representations*, 2024. 6, 2
- [44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 2
- [45] Nicholas Konz and Maciej A Mazurowski. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *The Twelfth International Conference on Learning Representations*, 2024. 3

- [46] Nicholas Konz and Maciej A Mazurowski. Reverse engineering breast mris: Predicting acquisition parameters directly from images. In *Medical Imaging with Deep Learning*, pages 829–845. PMLR, 2024. 1
- [47] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024. 2, 3, 1
- [48] Nicholas Konz, Yuwen Chen, Hanxue Gu, Haoyu Dong, and Maciej A Mazurowski. Rethinking perceptual metrics for medical image translation. In *Medical Imaging with Deep Learning*, 2024. 2, 3
- [49] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012. 2, 3
- [50] Christopher O Lew, Majid Harouni, Ella R Kirksey, Elianne J Kang, Haoyu Dong, Hanxue Gu, Lars J Grimm, Ruth Walsh, Dorothy A Lowell, and Maciej A Mazurowski. A publicly available deep learning model and dataset for segmentation of breast, fibroglandular tissue, and vessels in breast mri. *Scientific Reports*, 14(1):5383, 2024. 4, 5, 1
- [51] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 632–641. Springer, 2023. 2
- [52] Hui Li, Yitan Zhu, Elizabeth S Burnside, Karen Drukker, Katherine A Hoadley, Cheng Fan, Suzanne D Conzen, Gary J Whitman, Elizabeth J Sutton, Jose M Net, et al. Mr imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of mammaprint, oncotype dx, and pam50 gene assays. *Radiology*, 281(2): 382–391, 2016. 3
- [53] Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zeroshot medical image translation via frequency-guided diffusion models. *IEEE transactions on medical imaging*, 2023. 2, 3
- [54] Mengting Liu, Piyush Maiti, Sophia Thomopoulos, Alyssa Zhu, Yaqiong Chai, Hosung Kim, and Neda Jahanshad. Style transfer using generative adversarial networks for multisite mri harmonization. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27– October 1, 2021, Proceedings, Part III 24, pages 313–322. Springer, 2021. 2
- [55] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5188–5196, 2015. 7, 6
- [56] TorchVision maintainers and contributors. Torchvision: Py-

torch's computer vision library. https://github.com/
pytorch/vision, 2016. 2, 4

- [57] Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010. 4
- [58] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020. 2, 5
- [59] Jake McNaughton, Justin Fernandez, Samantha Holdsworth, Benjamin Chong, Vickie Shim, and Alan Wang. Machine learning for medical image translation: A systematic review. *Bioengineering*, 10(9):1078, 2023. 3
- [60] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315, 0. 2, 3
- [61] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 4, 5, 1
- [62] Gourav Modanwal, Adithya Vellal, Mateusz Buda, and Maciej A Mazurowski. Mri image harmonization using cycle-consistent generative adversarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, pages 259–264. SPIE, 2020. 2
- [63] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 7, 6
- [64] Matthew R Orton, Evan Hann, Simon J Doran, Scott TC Shepherd, Derfel Ap Dafydd, Charlotte E Spencer, José I López, Víctor Albarrán-Artahona, Francesca Comito, Hannah Warren, et al. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: Insights from the tracerx renal study. *Cancer Imaging*, 23(1):76, 2023. 2
- [65] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. 7
- [66] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 6, 2
- [67] Vu Minh Hieu Phan, Zhibin Liao, Johan W Verjans, and Minh-Son To. Structure-preserving synthesis: Maskgan for unpaired mr-ct translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023. 3, 6, 2

- [68] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022. 2
- [69] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2155–2162, 2023. 5
- [70] Amir L Rifi, Inès Dufait, Chaïmae El Aisati, Mark De Ridder, and Kurt Barbé. Interpretability and repeatability of radiomic features: Applied on in vivo tumor models. *IEEE Transactions on Instrumentation and Measurement*, 72:1–7, 2023. 2
- [71] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 2
- [72] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015. 2
- [73] Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Connie E Kim, Sujata V Ghate, Ruth Walsh, and Maciej A Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British journal of cancer*, 119(4):508–516, 2018. 4, 5, 1
- [74] Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111, 2022. 2
- [75] Sagar Saxena and Mohammad Nayeem Teli. Comparison and analysis of image-to-image generative adversarial networks: a survey. arXiv preprint arXiv:2112.12625, 2021. 2
- [76] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019. 5
- [77] Lulin Shi, Yan Zhang, Ivy HM Wong, Claudia TK Lo, and Terence TW Wong. Mulhist: Multiple histological staining for thick biological samples via unsupervised image-toimage translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 735–744. Springer, 2023. 2
- [78] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of imaging*, 6(6):52, 2020. 2
- [79] Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, Ke Yu, and Kayhan Batmanghelich. Hierarchical amortized gan for 3d high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 26(8):3966–3975, 2022.
   2
- [80] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception archi-

tecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

- [81] Gregory Szumel, Brian Guo, Darui Lu, Rongze Gui, Tingyu Wang, Nicholas Konz, and Maciej A Mazurowski. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. *arXiv preprint arXiv:2409.04368*, 2024. 5
- [82] Guillaume Thibault, Bernard Fertil, Claire Laure Navarro, Sandrine Pereira, Pierre Cau, Nicolas Lévy, Jean Sequeira, and Jean-Luc Mari. Texture indexes and gray level size zone matrix. application to cell nuclei classification. In 10th International Conference on Pattern Recognition and Information Processing, 2009. 3
- [83] Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging-a mini review. In Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021, pages 33–38. Springer, 2022. 1
- [84] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21): e104–e107, 2017. 2, 3
- [85] Eugene Vorontsov, Pavlo Molchanov, Matej Gazda, Christopher Beckham, Jan Kautz, and Samuel Kadoury. Towards annotation-efficient segmentation via image-to-image translation. *Medical Image Analysis*, 82:102624, 2022. 4
- [86] Matthias W Wagner, Khashayar Namdar, Asthik Biswas, Suranna Monah, Farzad Khalvati, and Birgit B Ertl-Wagner. Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology*, pages 1–11, 2021. 3
- [87] Zihao Wang, Yingyu Yang, Yuzhou Chen, Tingting Yuan, Maxime Sermesant, Hervé Delingette, and Ona Wu. Mutual information guided diffusion for zero-shot cross-modality medical image translation. *IEEE Transactions on Medical Imaging*, 2024. 2, 3
- [88] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiol*ogy: Artificial Intelligence, 5(5), 2023. 4, 5, 1
- [89] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P. Yung, Ankit B. Patel, and Kristy K. Brock. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024. 3
- [90] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019:*

22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, pages 255–263. Springer, 2019. 2, 3, 4

- [91] Li Yao, Jordan Prosky, Ben Covington, and Kevin Lyman. A strong baseline for domain adaptation and generalization in medical imaging. In *Medical Imaging with Deep Learning*, 2019. 2
- [92] Jing-Yuan Ye, Peng Fang, Zhen-Peng Peng, Xi-Tai Huang, Jin-Zhao Xie, and Xiao-Yu Yin. A radiomics-based interpretable model to predict the pathological grade of pancreatic neuroendocrine tumors. *European radiology*, 34(3):1994– 2005, 2024. 2
- [93] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019. 3
- [94] Stephen SF Yip and Hugo JWL Aerts. Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61 (13):R150, 2016. 2, 3
- [95] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018. 2, 3
- [96] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycleconsistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223– 2232, 2017. 6, 2
- [97] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. Image biomarker standardisation initiative. arXiv preprint arXiv:1612.07003, 2016. 3
- [98] Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. 3

# RaD: A Metric for Medical Image Distribution Comparison in Out-of-Domain Detection and Other Applications

Supplementary Material

# A. Dataset and Task Labeling Details

### A.1. Breast MRI

For breast MRI we use the 2D slices of the precontrast scan volumes from the Duke Breast Cancer dataset [73], using the same train/validation/test splits (by patient) and preprocessing of [47] from the 100 patient volumes with FGT and breast segmentation annotations (see the following paragraph). This results in train/validation/test splits with source, target domain subsplits of size  $\{4096, 7900\}/\{432, 1978\}/\{688, 1890\}$  images.

**FGT and breast segmentation.** FGT (fibroglandular/dense tissue) and breast segmentation masks for this dataset are provided from [50].

**Cancer classification/slice-level detection.** For the slice cancer classification task, we follow the same convention of [46], and label slice images as cancer-positive if they contain any tumor bounding box annotation, and negative if they are at least 5 slices away from any positive slices (ignoring the intermediate ambiguous slices). We then train a basic ResNet-18 [31] (modified for 1-channel input images) as our binary cancer classification model, on the positive and negative slices from the training set's GE scans. The model's evaluation datasets are otherwise unchanged from the other downstream tasks (besides the labels used for the images).

## A.2. Brain MRI

For brain MRI, we utilized the multi-modal brain tumor dataset from the BraTS 2018 challenge [61]. Since the BraTS's own validation set doesn't have masks available, we began by extracting the original shared training set and dividing the patients into training, validation, and test sets with a ratio of 0.7:0.15:0.15 for this paper. Next, we focused on the T1 and T2 sequence volumes along with their corresponding masks. Each slice of the image volume was normalized and saved as 2D PNG files to construct our 2D dataset.

**Tumor segmentation and detection.** The original mask contains multiple classes of segmentation, including: Background (Label 0), Enhancing Tumor (Label 4), Tumor Core (Label 1), Whole Tumor (Label 2), Peritumoral Edema (Label 3). We conducted a binary tumor/not-tumor segmenta-

tion by combing all pixels with label larger than 0. For tumor detection, the tumor bounding box is generated by the smallest box that covers the entire tumor region. For those cases without tumor shown in that slice, we excluded them during model training/validation/testing.

**Cancer classification/slice-level detection.** We also further modify the task into a binary tumor classification task: whether this slice contains tumor or not. For those slices that are near the boundary of the tumor, specifically the 5 slices before and after the tumor presence (switching between positive and negative in each volume), we excluded them from the classification as they are considered ambiguous slices.

## A.3. Lumbar Spine

The CT lumbar spine dataset is obtained from TotalSegmentator [88], and the T1 MRI data is private (to be revealed upon paper acceptance). We split the 2D source and target data in train/val/test as  $\{495, 1466\}/\{175, 409\}/\{158, 458\}.$ 

**Bone segmentation.** We perform binary classification on each pixel to determine whether it includes bone or not. The ground truth masks for MRI are reviewed by experts, while the CT masks are sourced from [88].

## A.4. CHAOS

We extract 2D CT and T1 in-phase MRI slices from the CHAOS dataset [39]. For each domain, we randomly split the data by patient in the ratio of 10:5:5, resulting in the 2D slices for the source and target domains being divided into train/val/test as  $\{1488, 322\}/\{926, 182\}/\{460, 182\}$ .

**Liver segmentation.** Liver masks for both modalities are provided by [39].

**Liver classification.** We assign positive labels to slices which contain the liver and negative labels to those that do not.

# **B. Model Training/Architectural Details**

In this section we describe the training details of all networks in the paper. All experiments were completed on four 48GB NVIDIA A6000 GPUs.

#### **B.1. Translation Models**

All six translation models (CycleGAN [96], MUNIT [34], CUT [66], GcGAN [22], MaskGAN [67], and UNSB [43]) were trained with their default settings (besides being modified to input and output 1-channel images), except for a few exceptions to be described shortly; these settings are shown in Table 5.

| Model         | Training time | Batch size |
|---------------|---------------|------------|
| CycleGAN [96] | 200 epochs    | 4          |
| MUNIT [34]    | 1M iters.     | 1          |
| CUT [66]      | 200 epochs    | 1          |
| GcGAN [22]    | 200 epochs    | 32         |
| MaskGAN [67]  | 200 epochs    | 4          |
| UNSB [43]     | 200 epochs    | 1          |

Table 5. Translation model training details.

The exceptions are that for MUNIT and CUT, training for too long resulted in drastic changes in image content for breast MRI and lumbar so we chose earlier model iterations of 10,000 and 20,000, respectively for MUNIT, and 20 epochs for both for CUT.

#### **B.2.** Downstream Task Models

In this section we describe the architectural and training details of all models trained for the downstream tasks of each dataset (Table 1) on its respective target domain data from the training set. All models are trained with Adam [44] and a weight decay strength of  $10^{-4}$  for 100 epochs.

**Segmentation.** For all segmentation downstream tasks we train a standard UNet [72] with five encoding blocks, at a batch size of 8 with a learning rate of 0.01. The model is trained with equally-weighted cross-entropy and Dice losses, the latter implemented with MONAI [9].

**Object Detection.** For detection downstream tasks, we trained a Faster-RCNN [25] with a batch size of 4 and a learning rate of 0.005. The model is implemented using Torchvision [56], with the number of predicted classes modified to 2. The loss function is the default loss from this built-in model.

**Classification/Slice-level Detection.** For classification tasks we train a standard ResNet-18 [31], modified to take in one-channel inputs and output one logit (as all tasks are binary classification). We use a batch size of 64 and a learning rate of 0.001, with a cross-entropy loss.

# **C.** Additional Experiments

## C.1. Evaluating Unconditional Image Generation

We will now study RaD in evaluating unconditional generative medical image models, similar to FID's typical use. We trained StyleGAN2-ADA [38] with default settings on four single-domain image generation tasks: (1) GE T1 breast MRI, (2) T1 brain MRI (BraTS), (3) lumbar spine CT, and (4) an abdominal CT dataset (CT-Organ [71])<sup>2</sup>.

Following [35], we evaluate each perceptual metric (RaD, RadFID, FID, CMMD, KID) by ranking samples from an early model iteration (**Model A**, kimg = 200) of visibly lower quality against a fully trained model (**Model B**, kimg = 2000), with results and sample generated images shown in Fig. 9. RaD successfully identifies the lower-quality model in all cases, aligning with prior metrics, except for CMMD, which fails for breast MRI.



Figure 9. Top: RaD and other perceptual metrics for evaluating unconditional generative models, comparing a poor model (A) to a better model (B). Bottom: example generated images.

#### C.2. Ablation Studies

#### C.2.1. Radiomic Feature Importance

To better interpret RaD, we assess the importance of different radiomic feature groups (textural/gray-level matrix, wavelet, first-order) by ablation: examining how removing each group affects the main translation model downstream task performance results (Fig. 6). Results are shown in Fig. 10.

Overall, wavelet and first-order features are most crucial for RaD, as excluding them significantly worsens correlation results. Textural features are somewhat important for breast MRI but have limited impact on other datasets. Breast MRI is generally the most sensitive to feature exclusion, suggesting that subtle domain shifts require a broader range of features for accurate analysis.

<sup>&</sup>lt;sup>2</sup>CHAOS was not large enough to train on for high generation quality.



Figure 10. Importance of different radiomic features for RaD. Pearson correlation r between RaD and downstream task performance metrics across all translation models (as in Fig. 6), comparing using standard RaD with all features (top row) to removing certain groups of features (lower rows).

#### C.2.2. Using MMD instead of Fréchet Distance

Perceptual metrics such as CMMD and KID use the MMD (Maximum Mean Discrepancy) distance metric  $d_{\rm MMD}$  [26] over the more common Fréchet distance, due to advantages such as lacking the Gaussianity assumption and being suitable for smaller datasets [7, 35]. Here we will evaluate calculating our proposed RaD distance using MMD (with a standard Gaussian RBF kernel) as

$$\operatorname{RaD-MMD}(D_1, D_2) := d_{\mathrm{MMD}}(f_{radio}(D_1), f_{radio}(D_2)),$$
(5)

instead of via Fréchet distance as  $d_{\rm radio}$  (Eq. (2)). We compare the two metrics (RaD and "RaD-MMD") in terms of (1) how much they correlate with downstream task performance metrics (as in Fig. 6), in Fig. 11, and (2) whether they rank translation models similarly (linearly or non-linearly), in Table 6.

| Corr.<br>type | Breast<br>MRI | Brain<br>MRI | Lumbar | CHAOS |
|---------------|---------------|--------------|--------|-------|
| Pearson       | -0.75         | -0.79        | 0.81   | -0.84 |
| Spearman      | -0.83         | -0.83        | 0.49   | -0.54 |

Table 6. Correlation r of RaD computed with MMD distance (Eq. (5)) with standard Fréchet distance RaD (Eq. (2)), across all translation models.

We first see that RaD-MMD is noticeably inferior to RaD in terms of its negative correlation to downstream task performance (Fig. 11); for all but one task, the correlation r is either close to zero, or in the wrong direction (positive r, as higher perceptual distance should correlate with worse performance, not better). Moreover, RaD-MMD is not consistent in terms of its relationship to standard RaD (Table 6). We hypothesize that these issues could poten-



Figure 11. Pearson correlation r between RaD and downstream task performance metrics across all translation models (as in Fig. 6), comparing using Fréchet or MMD distance for RaD.

tially be due to the dependence of MMD on the choice of kernel, which could require further tuning, or the fact that MMD does *not* have an assumption of Gaussianity unlike the Fréchet distance, which may result in a metric that is too unconstrained.

#### C.3. Robustness of RaD to Image Distortions

In this section we analyze the sensitivity of the perceptual metrics to image corruptions/distortions, including RaD and RadFID, also under the task of image translation. Given the importance of downstream task performance metrics for medical image translation models, as well as the typical increased sensitivity of medical image models to image corruptions compared to natural image models [45], it is important that any distortion or perturbation to a (translated) image that noticeably affects downstream task performance on that image, is also captured by the perceptual metrics.

More concretely, consider some image transformation/distortion  $T : \mathbb{R}^n \to \mathbb{R}^n$ . We model a preferable perceptual metric d as approximately following the inverse proportionality

$$\frac{d(D_t, T(D_{s \to t}))}{d(D_t, D_{s \to t})} \approx \left(\frac{\operatorname{Perf}(T(D_{s \to t}))}{\operatorname{Perf}(D_{s \to t})}\right)^{-1}, \quad (6)$$

evaluated on the test set's target domain images  $D_t$  and translated source-to-target images  $D_{s \to t}$ . In other words, if the perceptual distance increases by some positive multiplicative factor K (implying the distortion made the translated images more distant from the target domain), we would expect the performance to go worse by 1/K, up to a constant of proportionality—the sensitivity of the two metrics to image distortions should match.

We tested this on all datasets for downstream segmentation tasks that were sensitive to such distortions (FGT for breast MRI, tumor for brain MRI, bone for lumbar spine, and liver for CHAOS (Table 4)), for various translation models (CycleGAN, GcGAN, CUT, and MUNIT, respectively). We evaluate simple image distortions Tas Gaussian blurs with positive integer kernel  $k \in \mathbb{N}$  (blurk) or sharpness adjustments of non-negative factor  $\gamma$  (sharpness $\gamma$ ), from TorchVision [56]. We show the results in Fig. 12, plotting  $d(D_t, T(D_{s \to t}))/d(D_t, D_{s \to t})$  and  $\operatorname{Perf}(T(D_{s \to t}))/\operatorname{Perf}(D_{s \to t})$  for each distortion for all proposed and prior perceptual metrics d.



Figure 12. Sensitivity of RaD, RadFID (**blue**) and other perceptual metrics (**red**) to performance-affecting (**green**) distortions of translated images. Metric values (vertical axis) are relative to their un-distorted result ("None").

For breast MRI, we see that the distortion sensitivity of FID and RaD fairly well match the performance sensitivity, with the other perceptual metrics less so. For lumbar spine, both RaD and RadFID follow the performance sensitivity well, while other perceptual metrics are generally not as sensitive. For BraTS, performance is typically not sensitive to distortions, which RaD follows; other perceptual metrics are generally oversensitive, but all still increase when performance decreases due to blurring. CHAOS seems to be a failure mode where no perceptual metric has sensitivity aligning with performance sensitivity, likely due to challenges of training the downstream task model on such a small, and challenging dataset (Sec. 3.3). Overall, the sensitivities of our proposed metrics, especially RaD, align with performance sensitivity best. Other perceptual metrics are not as consistent over all datasets, which aligns with the results of Sec. 4.2.3.

#### C.4. OOD Performance Drop Prediction

In Table 7 we evaluate if images detected as out-of-domain with our OOD score thresholding approach (Sec. 4.1) also result in lowered performance compared to on detected ID cases.

#### C.5. OOD Performance Drop Severity Ranking

Here, we assess how well RaD and other perceptual metrics predict performance drops on out-of-domain (OOD) data. Given a model trained on target domain data  $D_t$  and two new OOD datasets  $D_{OOD,1}$  and  $D_{OOD,2}$ , we examine if a metric d can correctly indicate which OOD dataset will suffer a greater performance drop. Specifically, we test if  $Perf(D_{OOD,2}^{test}) < Perf(D_{OOD,1}^{test})$  aligns with  $d(D_{OOD,2}^{test}, D_t) > d(D_{OOD,1}^{test}, D_t)$ , and vice versa.

We evaluate this scenario with the datasets in Table 1 which possess additional data domains beyond the target domain and default source domain  $D_s$ , namely, BraTS using its T2-FLAIR data [61], and CHAOS using its T1 Dual Out-Phase and T2 SPIR MRI data [39]. We show these task performance vs. perceptual distance agreement results in Table 8 for each type of downstream task, and for each possible pair of  $D_{OOD,1}$  and  $D_{OOD,2}$  for each dataset (T1 MRI and T2 FLAIR MRI for BraTS, respectively, and all 2-combinations of {T1 Dual Out-Phase MRI, T2 SPIR MRI, and CT} for CHAOS). Shown in Tables 9 and 10 are the specific results that generated Table 8.

#### C.6. Towards Dataset-Level OOD Detection

In Sec. 4.1, we showed how RaD/radiomic features can be used for single image-level binary OOD detection. However, a more realistic scenario may be that some new dataset is acquired from an outside hospital/site, and we wish to know if the dataset is generally OOD relative to our own reference ID dataset  $D_{\rm ID}$  that we used to train some downstream task model, to get some idea of how our model will perform on the new dataset  $D_{\rm test}$ . For example, our ID dataset could be breast MRI collected from GE scanners, and the new dataset could potentially have OOD (*e.g.*, Siemens) images. Our goal is therefore to have a metric that returns an (approximately) standardized value if  $D_{\rm test}$ is OOD.

A naive prior approach to this could be to measure the FID or RadFID between  $D_{\rm ID}$  and  $D_{\rm test}$ , but as we will show, such distances are not clearly interpretable due to the distance value being noticeably affected by the specific dataset used, as well as the sample size (Sec. 6). To this end, we propose a RaD-based metric for dataset-level OOD detection which is designed to return 1 (or a value close to it) when the test set is completely OOD.

We do so by considering an ID reference point  $x_{\rm ID} \sim D_{\rm ID}$  and test set point  $x \sim D_{\rm test}$ , both randomly sampled.

|                            |                  | Breast MRI     |   |                |  | Brain MRI      |                |                                     |  |                  | Lumbar  | CHAOS          |                  |
|----------------------------|------------------|----------------|---|----------------|--|----------------|----------------|-------------------------------------|--|------------------|---|----------------|------------------|
|                            | Dic              | e (†)          | . | AUC (†)        |  | Dice (†)       | mIoU (†)       | $\mathrm{mAP}\left(\uparrow\right)$ |  | AUC $(\uparrow)$ | Dice $(\uparrow)$                             | Dice (†)       | AUC $(\uparrow)$ |
| Detected as:               | Breast           | FGT            |   | Cancer         |  | Tumor          | Tur            | nor                                 |  | Cancer           | Bone  | Liver          | Liver            |
| In-Domain<br>Out-of-Domain | n 0.904<br>0.731 | 0.698<br>0.473 |   | 0.670<br>0.535 |  | 0.434<br>0.498 | 0.179<br>0.215 | 0.175<br>0.223                      |  | 0.619<br>0.618   | $\begin{array}{c} 0.856 \\ 0.001 \end{array}$ | 0.848<br>0.129 | 0.789<br>0.463   |

Table 7. Downstream task performance on test points detected as ID vs. OOD using our thresholding method.

|                                     | Segm<br>Brain MRI:<br>trained on T2 | Segmentation (Dice)<br>ain MRI: CHAOS: trained on T1 Dual In-Phase |                  |                      |  | <b>Detection</b><br>Brain MRI:<br>trained on T2 |                  |  | Classification (AUC)<br>Brain MRI:<br>trained on T2 CHAOS: trained on T1 Dual In-Pha |                       |                            |                   |
|-------------------------------------|-------------------------------------|--|------------------|----------------------|--|---|------------------|--|--|-----------------------|----------------------------|-------------------|
|                                     | T1<br>vs. T2 FLAIR                  | T1 DOP<br>vs. T2 SPIR  | T1 DOP<br>vs. CT | T2 SPIR<br>vs. CT    |  | T1 vs. T<br>mIoU                                | 2 FLAIR<br>mAP   |  | T1<br>vs. T2 FLAIR   | T1 DOP<br>vs. T2 SPIR | T1 DOP<br>vs. CT           | T2 SPIR<br>vs. CT |
| RaD<br>RadFID<br>FID<br>KID<br>CMMD | /<br>/<br>X<br>-                    | ×  | \ \ \ \ \<br>\ \ | \$<br>\$<br>\$<br>\$ |  | ×<br>×<br>·                                     | ✓<br>✓<br>×<br>× |  | ✓<br>✓<br>×<br>×   | ✓<br>×<br>×<br>×      | \$<br>\$<br>\$<br>\$<br>\$ |                   |

Table 8. Can RaD predict OOD performance drop severity? For each downstream task type (sub-tables) trained on a given dataset's target domain data  $D_t$  (first row) and for two OOD test sets  $D_{\text{OOD},1}^{\text{test}}$  and  $D_{\text{OOD},2}^{\text{test}}$  (second row), whether  $\text{Perf}(D_{\text{OOD},2}^{\text{test}}) < \text{Perf}(D_{\text{OOD},2}^{\text{test}})$  does ( $\checkmark$ ) or does not ( $\checkmark$ ) correspond to  $d(D_{\text{OOD},2}^{\text{test}}, D_t) > d(D_{\text{OOD},1}^{\text{test}}, D_t)$  and vice-versa. "-" denotes that the given perceptual metric was only negligibly affected by the change of the OOD dataset. "DOP" is "Dual Out-Phase"

|                          | Downs       | tream ta       | sk perfo       | rmance         |              | Perceptual distance metrics |            |                |                |  |  |
|--------------------------|-------------|----------------|----------------|----------------|--------------|-----------------------------|------------|----------------|----------------|--|--|
| $D_{\rm OOD}^{\rm test}$ | Dice        | mIoU           | mAP            | AUC            | RaD          | RadFID                      | FID        | KID            | CMMD           |  |  |
| T1 MRI<br>T2 FLAIR MRI   | 0.005 0.286 | 0.152<br>0.144 | 0.065<br>0.108 | 0.727<br>0.885 | 6.18<br>5.09 | 0.25<br>0.19                | 108<br>117 | 0.089<br>0.088 | 0.179<br>0.394 |  |  |

Table 9. Downstream task performance  $Perf(D_{OOD}^{test})$  (left block) and perceptual distances  $d(D_{OOD}^{test}, D_t)$  (right block) on out-of-domain data  $D_{OOD}^{test}$  (each row) for downstream task models trained on (in-domain) BraTS T2 MRI data  $D_t$ , to supplement Table 8.

|                          | Downst | ream task performance | Perceptual distance metrics |        |     |       |       |  |  |  |
|--------------------------|--------|-----------------------|-----------------------------|--------|-----|-------|-------|--|--|--|
| $D_{\rm OOD}^{\rm test}$ | Dice   | AUC                   | RaD                         | RadFID | FID | KID   | CMMD  |  |  |  |
| T1 Dual Out-Phase MRI    | 0.779  | 0.853                 | 7.87                        | 0.09   | 143 | 0.096 | 0.205 |  |  |  |
| T2 SPIR MRI              | 0.262  | 0.867                 | 7.55                        | 0.20   | 189 | 0.126 | 0.507 |  |  |  |
| CT                       | 0.062  | 0.504                 | 60.6                        | 0.65   | 277 | 0.268 | 1.666 |  |  |  |

Table 10. Downstream task performance  $Perf(D_{OOD}^{test})$  (left block) and perceptual distances  $d(D_{OOD}^{test}, D_t)$  (right block) on out-of-domain data  $D_{OOD}^{test}$  (each row) for downstream task models trained on (in-domain) CHAOS T1 Dual In-Phase MRI data  $D_t$ , to supplement Table 8.

Now, we wish to have a metric that estimates the probability that the test set is OOD. The key insight here is that the higher this probability, the higher the chance that x is OOD, such that it's expected score/distance from  $D_{\rm ID}$ , s(x) (Eq. 3) will in turn be more likely to be larger than that of a typical ID point  $x_{\rm ID}$ . Assuming that OOD points will not be typically *closer* to D than ID points, which is true by the definition of OOD, then the minimum value of this probability is  $\Pr[s(x) > s(x_{\rm ID})] = 0.5$  if  $D_{\rm test}$  is 100% ID (no clear difference between the test set and reference set score distributions), and  $\Pr[s(x) > s(x_{\rm ID})] = 1$  if  $D_{\rm test}$  is 100% OOD.

We then convert this to a metric,  $nRaD_{group}$  (RaD for group-level OOD detection normalized to a fixed range) that

ranges from 0 to 1 with

$$nRaD_{group} := 2(Pr[s(x) > s(x_{ID})] - 0.5).$$
(7)

The final question is then how  $Pr[s(x) > s(x_{ID})]$  can be computed in practice; thankfully, the area under the ROC curve (AUC) by definition is this quantity [20], which can be easily computed, giving

$$nRaD_{group} := 2(AUC[S_{test}, S_{ID}] - 0.5), \qquad (8)$$

where  $S_{\text{test}} := \{s(x) : x \in D_{\text{test}}\}$  and  $S_{\text{ID}}$  is the reference distribution of ID scores,  $S_{\text{ID}} := \{s(x_{\text{ID}}; D_{\text{ID}} \setminus x_{\text{ID}}) : x_{\text{ID}} \in D_{\text{ID}}\}$ , as in Sec. 4.1.

We evaluate  $nRaD_{group}$  for OOD-scoring OOD test sets in Table 11, averaged over 10 randomly sampled test sets of size 100 for each trial, compared to using the FID or RaD-FID between  $D_{\text{test}}$  and  $D_{\text{ID}}$ . While all metrics assign a higher score for the OOD test set than the ID test set, we note that the scale of FID and RadFID OOD test set distance values changes noticeably depending on the dataset, a factor which would be even more pronounced if considering datasets of different sizes, as those metrics can be highly unstable for different sample sizes (Sec. 6). On the other hand, nRaD<sub>group</sub> is  $\approx$  1 for the OOD test set in 3/4 datasets (besides BraTS, due to it generally proving difficult for disentangle the ID and the OOD distributions (Sec. 4.1)), making it a more standardized, interpretable and practical metric. This enables us to posit that a nRaD<sub>group</sub> score of  $\approx$  1 for some new dataset means that the dataset is likely OOD.

We similarly see that for completely ID test sets (Table 12), nRaD<sub>group</sub> is  $\approx 0$  in 3/4 cases. While RadFID does so for 4/4 cases, this doesn't account for the fact that Rad-FID is still highly sensitive to sample size, hurting its interpretable, standardized, realistic use in this case. Finally, we also show ablation studies in these two tables of using ImageNet or RadImageNet features to compute nRaD<sub>group</sub> instead of radiomic features, where we see that using these learned features results in less stable OOD test set distance values.

| Metric         | Breast<br>MRI | Brain<br>MRI | Lumbar | CHAOS |
|----------------|---------------|--------------|--------|-------|
| FID            | 178           | 223          | 277    | 338   |
| RadFID         | 0.22          | 0.35         | 1.23   | 1.64  |
| $nRaD_{group}$ | 1.00          | 0.62         | 0.95   | 1.00  |
| +ImageNet      | 0.01          | 0.84         | 0.75   | 0.94  |
| +RadImageNet   | 0.14          | 0.32         | 0.99   | 0.94  |

Table 11. Dataset-level OOD detection scores for OOD test sets.

| Metric       | Breast<br>MRI | Brain<br>MRI | Lumbar | CHAOS |
|--------------|---------------|--------------|--------|-------|
| FID          | 92            | 73           | 77     | 48    |
| RadFID       | 0.09          | 0.06         | 0.04   | 0.04  |
| nRaDgroup    | 0.00          | 0.04         | 0.07   | 0.44  |
| +ImageNet    | 0.22          | -0.04        | 0.05   | -0.05 |
| +RadImageNet | 0.1           | -0.05        | 0.07   | -0.13 |

Table 12. Dataset-level OOD detection scores for ID test sets.

## C.7. Attempting to Interpret Differences between Medical Image Distributions with Learned Features

We applied feature inversion [55, 63] to visualize  $\Delta h$  using ImageNet and RadImageNet features (via Lucent [41]), for breast MRI and brain MRI UNSB translation models, shown in Fig. 13. While the results hint at general textural and shape changes from translation, they lack clear, quantitative insights useful for clinical interpretation.



Figure 13. Attempts at medical image translation interpretability via learned feature inversion.

# **D.** Additional Discussion

# D.1. Downstream Task Metrics as Image Distribution Distance Metrics

Here we will discuss how segmentation performance metrics are themselves distance functions between two distributions of image features. If the predictions of the downstream task model on its test set and the corresponding ground truth labels/segmentations are taken as "features" of the images, then performance metrics such as Dice segmentation coefficient, IoU, etc. are image distribution metrics that clearly follow the topological requirements of distance metrics: reflexivity, non-negativity, symmetry, and the triangle inequality (Sec 2).

#### **D.2. Best Medical Image Translation Models**

Our comprehensive experiments notably suggest general recommendations for medical image translation models. For severe domain shifts (*e.g.*, lumbar spine and CHAOS), CUT performs best in both downstream tasks and perceptual metrics of RadFID and RaD, likely due to its contrastive learning approach that preserves image structure. For more subtle shifts (breast MRI and brain MRI), GcGAN performs well, with UNSB and CycleGAN also effective for breast MRI. We recommend CUT for high domain shifts and Gc-GAN for moderate shifts.