# Synthesis: Image Generation and Hole Filling

**Danna Gurari**
University of Colorado Boulder

Fall 2021



https://home.cs.colorado.edu/~DrG/Courses/RecentAdvancesInComputerVision/AboutCourse.html

# Review

- Last lecture topic:
  - Synthesis: style transfer
- Assignments
  - Final project outline due tonight
  - Final project presentation due in three weeks
  - Peer evaluation due in three weeks
  - Final project report due in four weeks
- Questions?

# Today's Topics

- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# Today’s Topics

- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# Synthesis (With and Without Context)

Generation &
Alteration:

Hole Filling
(constrained
generation):

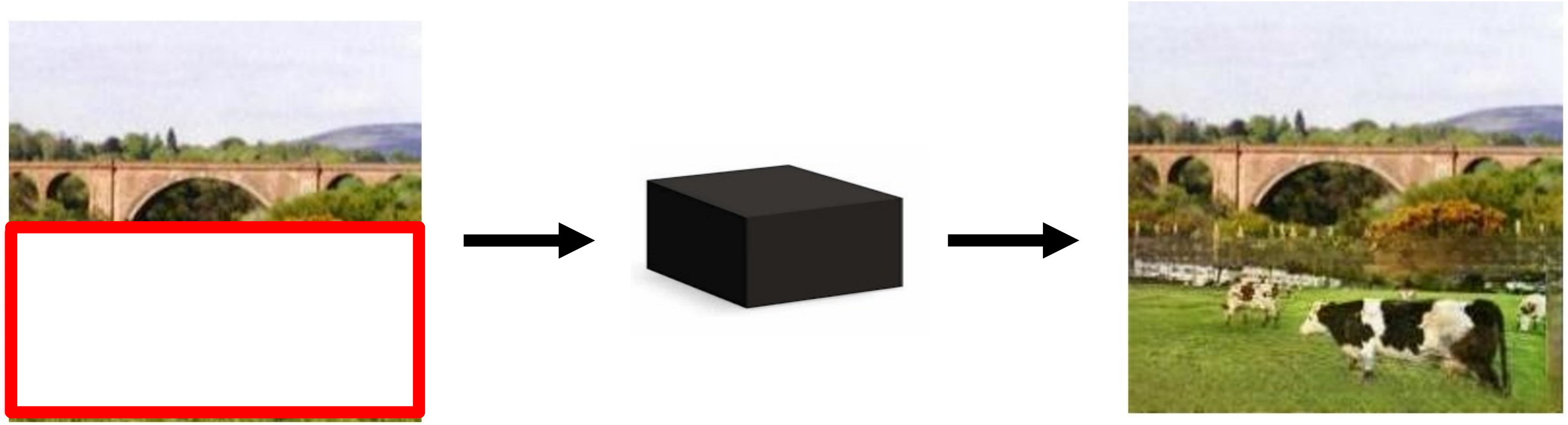


Image source: https://medium.com/image-recreation-a-method-to-make-learning-of-gan/image-generation-text-to-image-d7c4210ecb90

# Today’s Topics

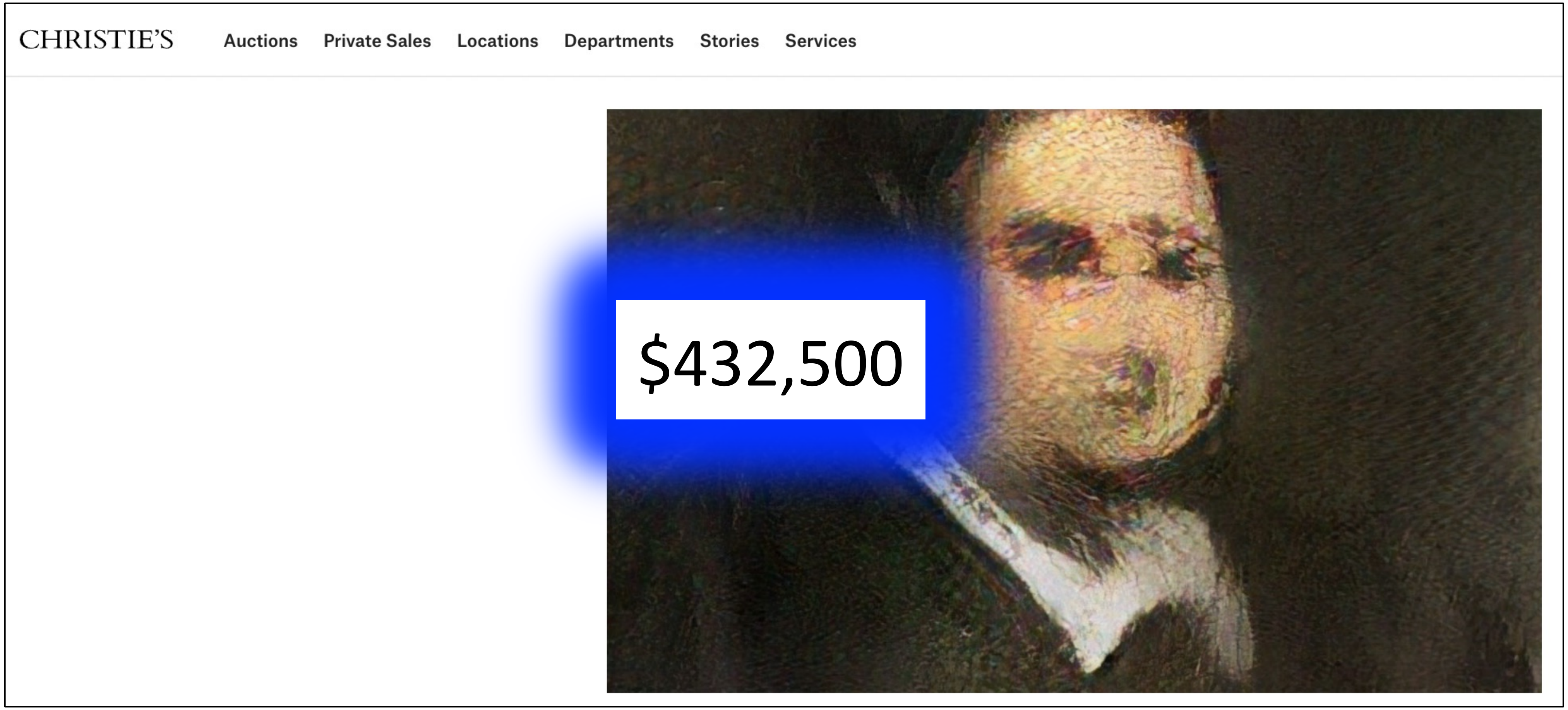
- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# Refinement (e.g., enhance, avoid payment, and/or rewrite history)

- Damaged regions from camera
- Blurred areas
- Watermarks
- Undesired content

e.g., remove ex-partner

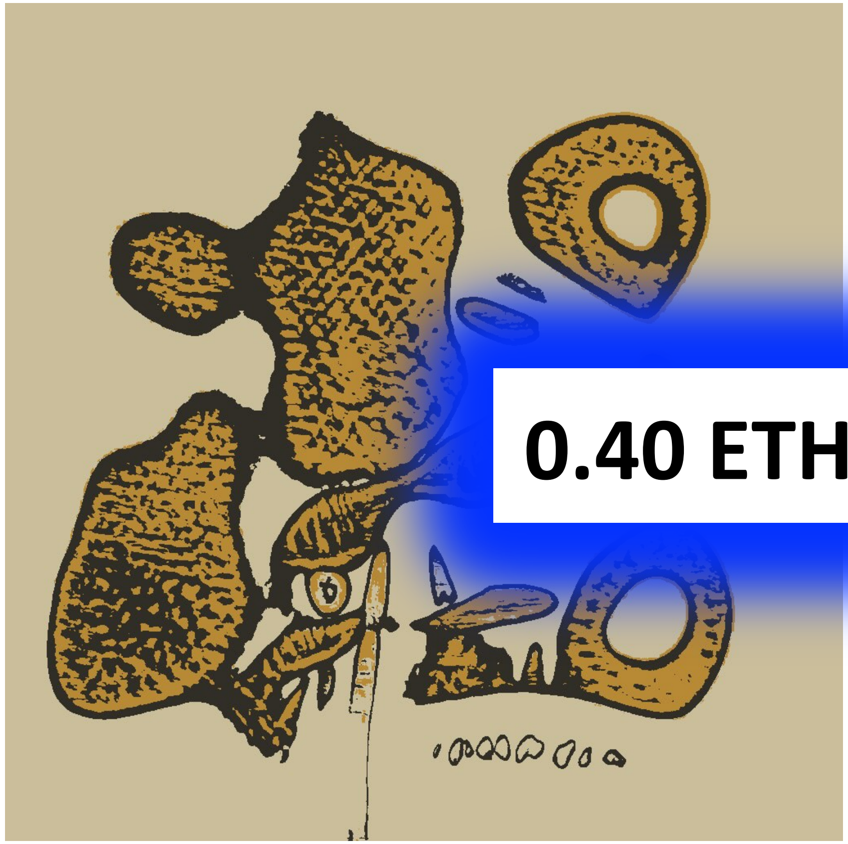
# Commercial Art

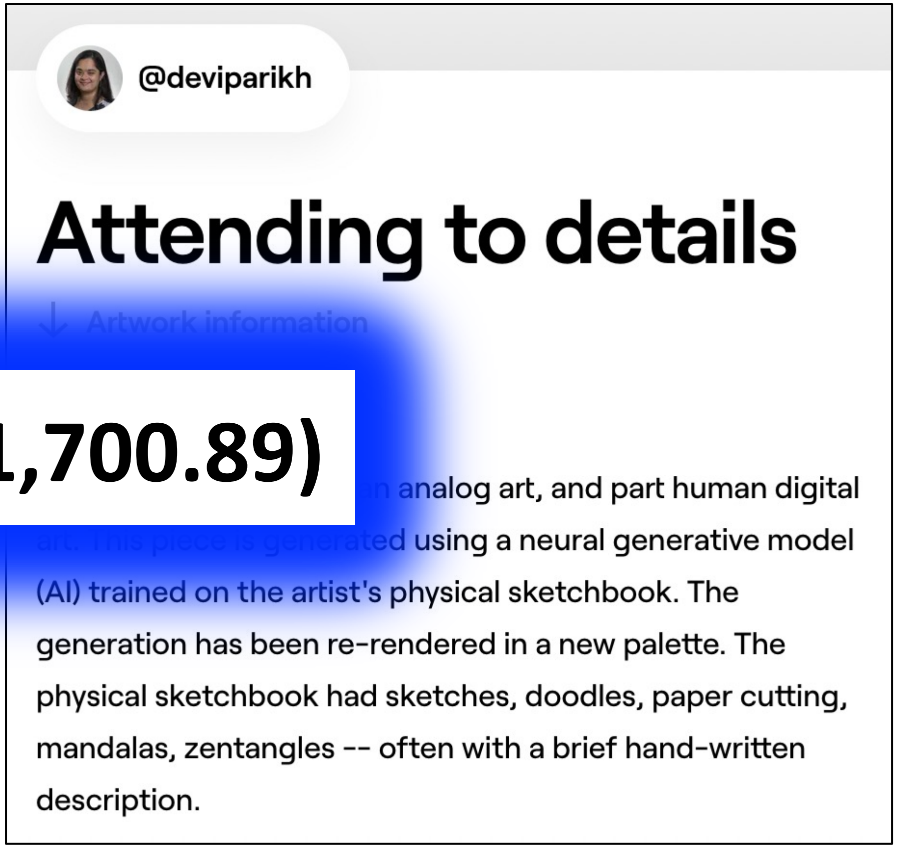



How much do you think this sold for?

https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

# Commercial Art



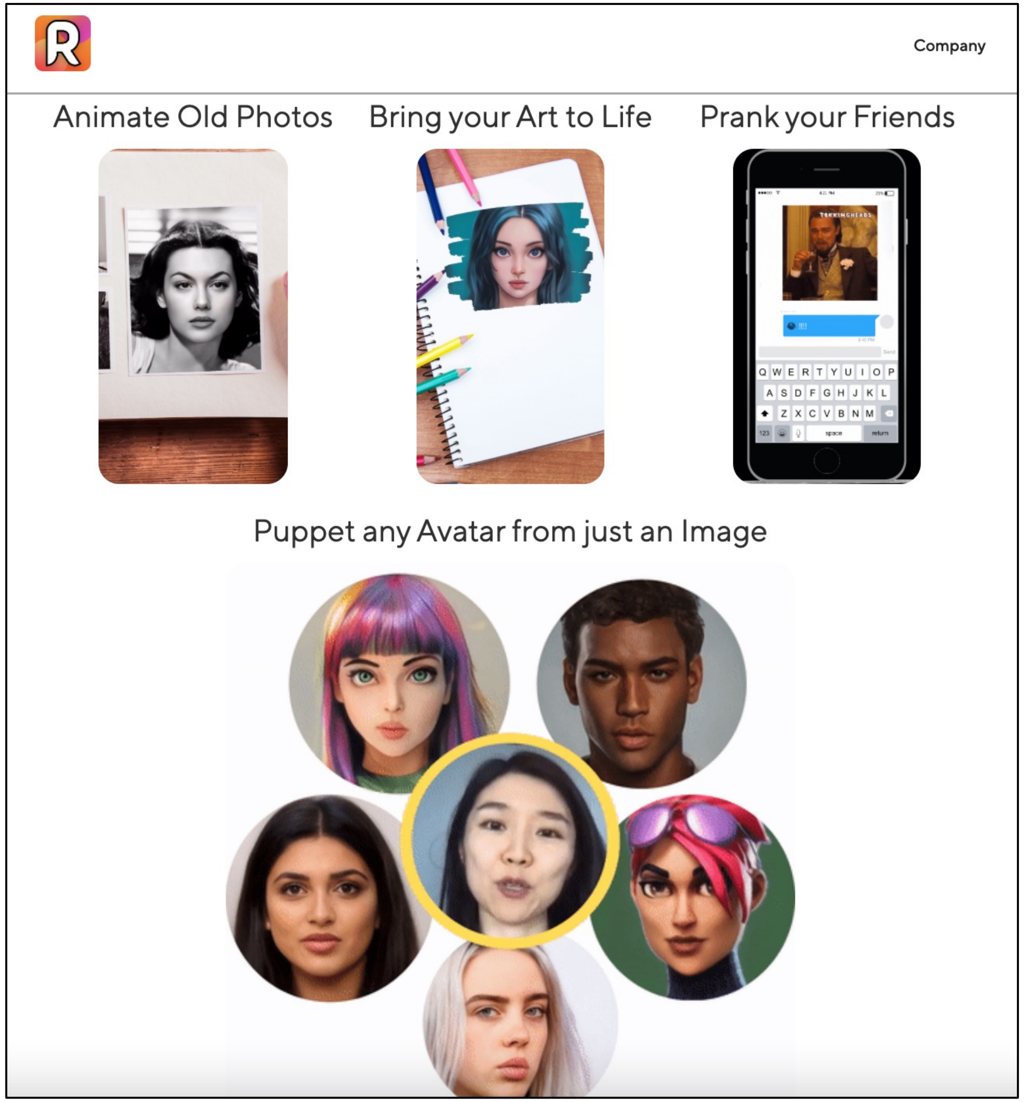



**0.40 ETH ($1,700.89)**

How much do you think this sold for?

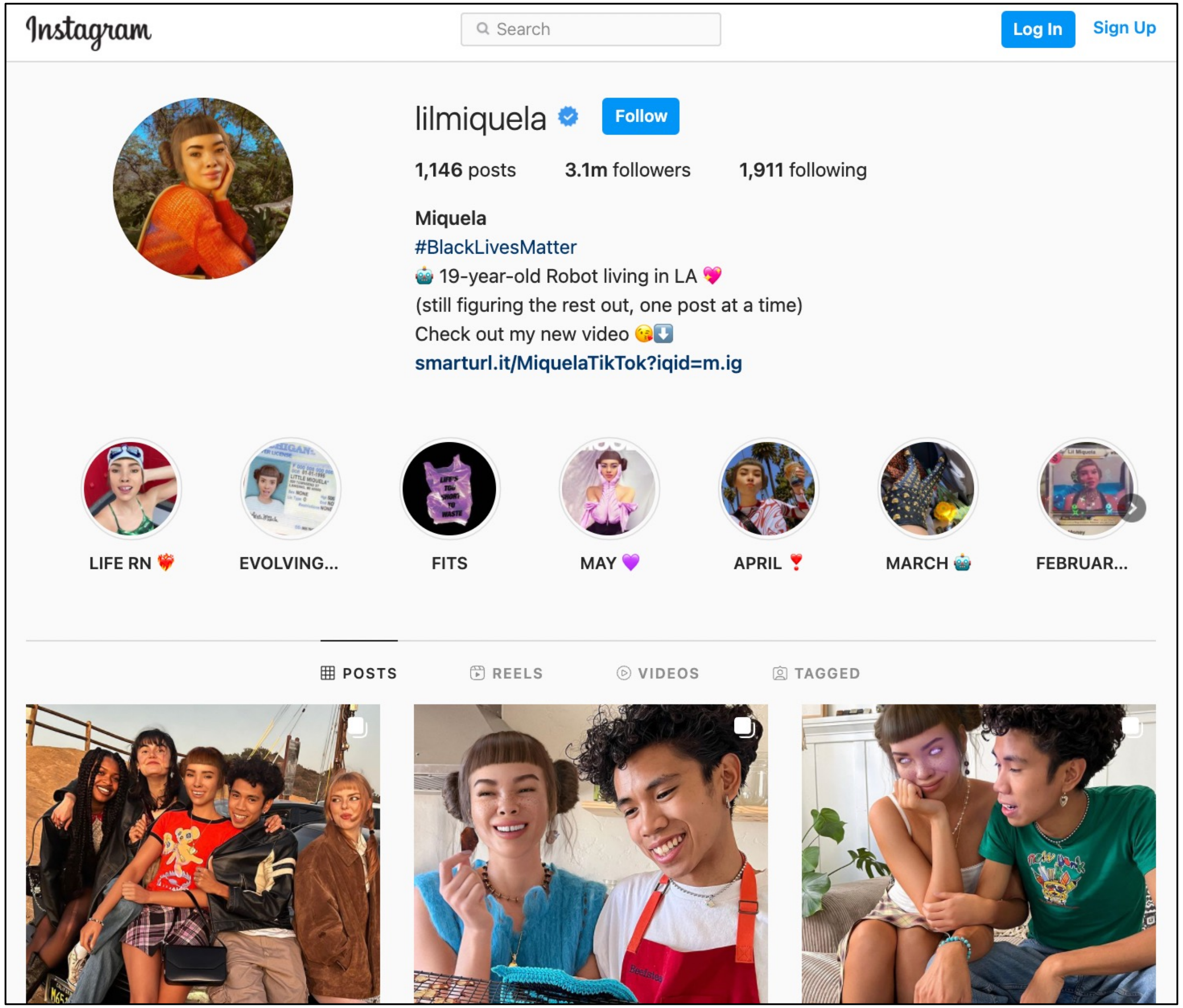
https://foundation.app/@deviparikh/~/79149

# Entertainment




https://www.rosebud.ai/

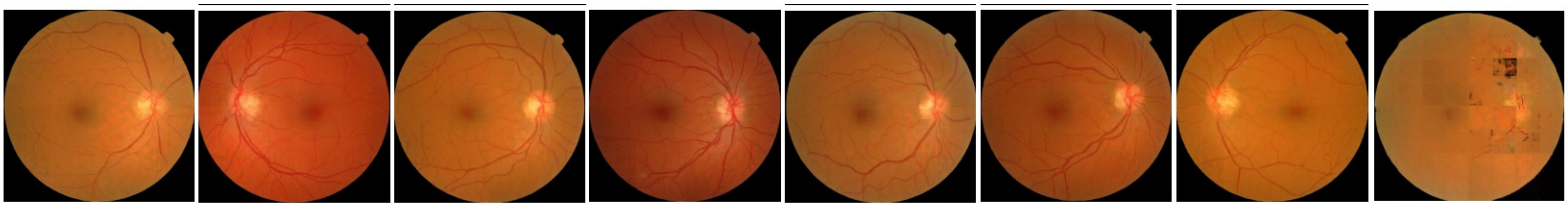
# Social Media



# News

e.g., face re-enactment
Demo: https://www.youtube.com/watch?v=ttGUiwfTYvg

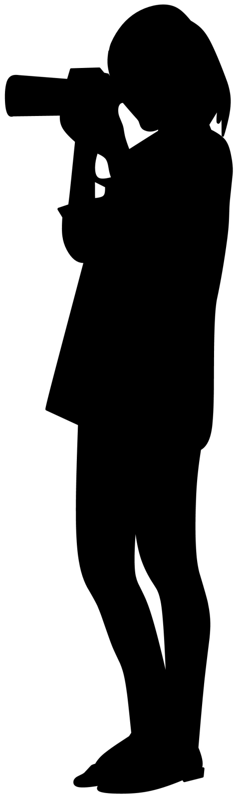
# Training Data Creation

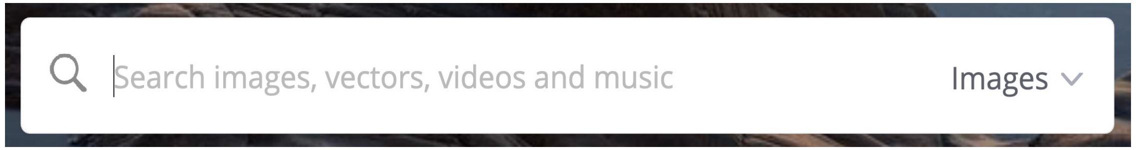


Costa et al. Towards Adversarial Retinal Image Synthesis. arXiv 2017.

# Improved Messaging via Visual Content

- Marketing
- Artwork
- Presentations
- Blogs
- Websites






Potential sources:

Photographer
(self or hired)

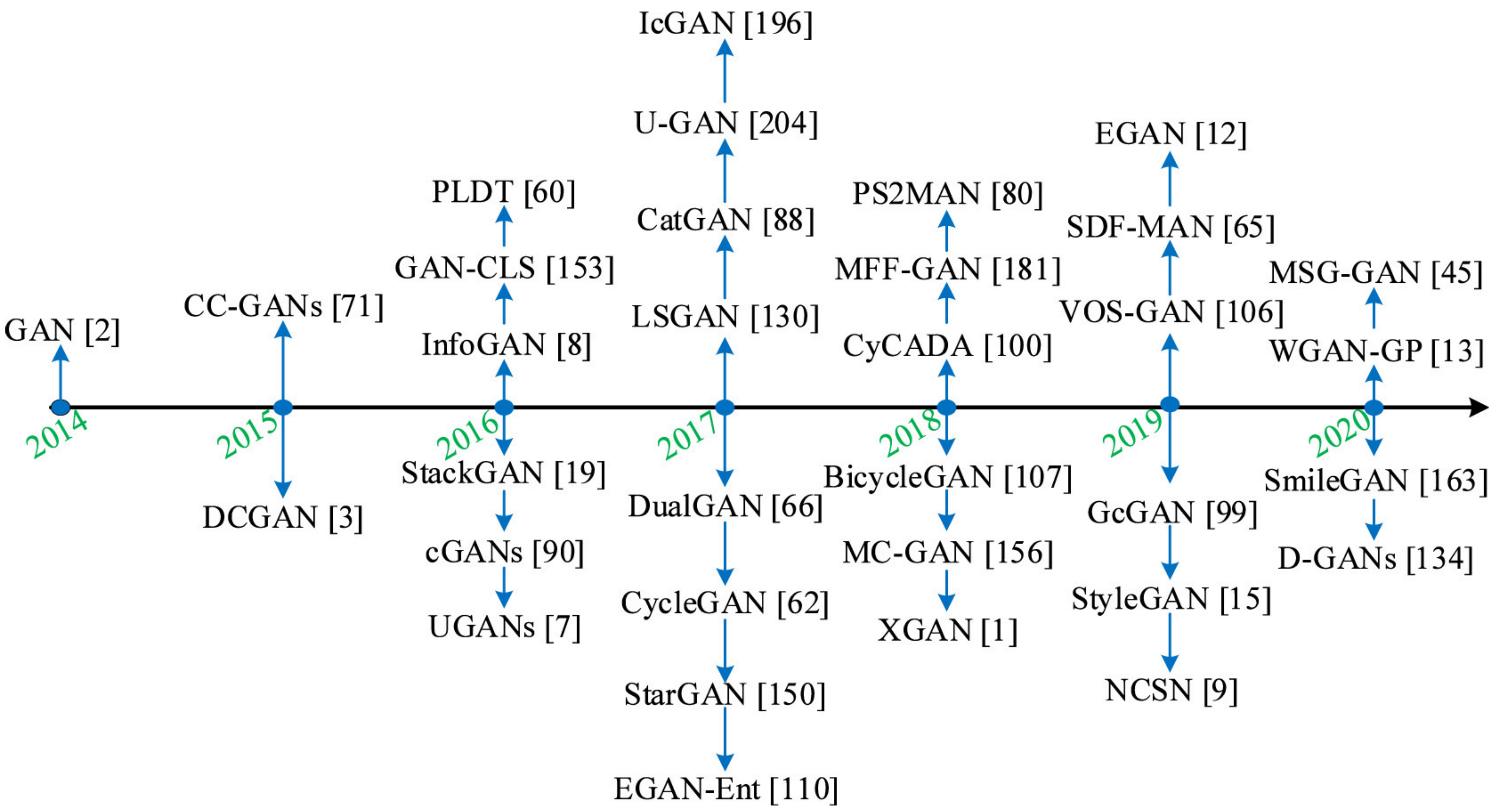
Stock photos

What are other possible applications of image synthesis, for good and harm?

# Today’s Topics

- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# Generative adversarial networks




https://ars.els-cdn.com/content/image/1-s2.0-S1566253521000385-gr2_lrg.jpg

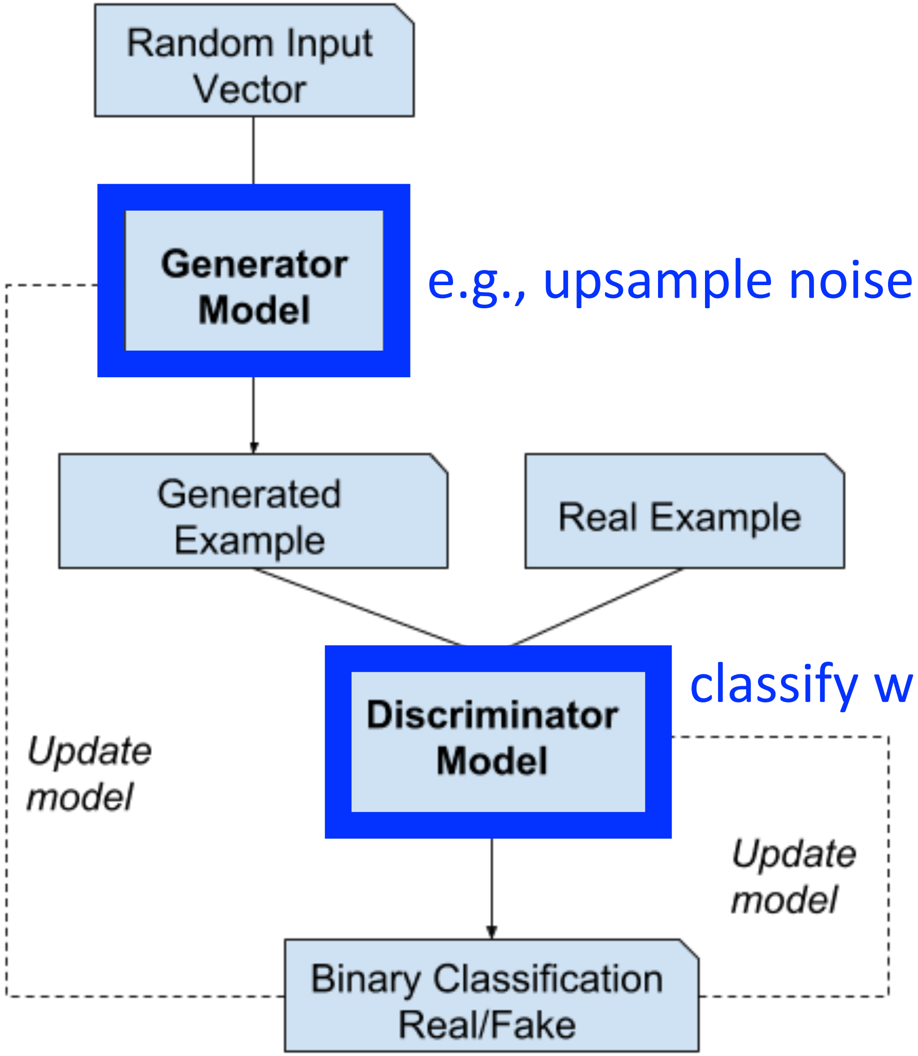
# Generative adversarial networks

- Generative adversarial networks (GANs)
- Deep convolutional generative adversarial networks (DCGANs)
- GIRAFFE

# Generative adversarial networks

- Generative adversarial networks (GANs)
- Deep convolutional generative adversarial networks (DCGANs)
- GIRAFFE

# GAN: Basic Architecture



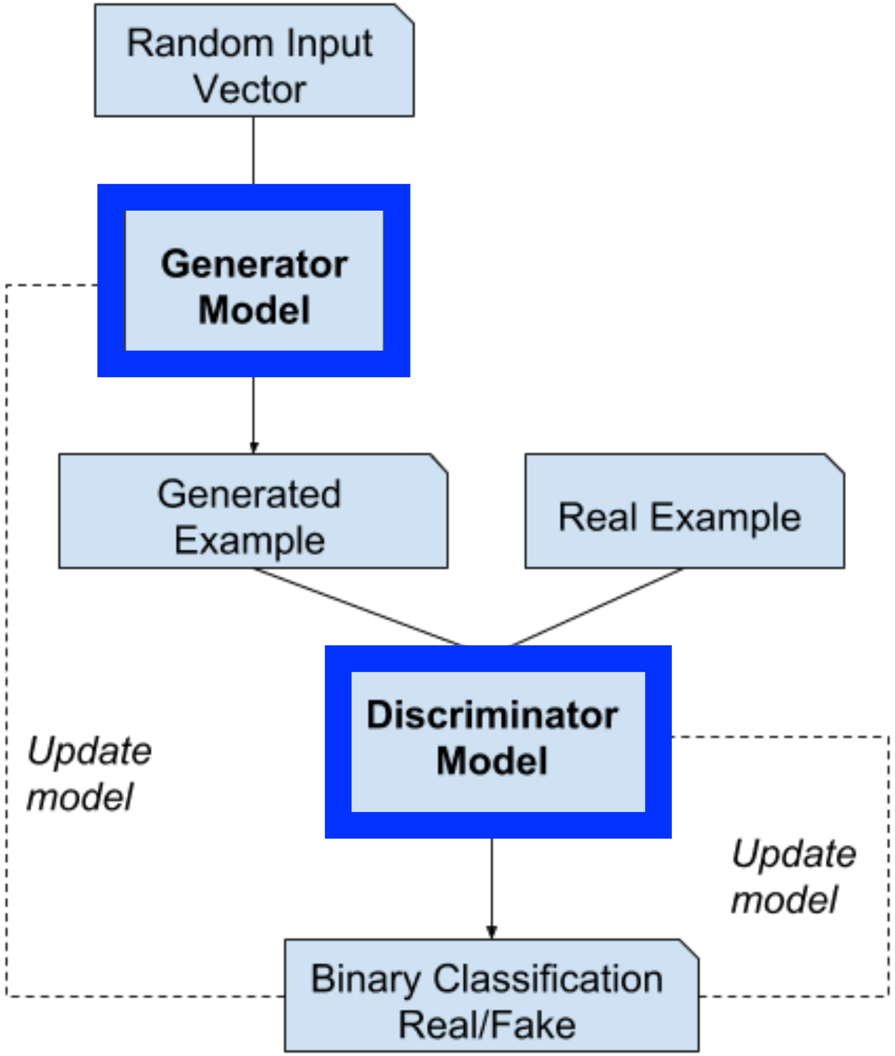

e.g., upsample noise to image of desired size

classify which examples are real

Consists of two models that compete against each other

https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

# GAN: Training




The two models are iteratively trained separately

- Train discriminator using fake and real images
- Train generator using just fake images and penalize it when the discriminator recognizes images are fake

https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

# GAN: Discriminator Loss Function

Discriminator tries to minimize classification error

Discriminator wants a value of 1 for real images

Discriminator wants a value of 0 for fake images

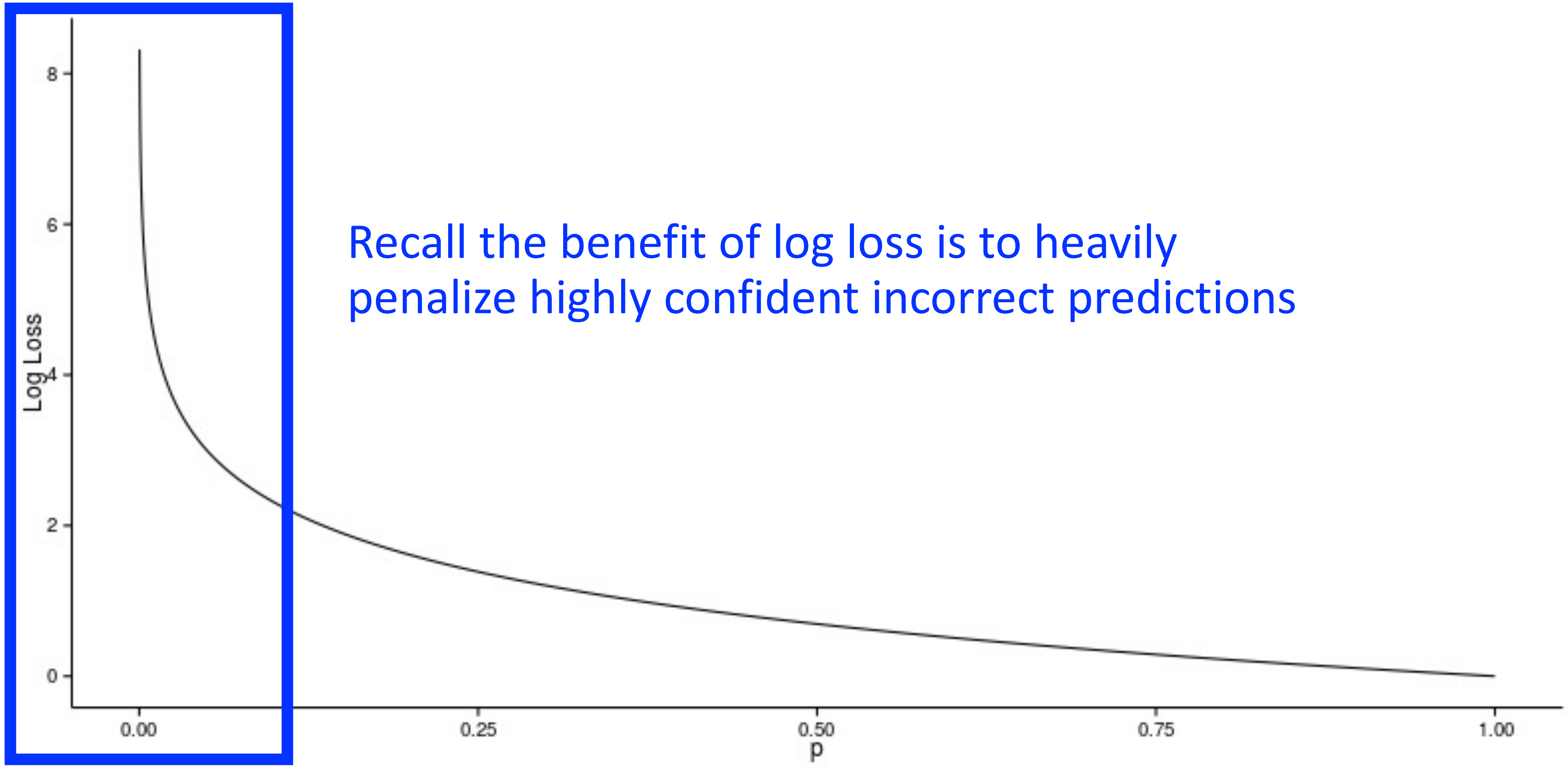
$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boxed{\boldsymbol{x}\sim p_{\text{data}}}} \log \boxed{D(\boldsymbol{x})} - \frac{1}{2}\mathbb{E}_{\boxed{\boldsymbol{z}}} \log\left(1 - \boxed{D\left(G(\boldsymbol{z})\right)}\right)$$

Real image

Input noise

https://arxiv.org/pdf/1701.00160.pdf

# GAN: Discriminator Loss Function




Recall the benefit of log loss is to heavily penalize highly confident incorrect predictions

https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f

# GAN: Generator Loss Function

Generator tries to maximize classification error

$$J^{(G)} = -J^{(D)}$$

Want the discriminator to mistakenly arrive at a value of 1 for fake images

$$J^{(G)} = -\frac{1}{2}\mathbb{E}_{\boxed{\boldsymbol{z}}} \log \boxed{D\left(G(\boldsymbol{z})\right)}$$

Input noise

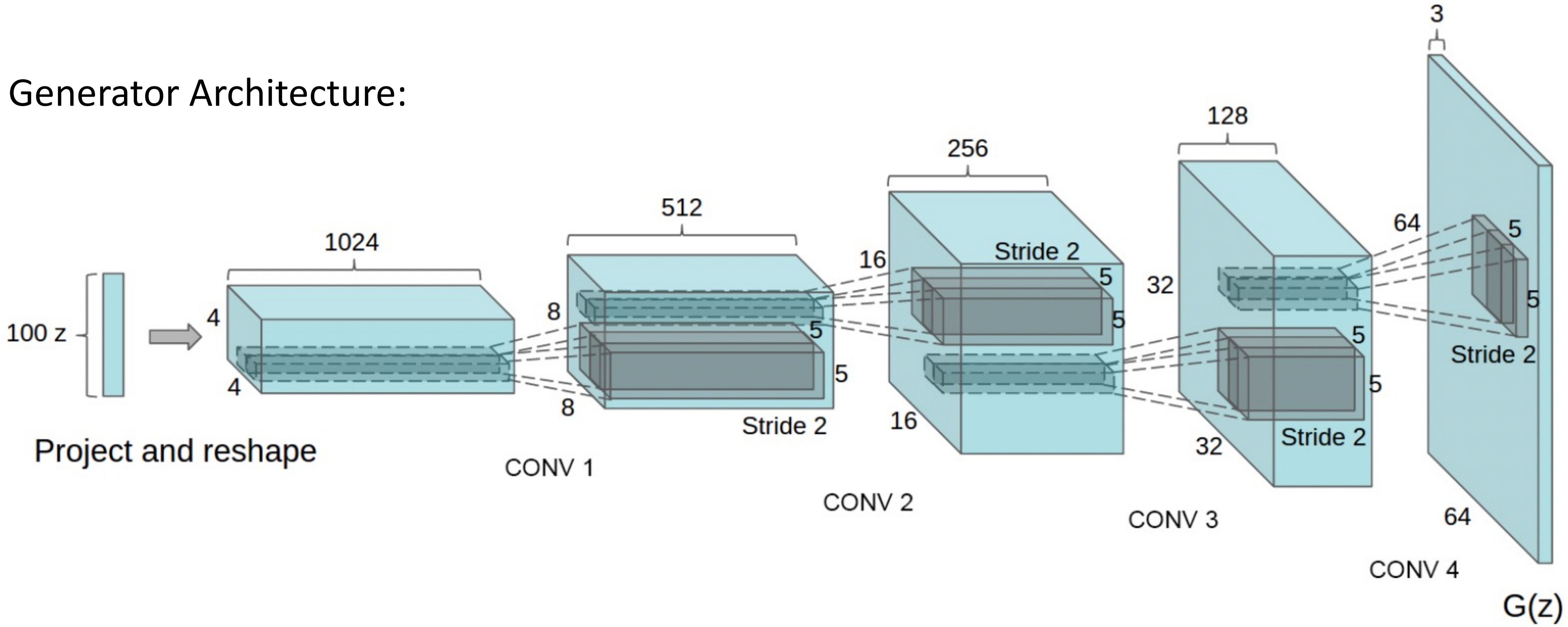
https://arxiv.org/pdf/1701.00160.pdf

# Generative adversarial networks

- Generative adversarial networks (GANs)
- Deep convolutional generative adversarial networks (DCGANs)
- GIRAFFE

# DGANs: GANs that Use Convolutional Layers

Generator Architecture:




What is the resolution of the image generated by this network?

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

# DGANs: Qualitative Results

Bedrooms generated by observing over 3M bedroom images

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

# DGANs: Qualitative Results

What objects does it learn to generate?

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

# DGANs: Qualitative Results

What objects may it not have learned to generate?

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

# DGANs: Qualitative Results

Faces generated by observing over 3M images of 10K people

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

# DGANs: Qualitative Results

What does it generate poorly or not all?

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

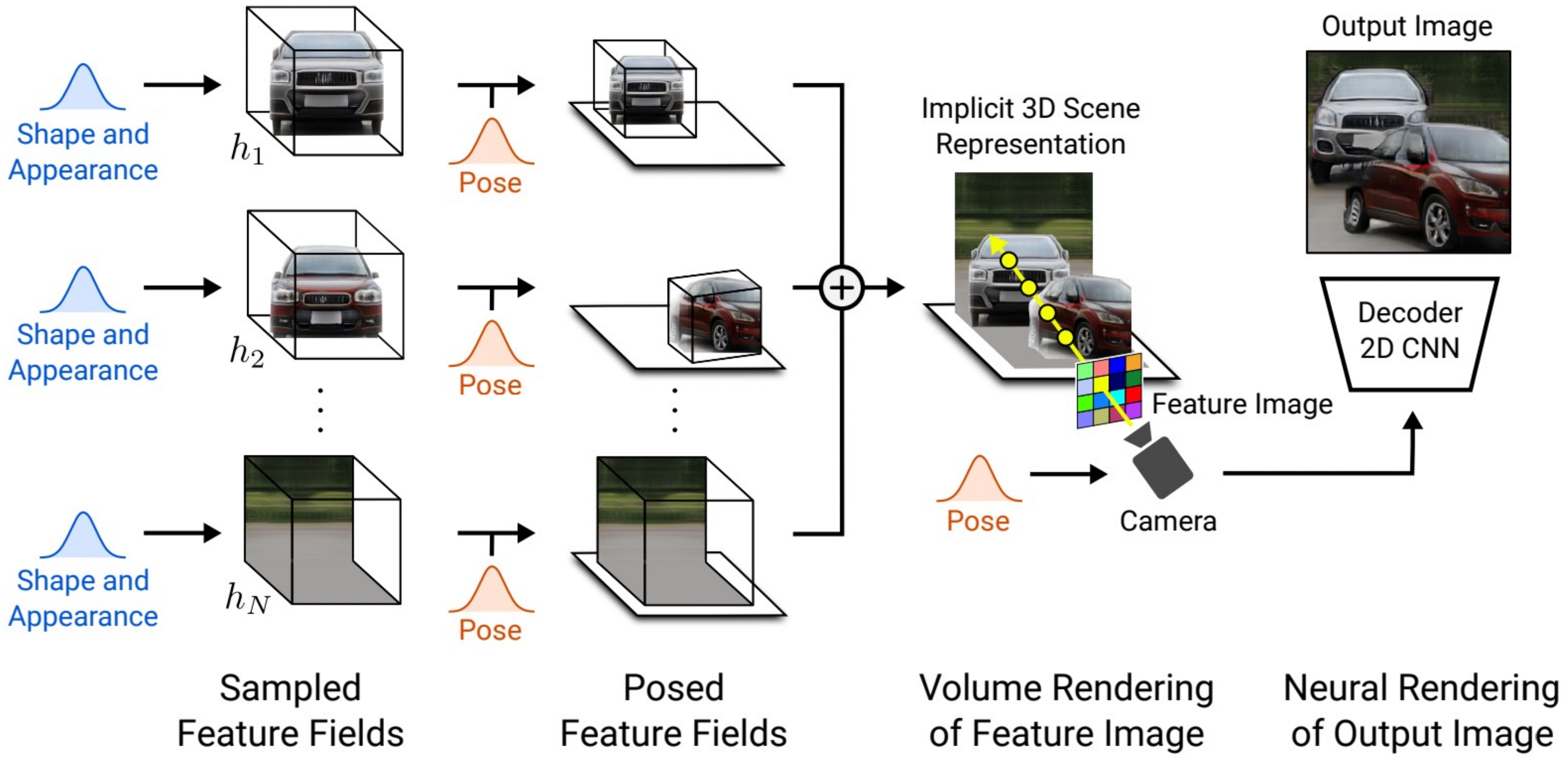
# DGANs: Limitation

Cannot control what is generated

# Generative adversarial networks

- Generative adversarial networks (GANs)
- Deep convolutional generative adversarial networks (DCGANs)
- GIRAFFE

# GIRAFFE: Idea

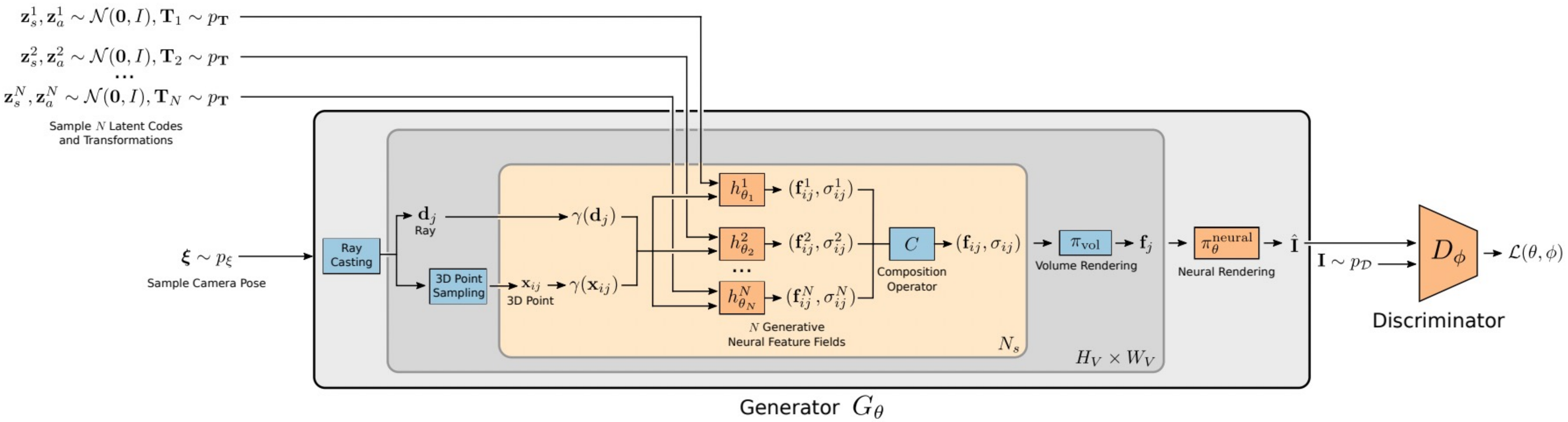



Key idea: control what is synthesized using a 3D scene representation in the generator

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Architecture

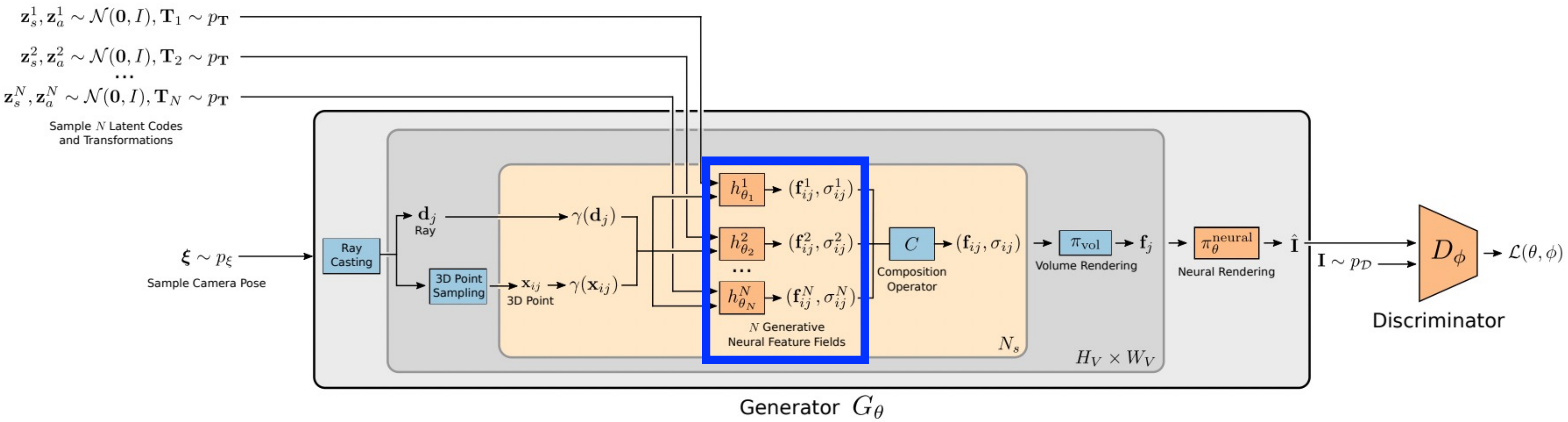



Key idea: control what is synthesized using a 3D scene representation in the generator

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Architecture

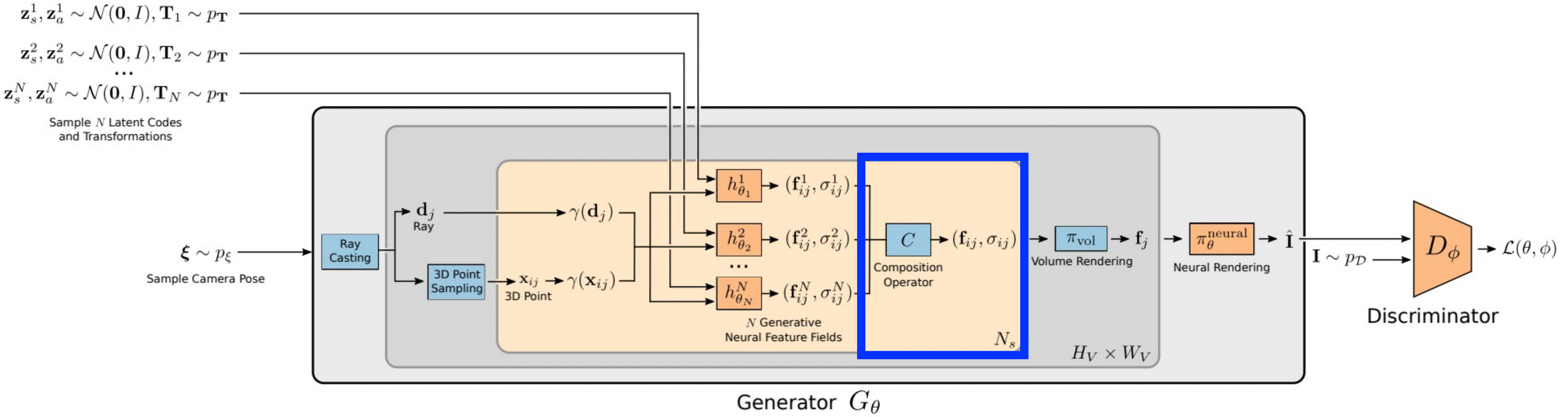



The *N-1* objects and background to appear in the scene are not only represented separately but also separate from their shape and appearance

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Architecture

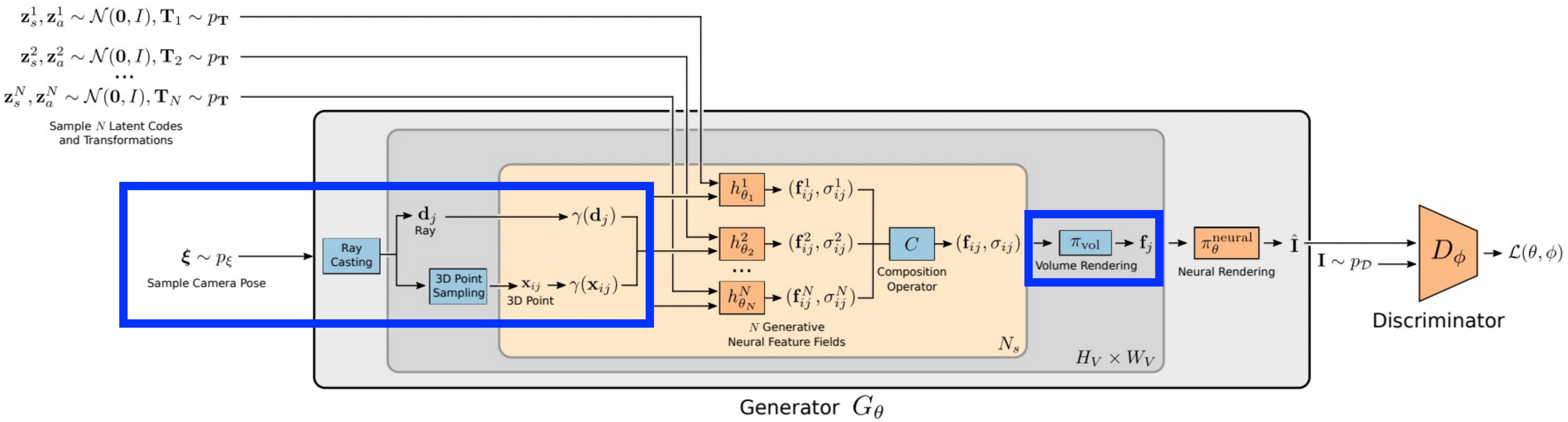



These $N$ entities are then incorporated into a scene representation

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Architecture

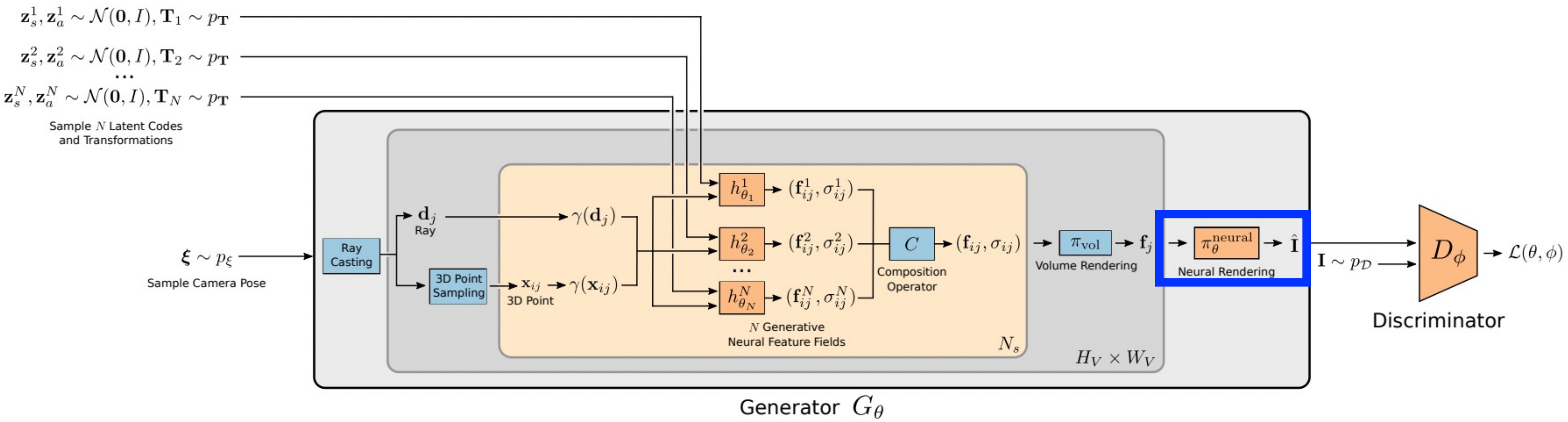



Knowledge about the camera pose is used to render a high dimensional feature vector for each pixel

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Architecture

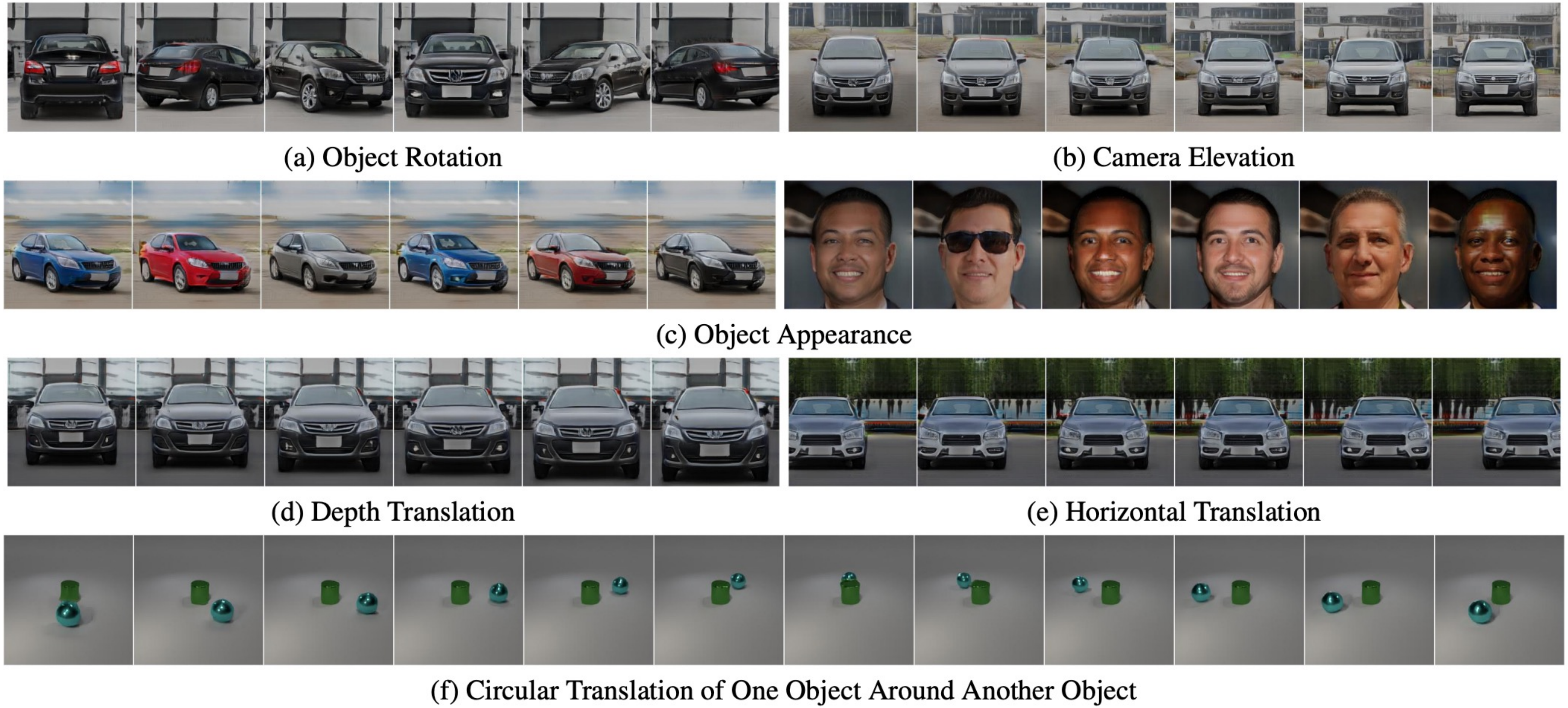



The final 2D image is then rendered

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Qualitative Results




Can control synthesized results!

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# GIRAFFE: Qualitative Results

Can control synthesized results!

(Recognized with Best Paper Award)

Niemeyer and Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021.

# Generative adversarial networks

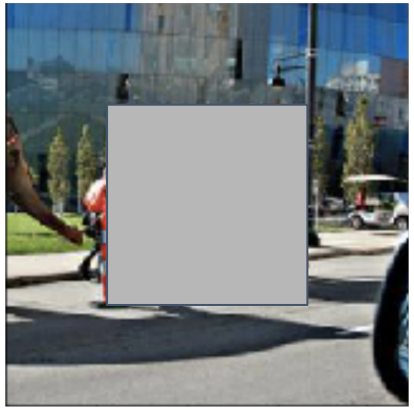
- Generative adversarial networks (GANs)
- Deep convolutional generative adversarial networks (DCGANs)
- GIRAFFE

# Today’s Topics

- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# Key Challenge

- What might fit into this hole?



- Many items may plausibly fit into the hole:

- Challenge: have up to 1 known ground truth region per hole

# Methods

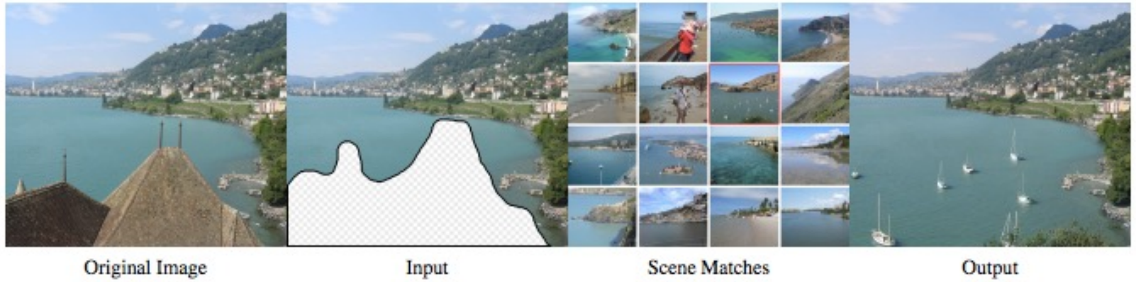
- Before deep learning era: cut-paste from nearest neighbors
- Context encoder
- Guided image inpainting

# Methods

- Before deep learning era: cut-paste from nearest neighbors
- Context encoder
- Guided image inpainting

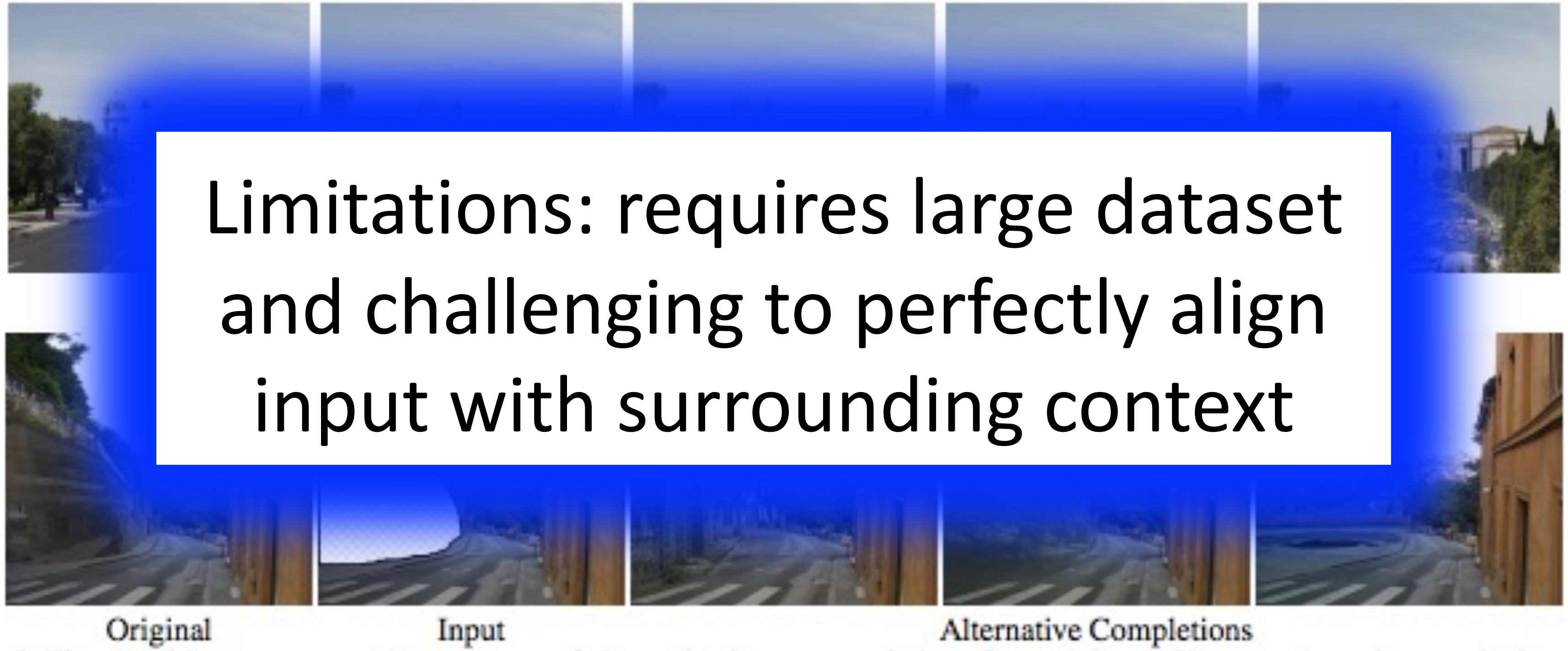
# Cut-Paste from Nearest Neighbors

Idea:




Example:

Example:




Limitations: requires large dataset and challenging to perfectly align input with surrounding context

James Hays & Alexei A. Efros, SIGGRAPH 2007

# Methods

- Before deep learning era: cut-paste from nearest neighbors
- Context encoder
- Guided image inpainting

# Architecture

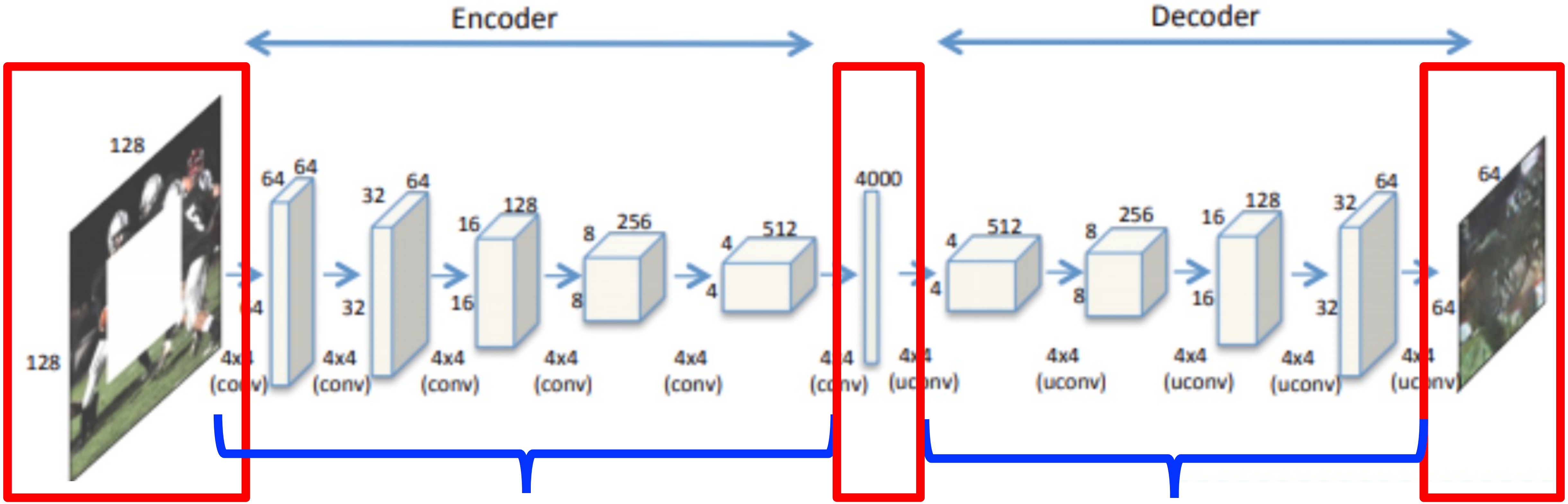



Image with hole

Follows AlexNet architecture

Encoded context

Transposed convolutions

Patch for hole

Figure source: https://medium.com/knowledge-engineering-seminar/context-encoder-image-inpainting-using-gan-ccd6a1ea5fb7

# Training: Loss Functions ($\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\boxed{\mathcal{L}_{rec}}$)

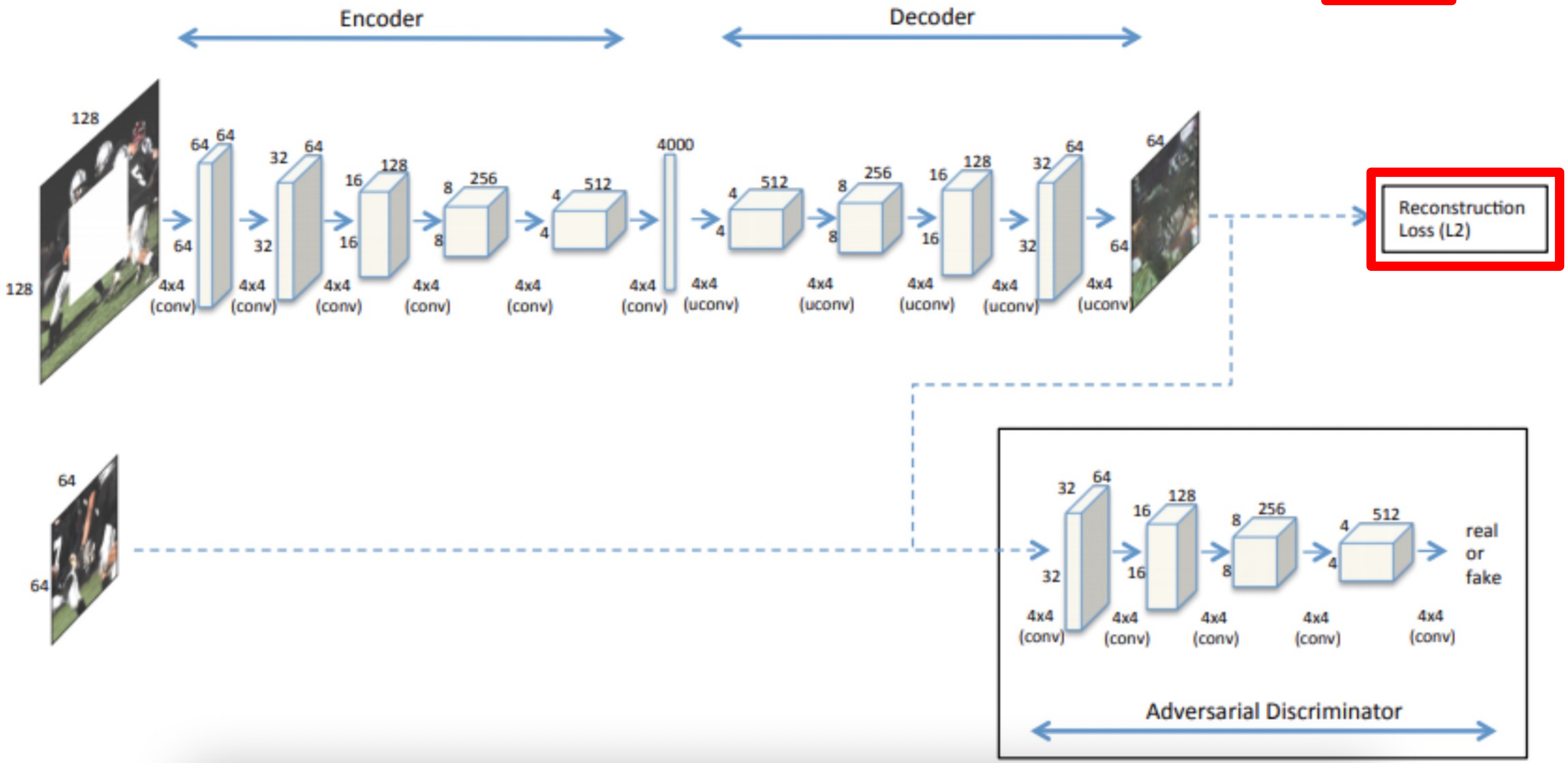
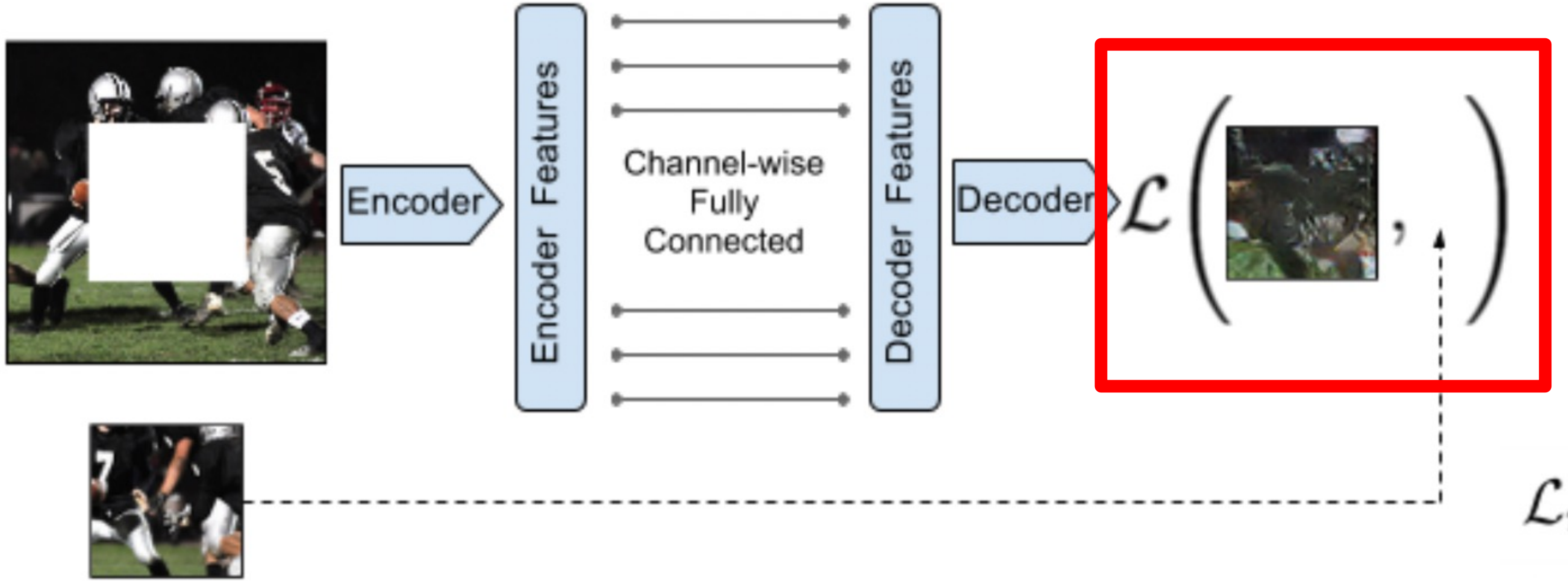



Figure source: https://medium.com/knowledge-engineering-seminar/context-encoder-image-inpainting-using-gan-ccd6a1ea5fb7

# Training: Reconstruction Loss (i.e., Self-Supervised Learning Approach)




Learns to fit into the context by computing the L2 loss to compare the original patch content (P) to the predicted patch content created by the model when given the image with hole (CE(X')).

$$\mathcal{L}_{rec} = \|\mathbf{P} - CE(\mathbf{X}')\|_2^2$$

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

# Training: Reconstruction Loss (i.e., Self-Supervised Learning Approach)

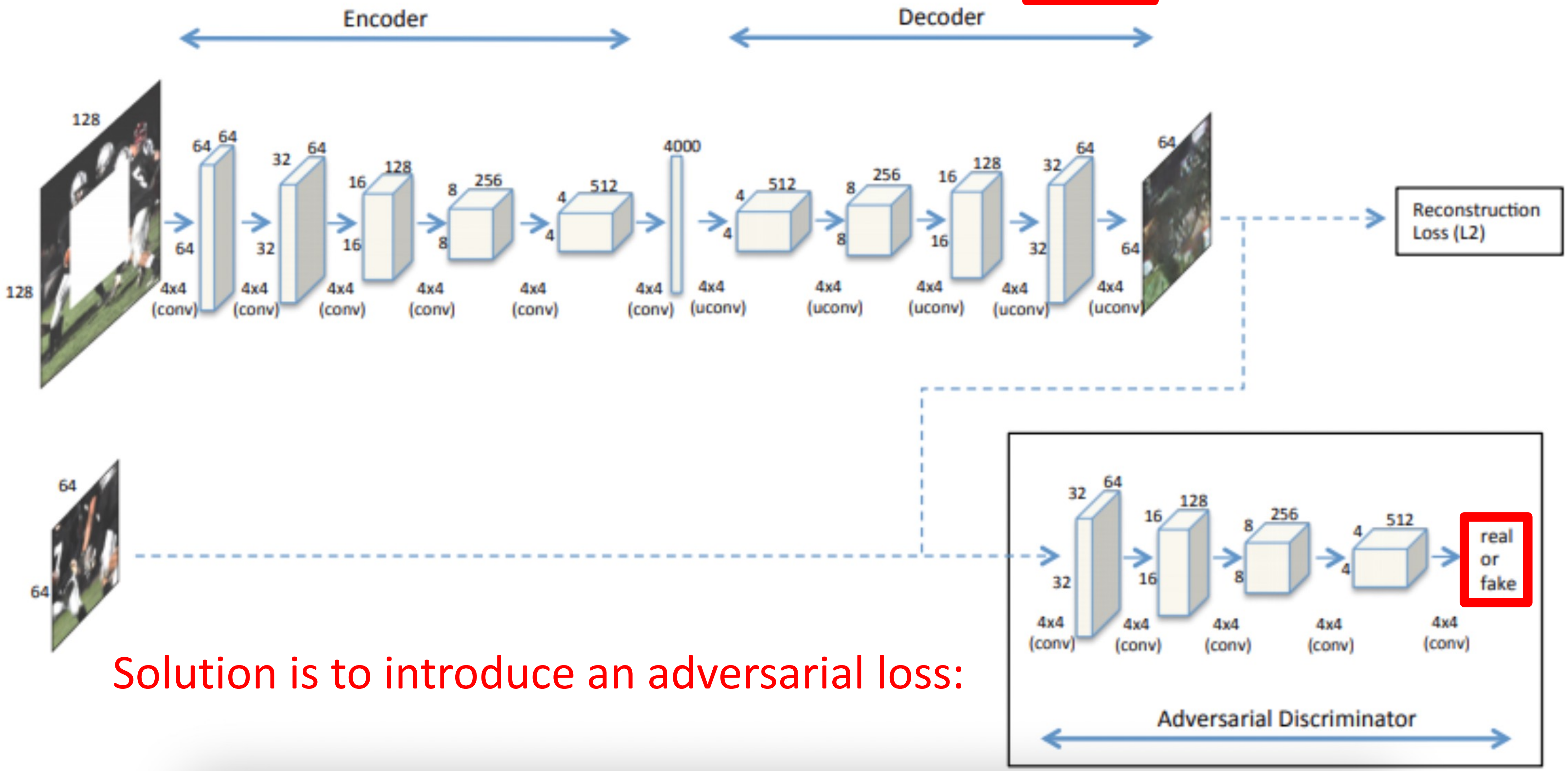
(a) Input context

(c) Context Encoder
($L2$ loss)

Why might training with this loss function alone lead to blurry results?
- It averages the multiple plausible inpaintings for a hole

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

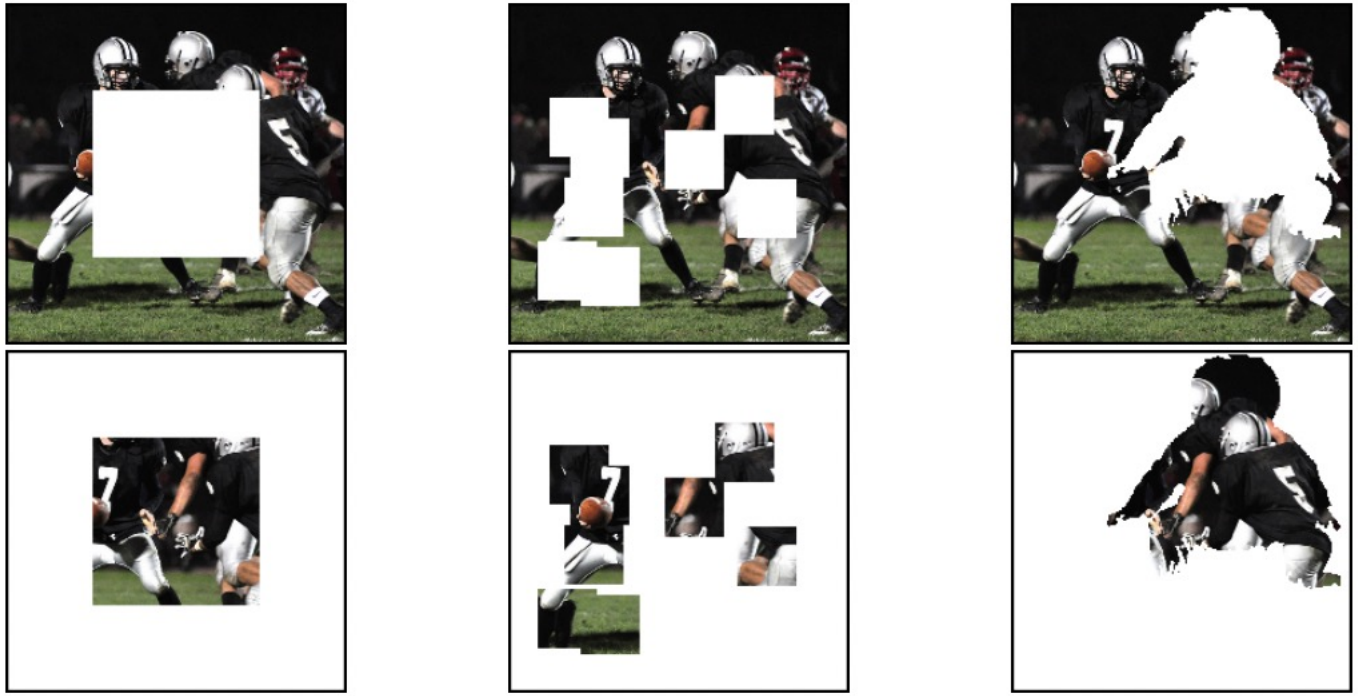
# Training: Loss Functions ($\mathcal{L} = \lambda_{adv}\boxed{\mathcal{L}_{adv}} + \lambda_{rec}\mathcal{L}_{rec}$)




Solution is to introduce an adversarial loss:

Figure source: https://medium.com/knowledge-engineering-seminar/context-encoder-image-inpainting-using-gan-ccd6a1ea5fb7

# Training: Datasets




(a) Central region (b) Random block (c) Random region

Training completed on ImageNet (all 1.2M and a 100K subset) for three hole types

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

# Results Demo

https://www.cs.cmu.edu/~dpathak/context_encoder/

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

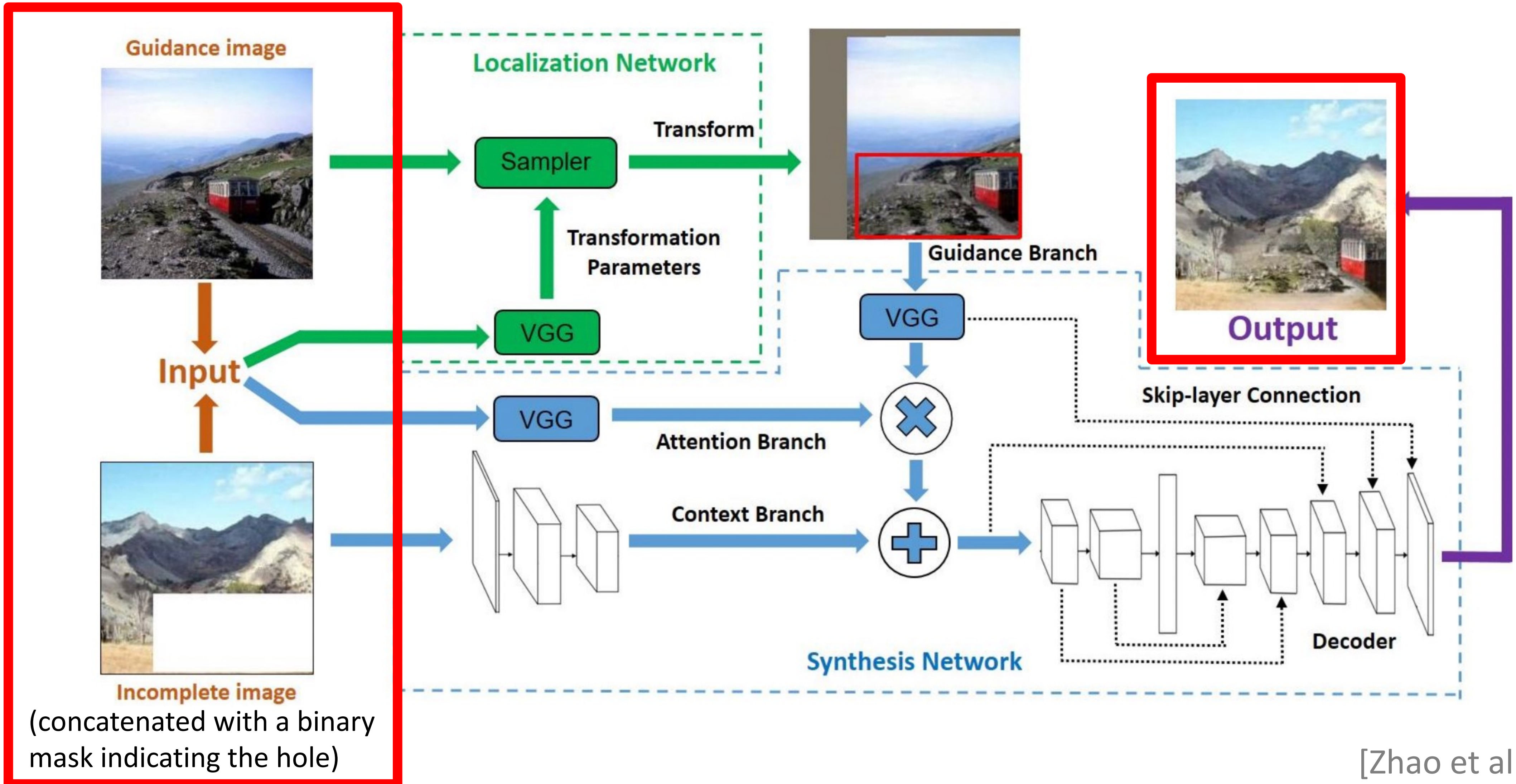
# Key Limitation

Users cannot control what content to insert in the hole

# Methods

- Before deep learning era: cut-paste from nearest neighbors
- Context encoder
- Guided image inpainting

# Architecture



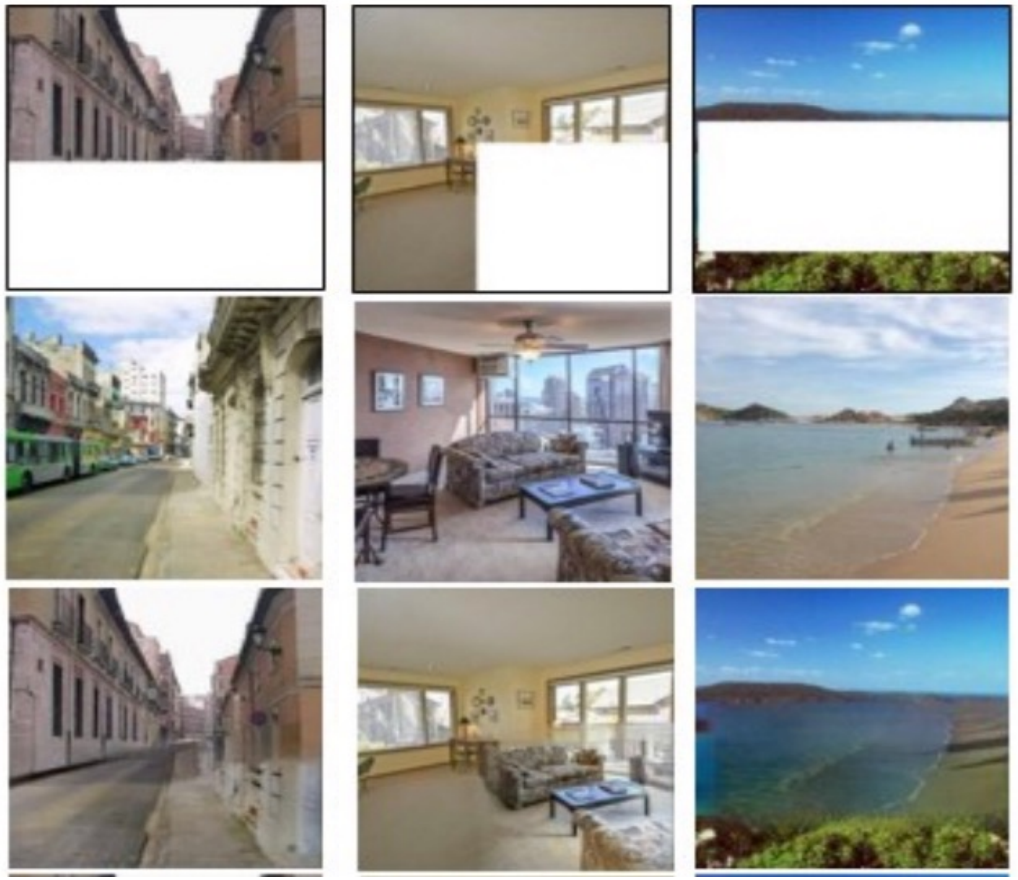

[Zhao et al. WACV 2019]

# Examples of Input and Output

Input Image:

Input user content:

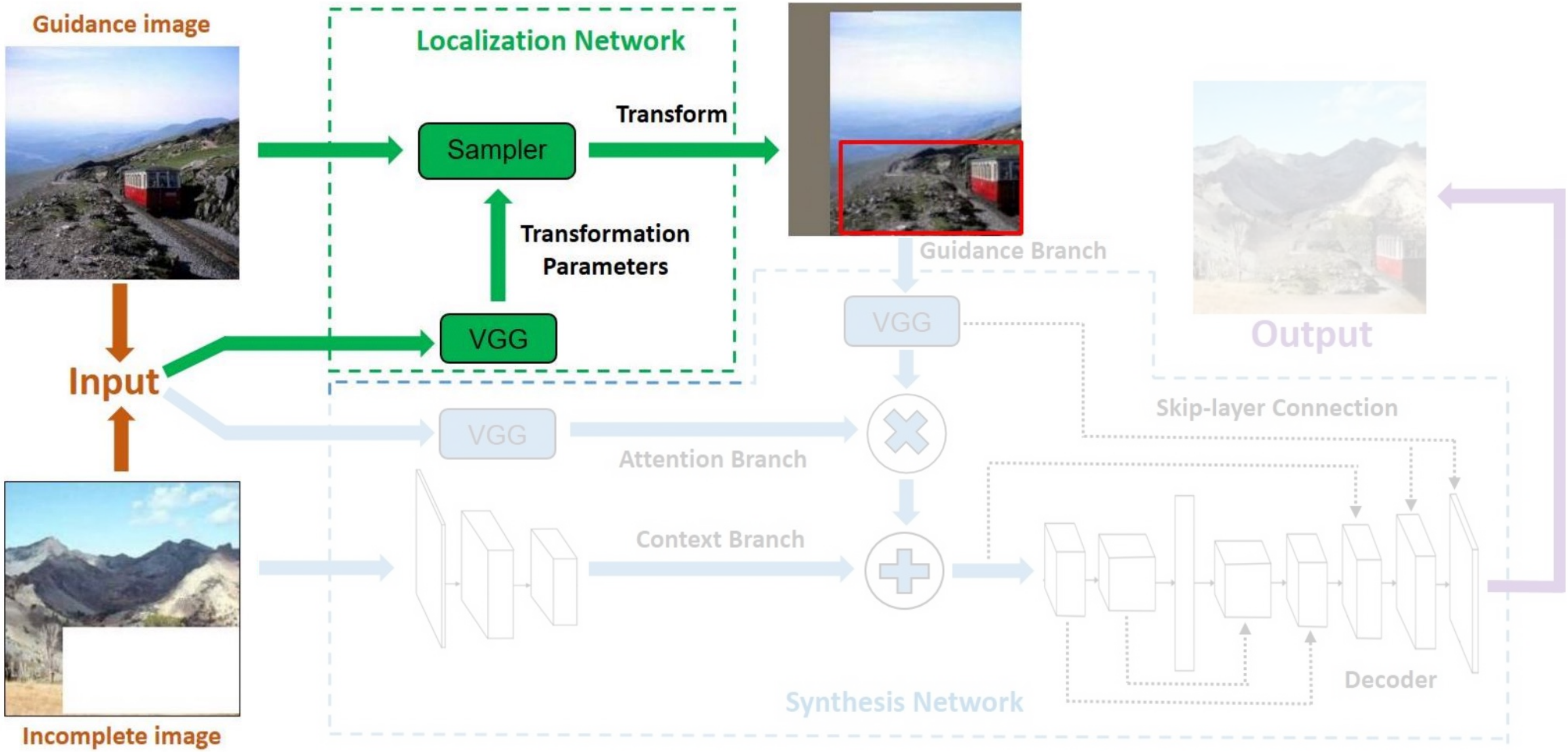
Result:



[Zhao et al. WACV 2019]

# Architecture

Transforms the guidance image with (predicted) transformation parameters to locate which patch to align with the hole of the incomplete image

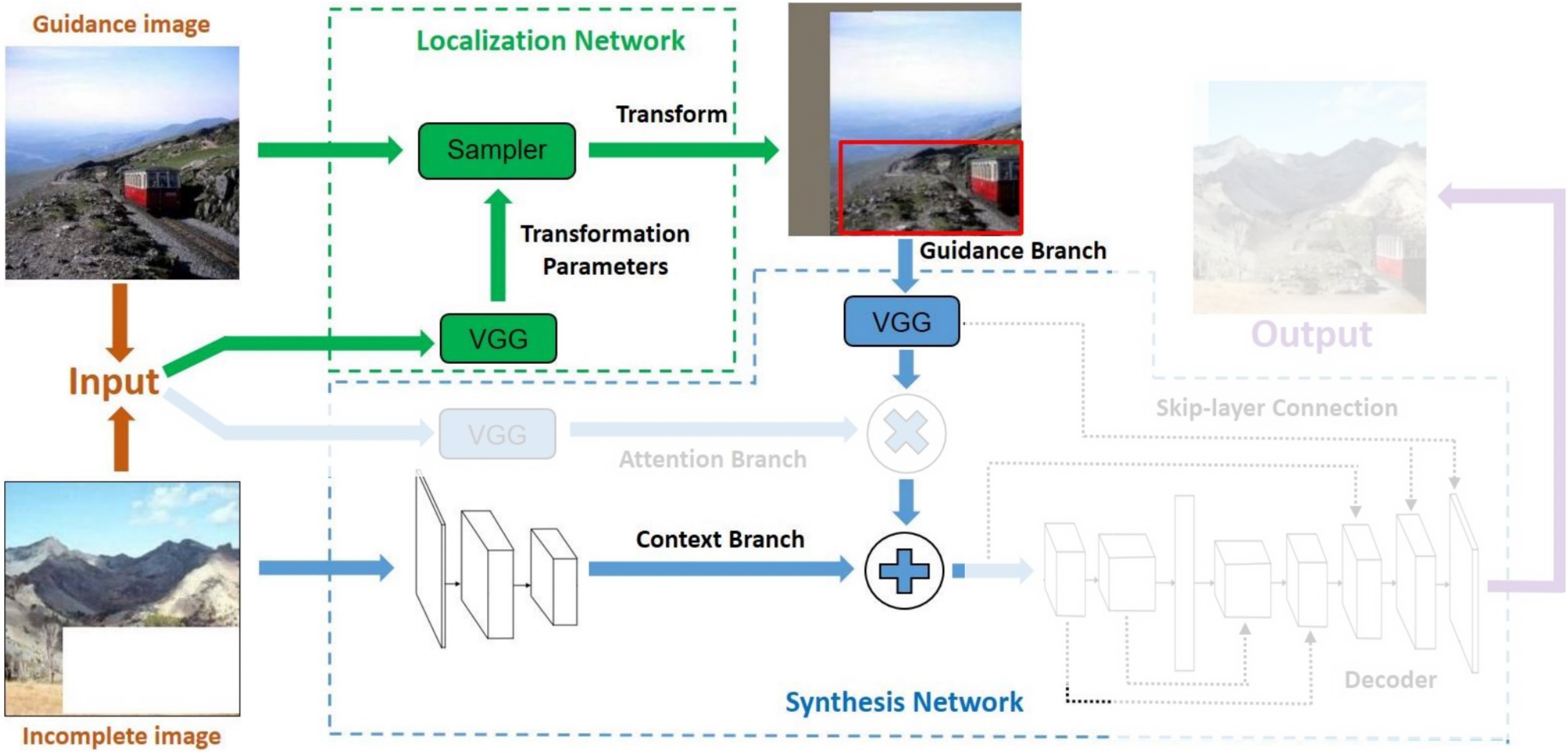



(concatenated with a binary mask indicating the hole)

[Zhao et al. WACV 2019]

# Architecture

High-level features describing the aligned guidance image fused with features describing the incomplete image

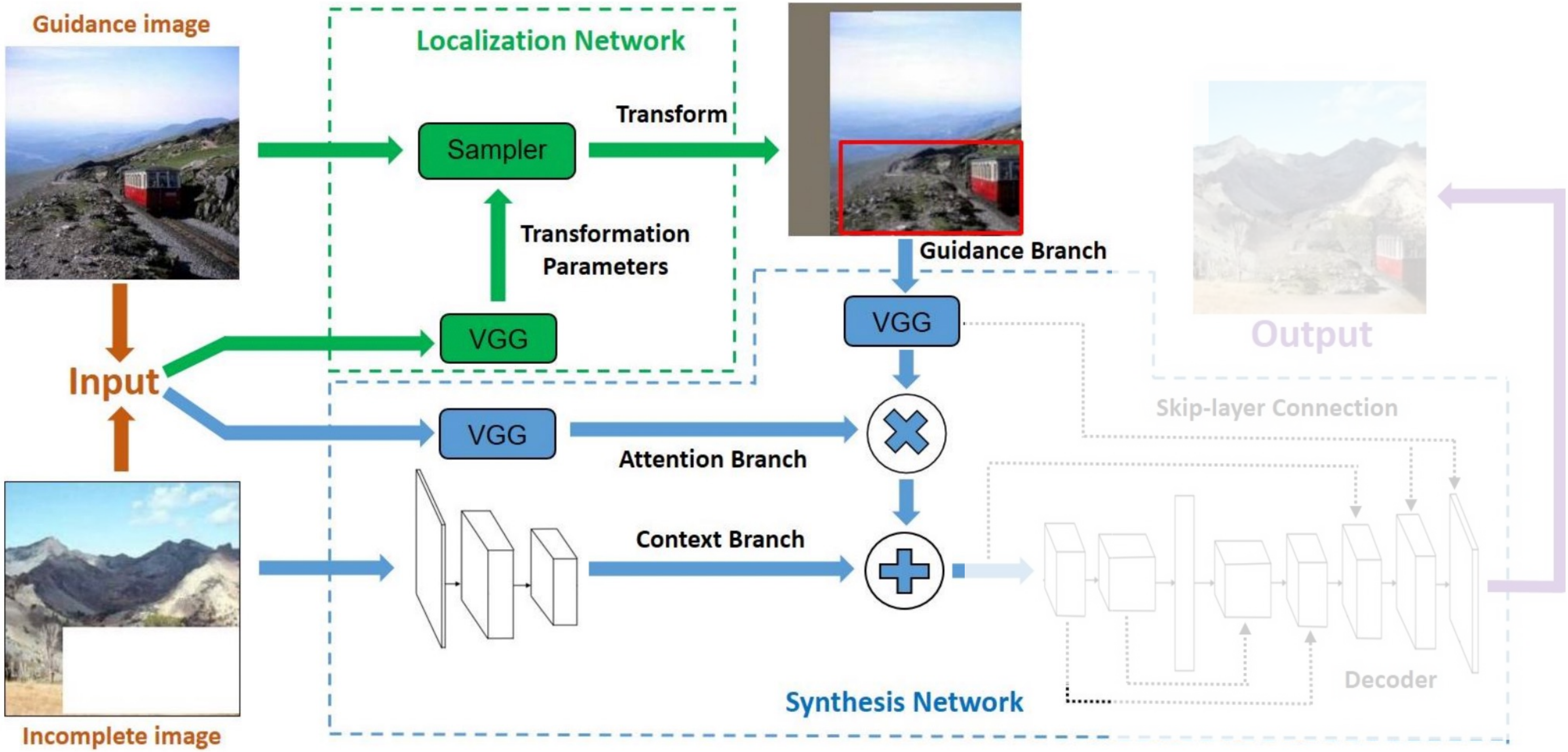



(concatenated with a binary mask indicating the hole)

[Zhao et al. WACV 2019]

# Architecture

Attention branch indicates to what extent regions in the guidance image patch are inconsistent with the context of the incomplete image to indicate whether the decoder should synthesize new content

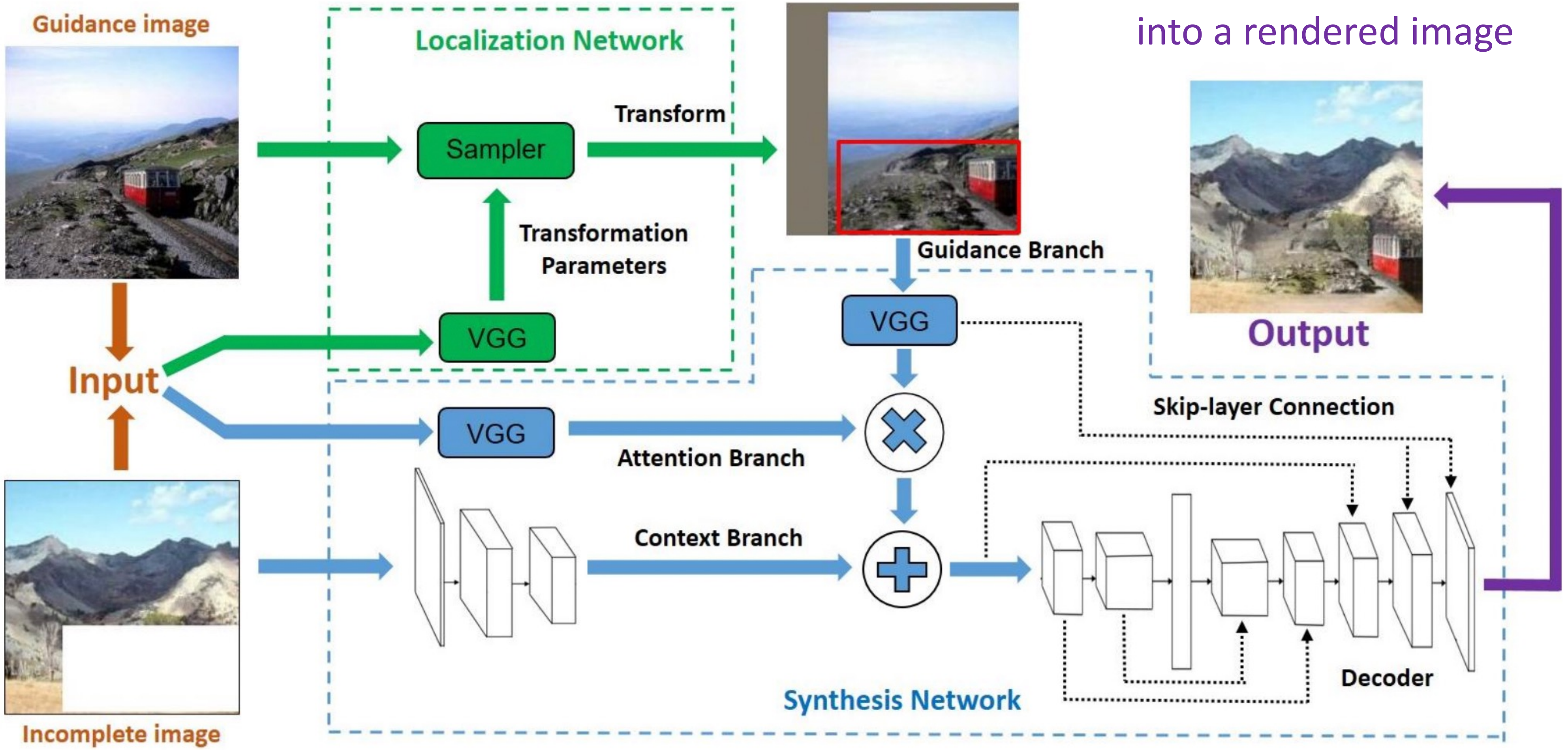



(concatenated with a binary mask indicating the hole)

[Zhao et al. WACV 2019]

# Architecture

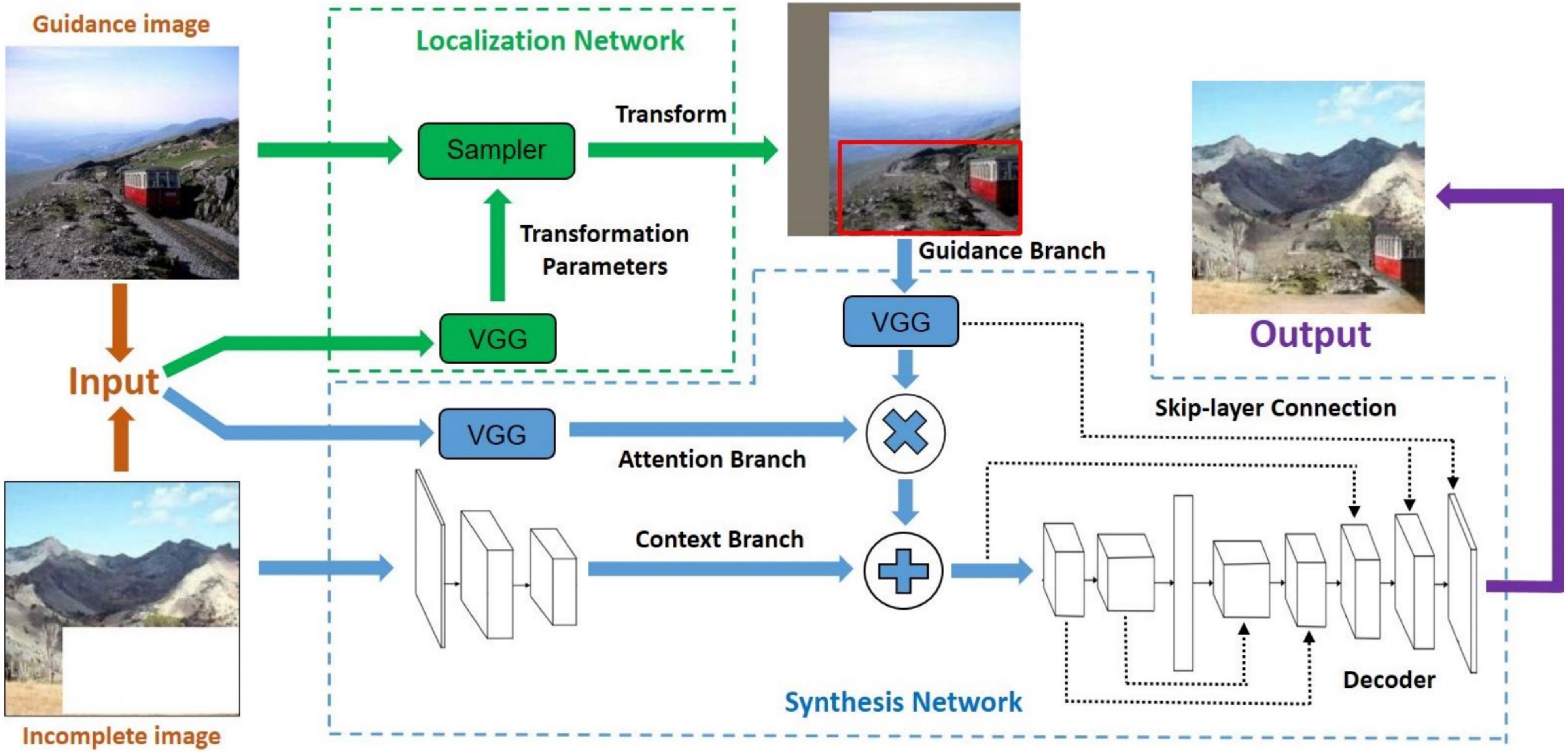
Fused features are decoded into a rendered image




(concatenated with a binary mask indicating the hole)

[Zhao et al. WACV 2019]

# Key Challenge: How to Collect Training Data

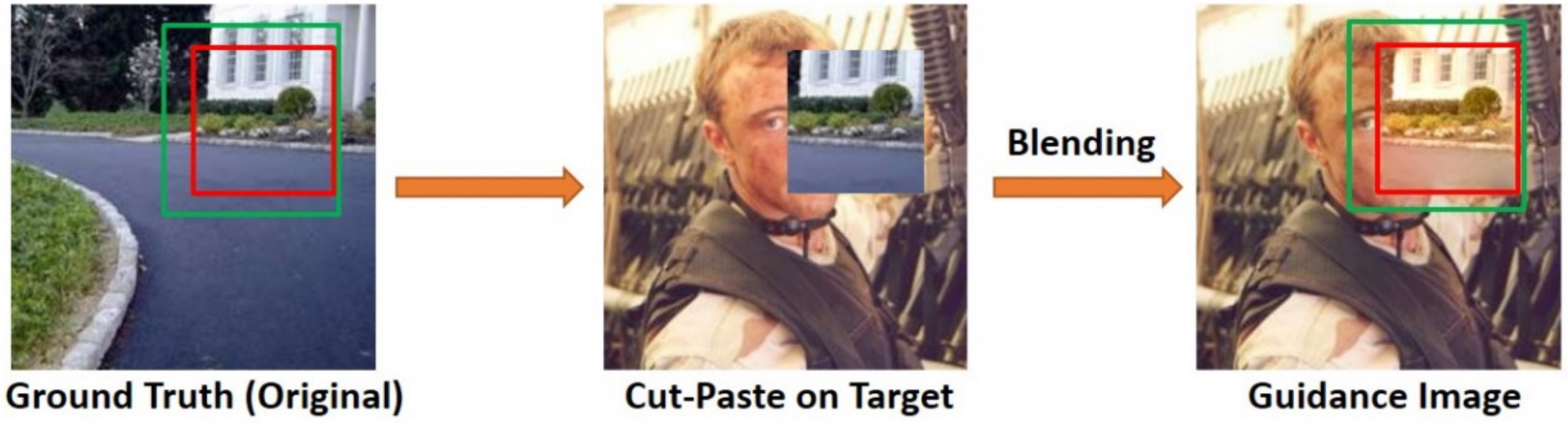



(concatenated with a binary mask indicating the hole)

[Zhao et al. WACV 2019]

# Key Challenge: How to Collect Training Data

Synthesize training data such that we know the true inpainting (i.e., self-supervised learning)




Hole to remove from the original image

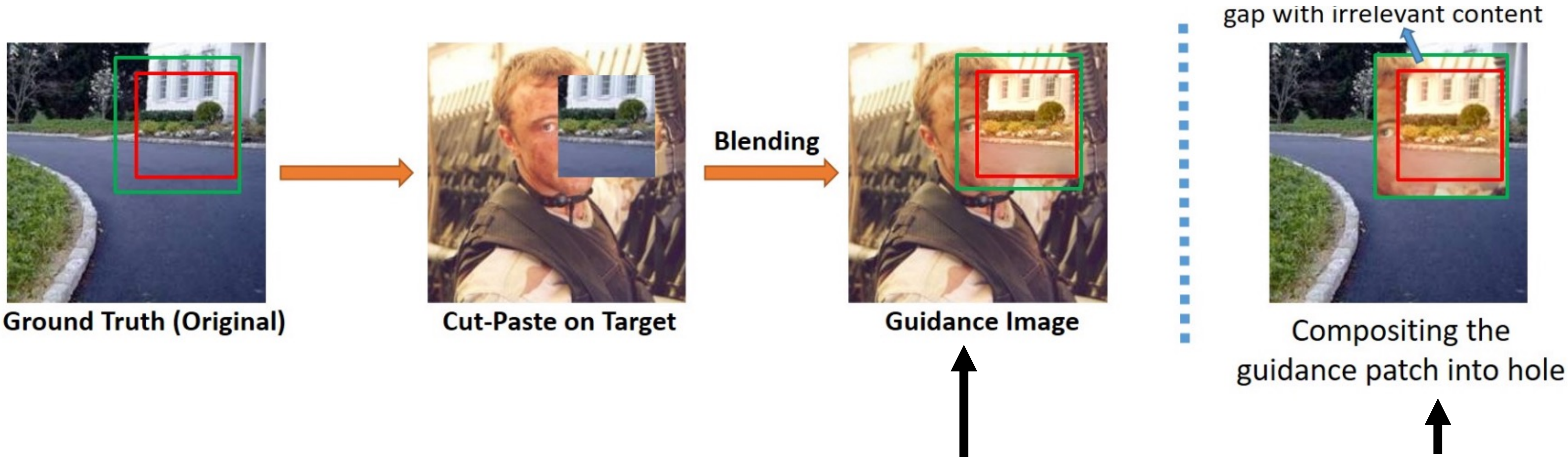
Patch of the original image to corrupt

Guidance image contains original content that is corrupted and content irrelevant to the original image (gap between red and green regions)

[Zhao et al. WACV 2019]

# Key Challenge: How to Collect Training Data

Synthesize training data such that we know the true inpainting (i.e., self-supervised learning)




The localization network must find the patch to use from the guidance image while the synthesis network must then recover the original content, including by synthesizing new content in the gap containing irrelevant content

[Zhao et al. WACV 2019]

# What is a Key Limitation?


Does not account for lighting!

# Methods

- Before deep learning era: cut-paste from nearest neighbors
- Context encoder
- Guided image inpainting

# Today’s Topics

- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# Experiments

| Method | HR[10] | PB[23] | DH[24] | CAF[13] | CE[9] | IM[25] | GLCIC[15] |
|---|---|---|---|---|---|---|---|
| Retrieval (a) | 76% | 76% | 71% | 70% | 70% | 67% | 66% |
| Retrieval (b) | 71% | 73% | 72% | 70% | 73% | 67% | 70% |

Crowd workers rate which generated images look more realistic between the method and baselines

(chance score is 50%)

[Zhao et al. WACV 2019]

# Experiments

| Method | NI | CE[9] | HR[10] | CAF[13] | PB[23] | DH[24] | GLCIC[15] | IM[25] | Ours |
|---|---|---|---|---|---|---|---|---|---|
| Retrieval (a) | 97.7% | 10.0% | 14.0% | 31.0% | 18.0% | 23.0% | 14.0% | 23.0% | **33.0%** |
| Retrieval (b) | 97.7% | 22.0% | 15.0% | 16.0% | 20.0% | 22.0% | 12.0% | 27.0% | **36.0%** |

Crowd workers indicate if images look realistic independently for the method and baselines when also shown real images

[Zhao et al. WACV 2019]

# Today’s Topics

- Problem
- Applications
- Image generation methods
- Hole filling methods
- Evaluation approaches

# The End