

Cycle-Consistent Counterfactuals by Latent Transformations

Saeed Khorram and Li Fuxin

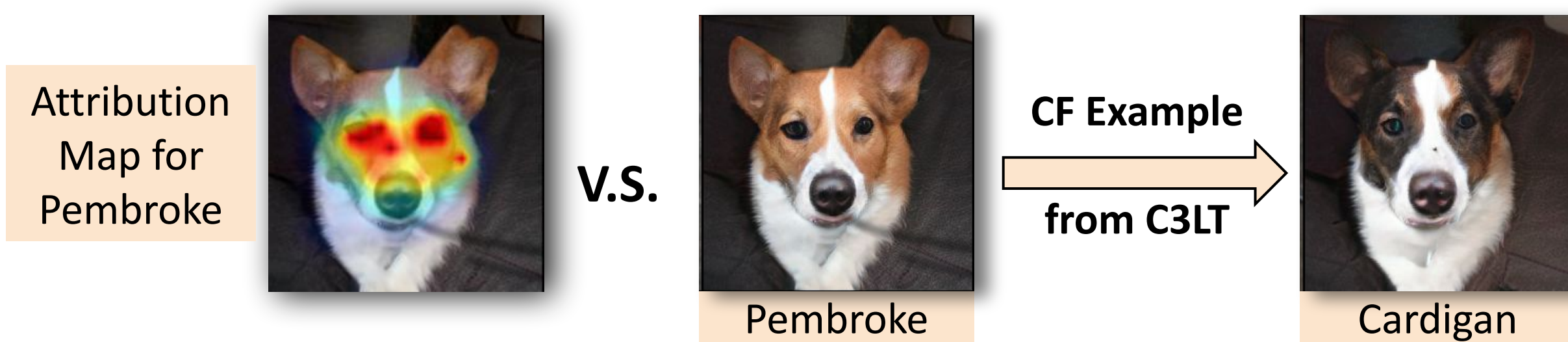
Collaborative Robotics and Intelligent Systems Institute, Oregon State University



CounterFactual (CF) Visual Explanation

- Motivation** Humans prefer to see natural images (e.g. nearest neighbor images from the training set, Cfs) rather than explanations such as attribution maps.

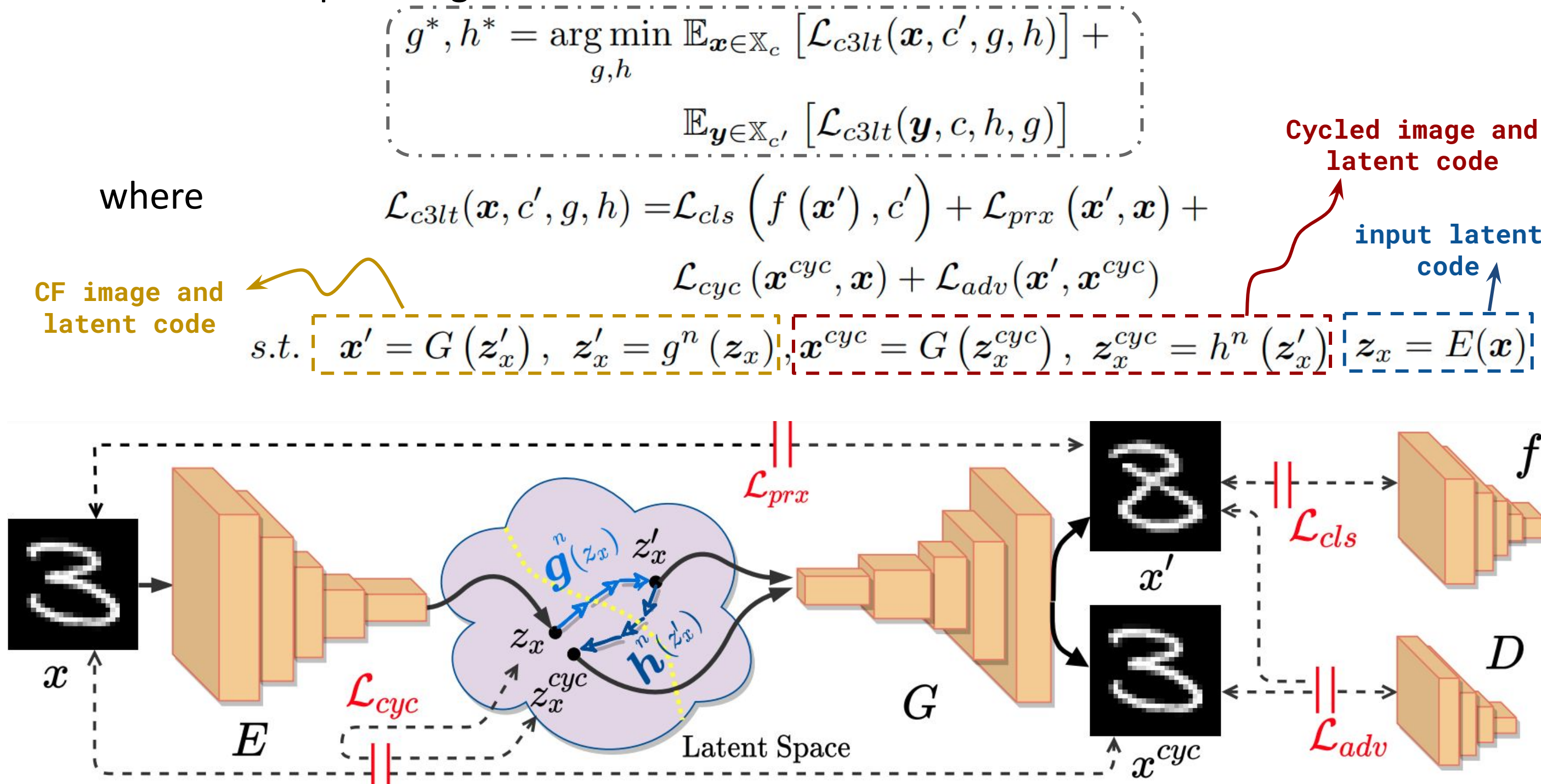
CFs are similar to the query (input) image but change the decision of a vision system to a specified outcome.



- Limitations of previous work** in generating CFs:
 - **Lie off the data manifold** (e.g. pitfall of adversarial solutions)
 - **Difficult to use/integrate** (e.g. require architectural modification in GANs)
 - **Slow to generate** (solving individual optimization to generate each CF)
 - **Limited to trivial/low-resolution datasets** (e.g. Mnist)

Cycle-Consistent Counterfactuals by Latent Transformations (C3LT)

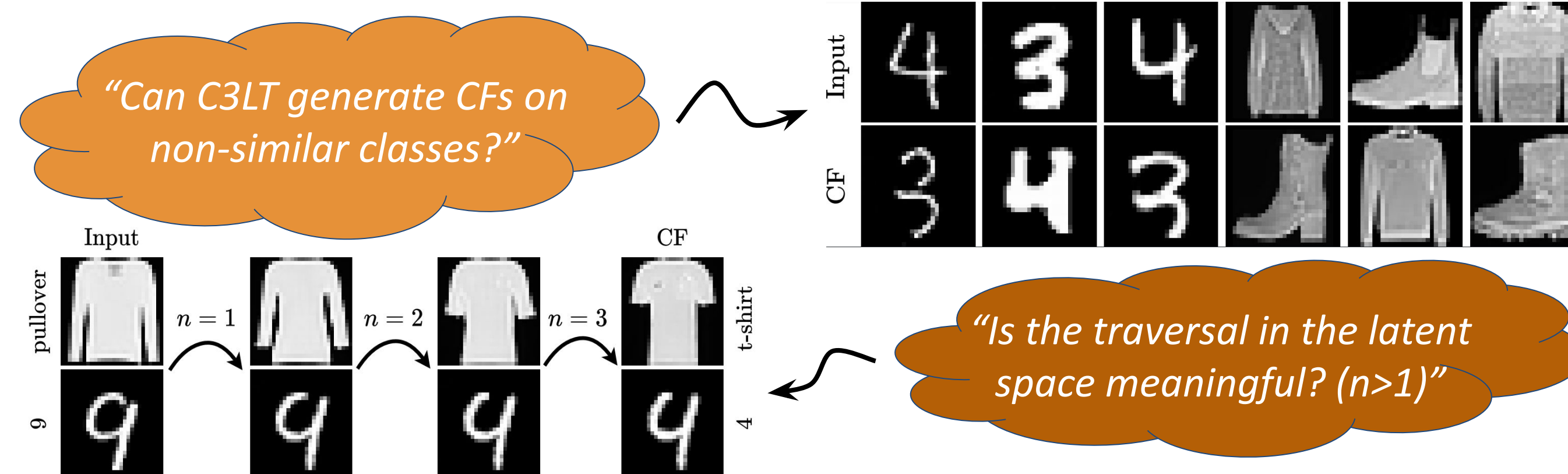
- C3LT re-defines finding CFs as learning non-linear cycle-consistent mappings (g and h) in the latent space of generative models:



Notes on C3LT

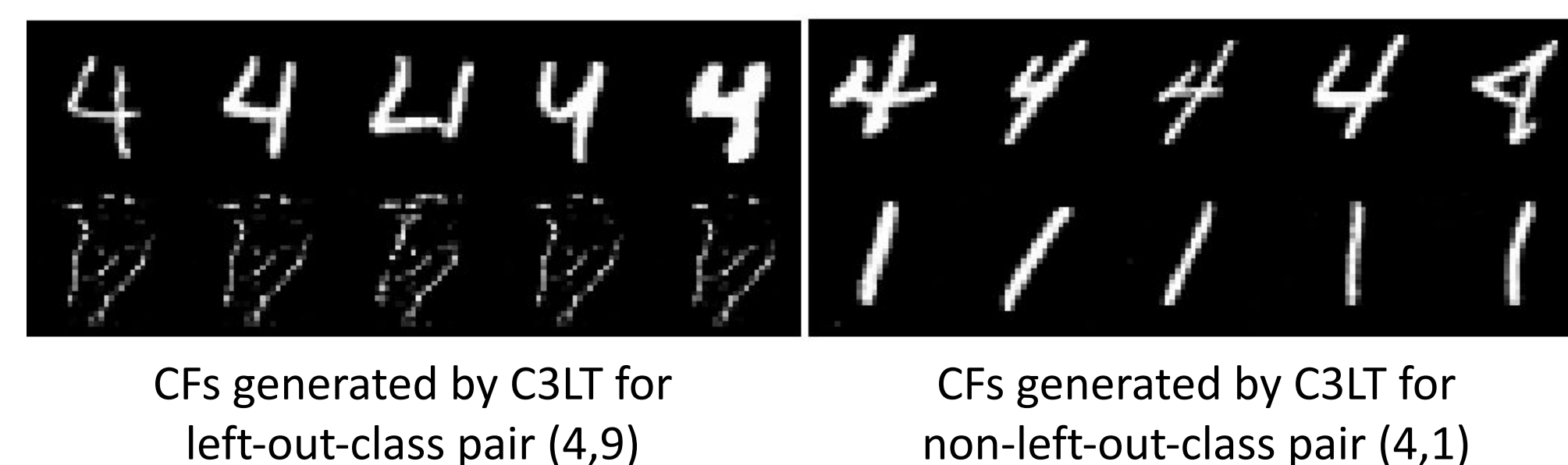
- **No optimization for individual images.** (on-the-fly generation at inference!)
- **Plug and play** using any pre-trained GAN/VAE. (no retraining needed!)
- **Cycle-Consistency** regularizes mapping finding. (highly under-constrained problem)
- **Adv. Loss** to help staying on the data manifold.
- **Explains a classifier f** compared to conditional GANs. (see below)

Qualitative Experiments



Debugging a Faulty Classifier

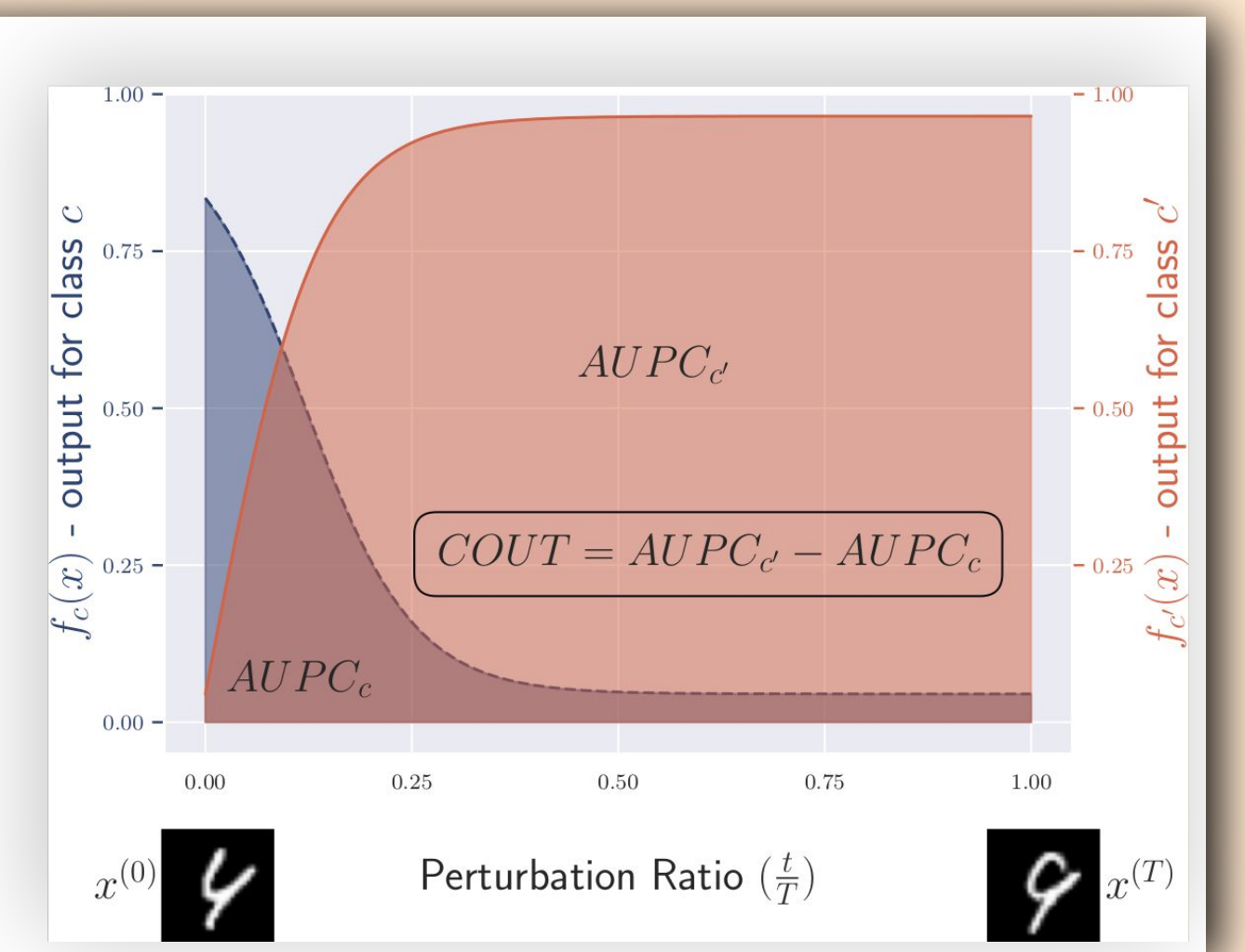
- ❖ MNIST classifier with left-out class 9. (~89% acc.)
- ❖ Cycle-GAN and other conditional GANs can't debug.



COUnterfactual Transition (COUT) Metric

- ❖ Automatic evaluation considering the changes in the output score of the query and CF classes simultaneously.

$$AUPC_k = \frac{1}{T} \left\langle \sum_{t=0}^{T-1} \frac{1}{2} \left(f_k(x^{(t)}) + f_k(x^{(t+1)}) \right) \right\rangle_{p_{data}}$$



Quantitative Evaluation

Methods	ExpGAN		CEM		CVE		C3LT (ours)	
	Mnist	FMnist	Mnist	FMnist	Mnist	FMnist	Mnist	FMnist
$AUPC_{c'} \uparrow$	0.967	0.920	0.301	0.347	0.209	0.275	0.980	0.958
$AUPC_c \downarrow$	0.040	0.062	0.555	0.427	0.767	0.638	0.031	0.052
$COUT \uparrow$	0.927	0.858	-0.253	-0.080	-0.557	-0.363	0.948	0.906

The **COUT** metric evaluates the validity and sparsity of CF examples.

Methods	ExpGAN		CEM		CVE		C3LT (ours)	
	Mnist	FMnist	Mnist	FMnist	Mnist	FMnist	Mnist	FMnist
$IM1 \downarrow$	0.72	0.77	1.68	1.63	1.44	1.24	0.70	0.74
$IM2 \times 10 \downarrow$	0.43	0.14	1.08	0.26	1.38	0.37	0.36	0.093
$FID \downarrow$	41.12	76.52	50.03	96.87	47.53	83.77	22.83	62.31
$KID \times 1e3 \downarrow$	37.27	70.44	44.88	91.71	37.24	72.71	13.39	52.71

Realism of the generated CF examples.

Adversarial!

Methods	ExpGAN		CEM		CVE		C3LT (ours)	
	Mnist	FMnist	Mnist	FMnist	Mnist	FMnist	Mnist	FMnist
$Prox \downarrow$	0.074	0.135	0.016	0.013	0.055	0.054	0.072	0.116
$Val \uparrow$	0.997	0.998	0.469	0.620	0.231	0.145	0.999	1.0

Validity and Proximity of the CF examples.

Conclusion

- ❖ We proposed a novel framework to generate realistic CFs by learning cycle-consistent transformations in the latent space of pre-trained generators.
- ❖ C3LT is able to generate CFs at high-res (e.g. 256x256) images (e.g. from ImageNet) with real-time speed at inference time.
- ❖ C3LT can be readily plugged into existing pre-trained generative algorithms without modifying the architecture/retraining.
- ❖ We showed the effectiveness of C3LT through qualitative and quantitative experiments.

This work is partially supported by DARPA contract N66001-17-2-4030 and NSF Award #1927564.