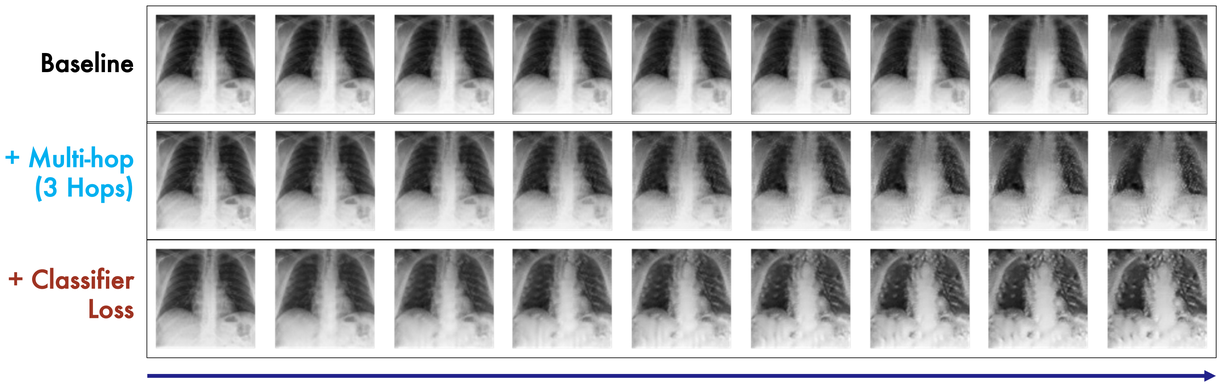
Feature Exaggeration

Berthy Feng, James Bowden, Katie Bouman

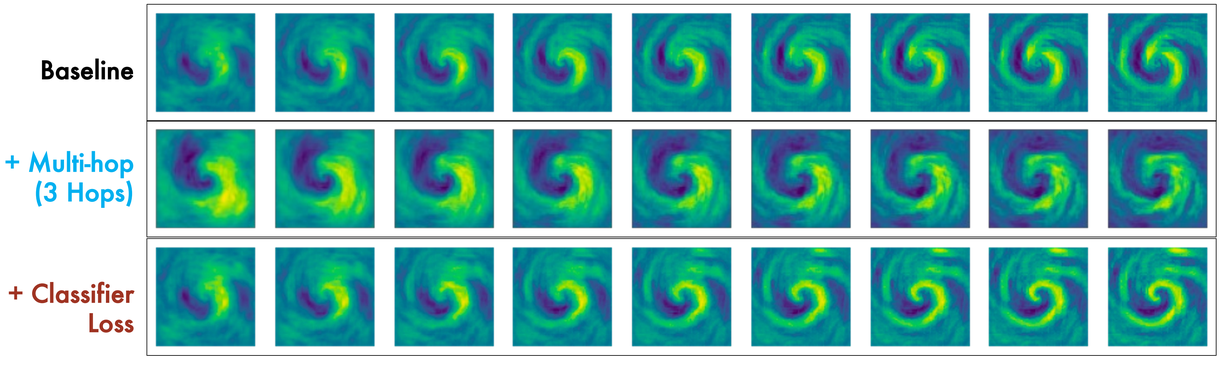
Caltech

Initial Results

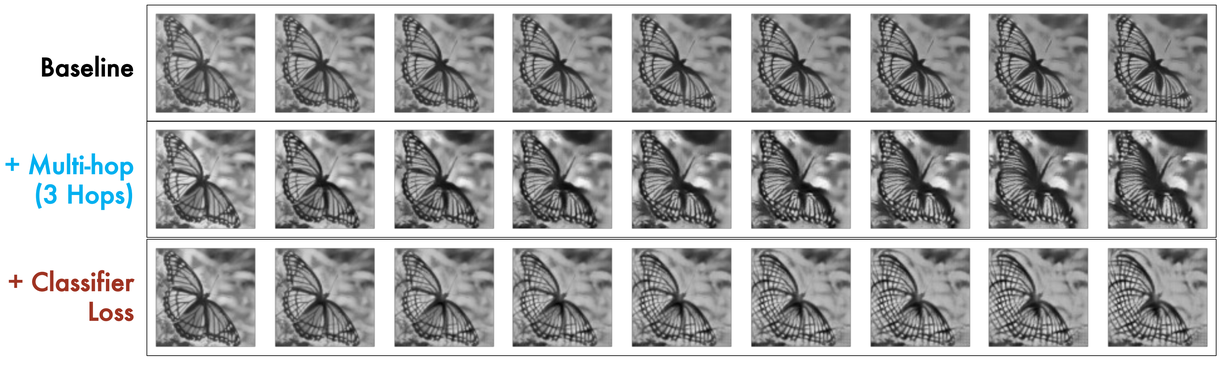
Chest X-Rays: Normal vs. Enlarged Heart



Fluid-Flow Simulations: Smaller vs. Larger Opening Angle



Butterflies: Monarch vs. Viceroy

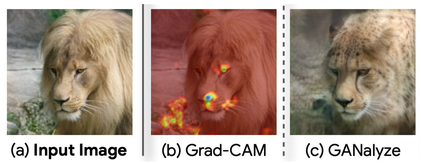


Acknowledgments

This work is done in collaboration with Google-affiliated researchers Miki Rubinstein, Oran Lang, and Inbar Mosseri. B.F. is supported by a Kortschak Scholarship and an NSF GRFP Fellowship.

Motivation

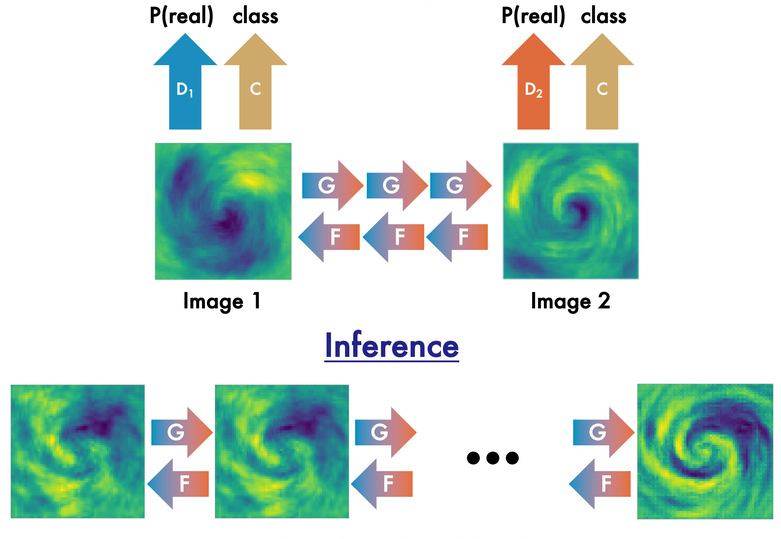
- Humans need *interpretable* features to do visual classification
- Existing explainability techniques (e.g., Grad-CAM, GANalyze) struggle to reveal the simplest semantic feature



Lion vs. Cheetah explanation (StylEx paper, Fig. 2)

- **Our goal:** help humans discover important visual features by exaggerating them in the *simplest, most interpretable* way

CycleGAN-Based Approach



*Ask me for technical details!

Related Work

- Baseline CycleGAN approach: "Scientific Discovery by Generating Counterfactuals Using Image Translation," Narayanaswamy et al., arXiv 2020.
- StylEx: "Explaining in Style: Training a GAN to explain a classifier in StyleSpace," Lang et al., ICCV 2021.
- Multi-hop GAN
- A whole bunch of explainability work!
- A whole bunch of GAN work!

Open Questions

Interpretable features:

- Why does CycleGAN sometimes learn an interpretable feature when the classifier does not?
- What to do when neither CycleGAN nor classifier learns an interpretable feature?
  - Other dataset statistics might give stronger classification signal, but are not necessarily interpretable/simple
  - Could restrict to a set of allowed transformations?
- Could do exaggeration in VAE latent space, but how to ensure the decoder can express exaggerated images?

Evaluation:

- How to evaluate classification score of exaggerated images?
- How to evaluate probabilities on exaggerated images?