Rethinking Perceptual Metrics for Medical Image Translation

Nicholas Konz, Yuwen Chen, Hanxue Gu,
Haoyu Dong, Maciej A. Mazurowski, Duke University

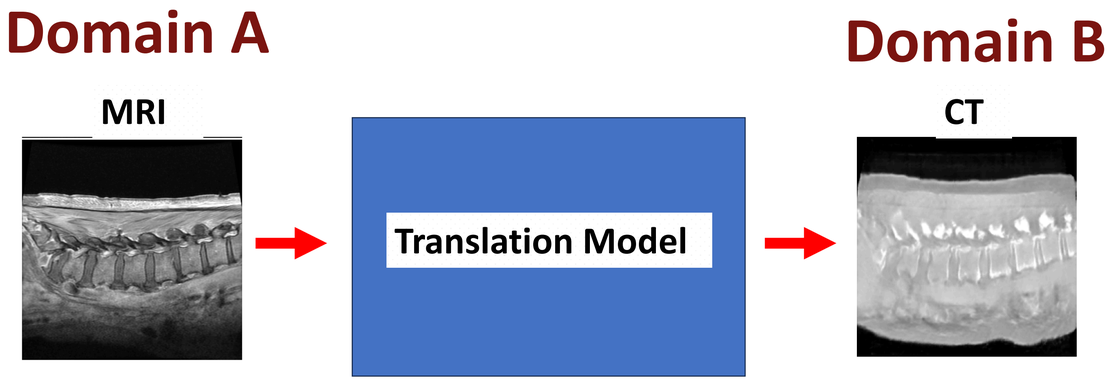
MIDL Paris 2024

Are existing perceptual metrics for image translation models actually useful for medical image translation?

With what criteria should we evaluate them?

Background

- Unpaired image-to-image translation: a common task in medical image computing.



- For example: transforming a lumbar spine MRI **(source domain)** to make it look like a CT **(target domain)**

What are the desiderata for medical image translation?

1. ***Anatomical consistency w.r.t. the input image.***
   1. Commonly measured with a **segmentation model trained in the target domain**, and applied to translated images.
      1. **A standard metric in medical image translation papers.**
   2. Limitations: need for labels and resources to train the segmentation model, **bias towards the task**/object, etc.

2. ***Overall perceptual quality and realism.***
   1. Typically measured with **perceptual metrics** from **mainstream computer vision:** FID, IS, etc.
   2. These metrics are **task-agnostic**, **but may fail to capture local or global anatomical consistency and realism** in medical images!

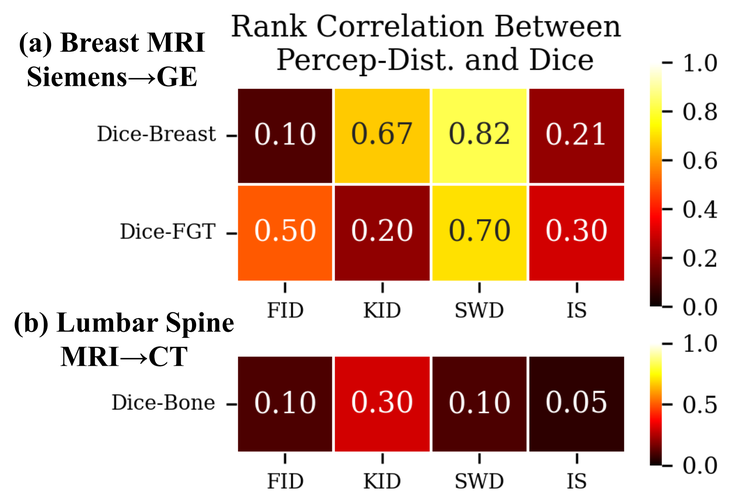
Contact us

- nicholas.konz@duke.edu
- maciej.mazurowski@duke.edu
- @nick_konz on twitter

Are common perceptual metrics useful for medical image translation?

- Do **any** task-agnostic **perceptual** metrics also reliably correlate with **anatomical consistency**?

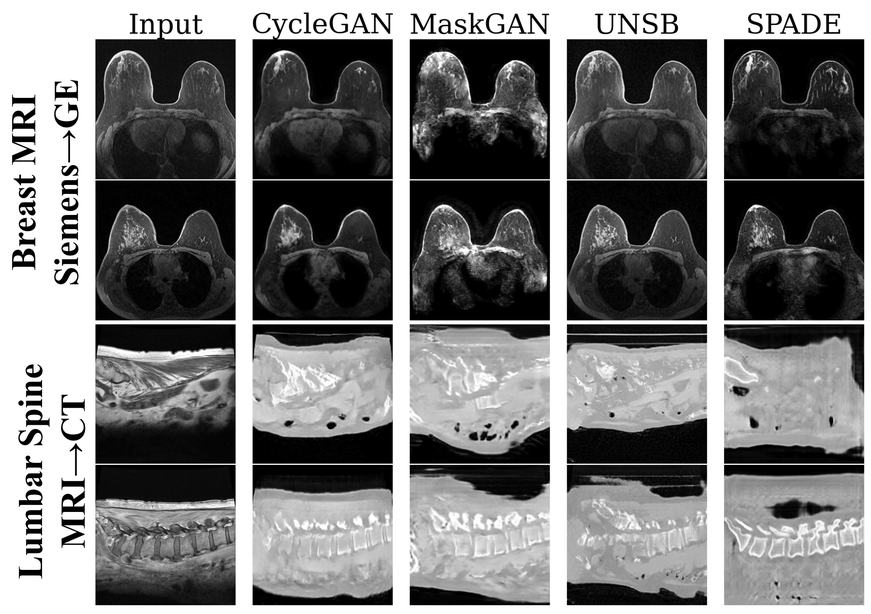
Sadly, **no.** ☹



Breast MRI Siemens→GE Translation							Lumbar Spine MRI→CT Translation				
	Dice (↑)		Perceptual Metrics				Dice (↑)	Perceptual Metrics			
Method	Breast	FGT	FID*	KID	SWD	IS	Bone	FID*	KID	SWD	IS
None	0.927	0.651	144	0.069	705	2.58	0.007	323	0.300	1553	**2.93**
CycleGAN	0.934	0.529	**107**	**0.049**	556	2.73	0.229	210	**0.161**	960	2.29
MaskGAN	0.865	0.277	118	0.089	1037	**3.00**	0.158	248	0.217	1114	2.22
UNSB	0.934	0.646	156	0.079	756	2.46	0.138	**208**	0.172	**932**	2.14
SPADE†	**0.950**	**0.707**	119	0.067	**500**	2.91	**0.942**	251	0.242	1359	2.29

Table 1: Quantitative results for both translation tasks. Best and runner-up models are shown in bold and underlined according to each metric, respectively.

- **FID is especially inconsistent!**
- SWD (**pixel-level**, not learned feature-level) may be good for certain datasets, **but still isn't consistent.**



Conclusion: ***we need better metrics for medical image translation*** *that satisfy these desiderata!*