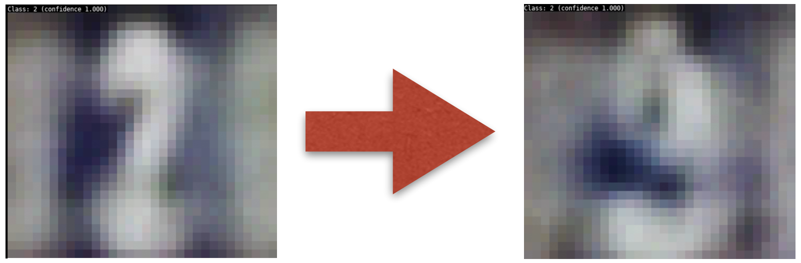
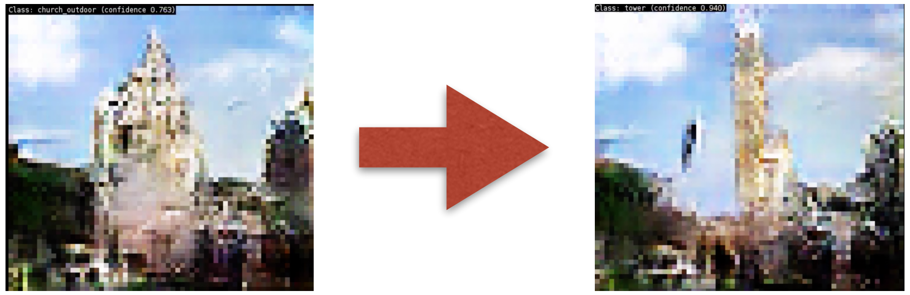
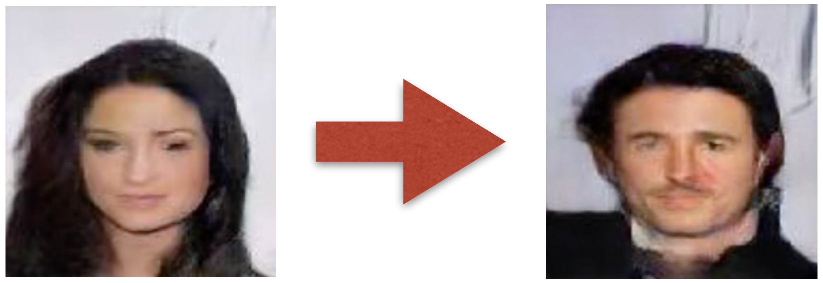
Counterfactual Image Generation

- Train two networks: encoder **E** and generator **G**, with GAN loss
- Train a classifier **C** on the features learned by **E**
- Ask a “What-If” question: “What if input *x* was of class *P* ?”
- Optimization: Find the vector ***z*** closest to *E(x)* that is classified as *P*

$$\min_{z} ||E(x) - z||_2$$
$$\text{subject to} \quad \arg\max_{p} \, C(z)_p = P$$

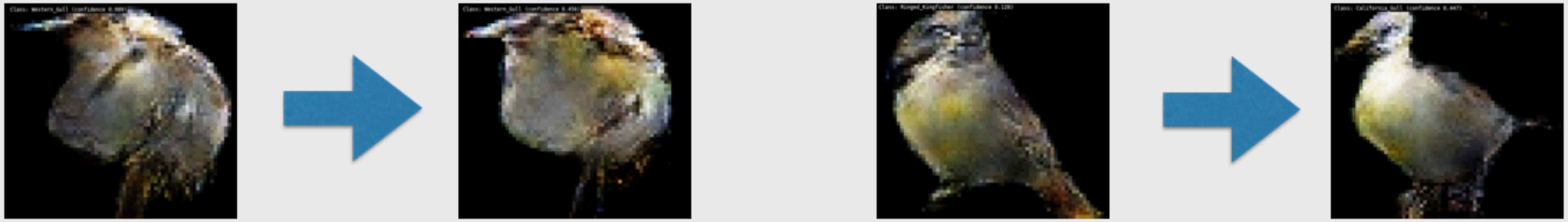



Class: church_outdoor (confidence 0.763)
Class: tower (confidence 0.940)

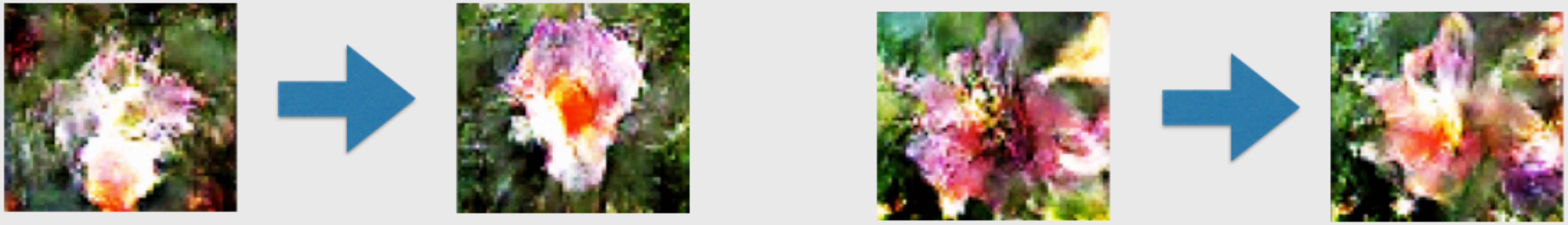


What if **this image looked like class** ***X*** **?**

- Bird Species Classification: Visualizing Class “Gull” (Dataset: CUB200)


Class: Western_Gull (confidence 0.009)
Class: Western_Gull (confidence 0.456)
Class: Ringed_Kingfisher (confidence 0.120)
Class: California_Gull (confidence 0.447)

- Plant Species Classification: Visualizing “Hard-Leaved Pocket Orchid” (Dataset: Oxford Flowers 102)



- Human face attribute classification: Visualizing attribute “male/female” (Dataset: CelebA)