Interpreting and Explaining Deep Neural Networks: A Perspective on Time Series Data – Part 2/3

Jaesik Choi

Explainable Artificial Intelligence Center
Graduate School of Artificial Intelligence
KAIST

Some slides courtesy of David Bau and M. Pawan Kumar

Interpreting and Explaining Deep Neural Networks: A Perspective on Time Series Data

Agenda (150 min)

Overview to Explainable Artificial Intelligence (XAI) – 15 min

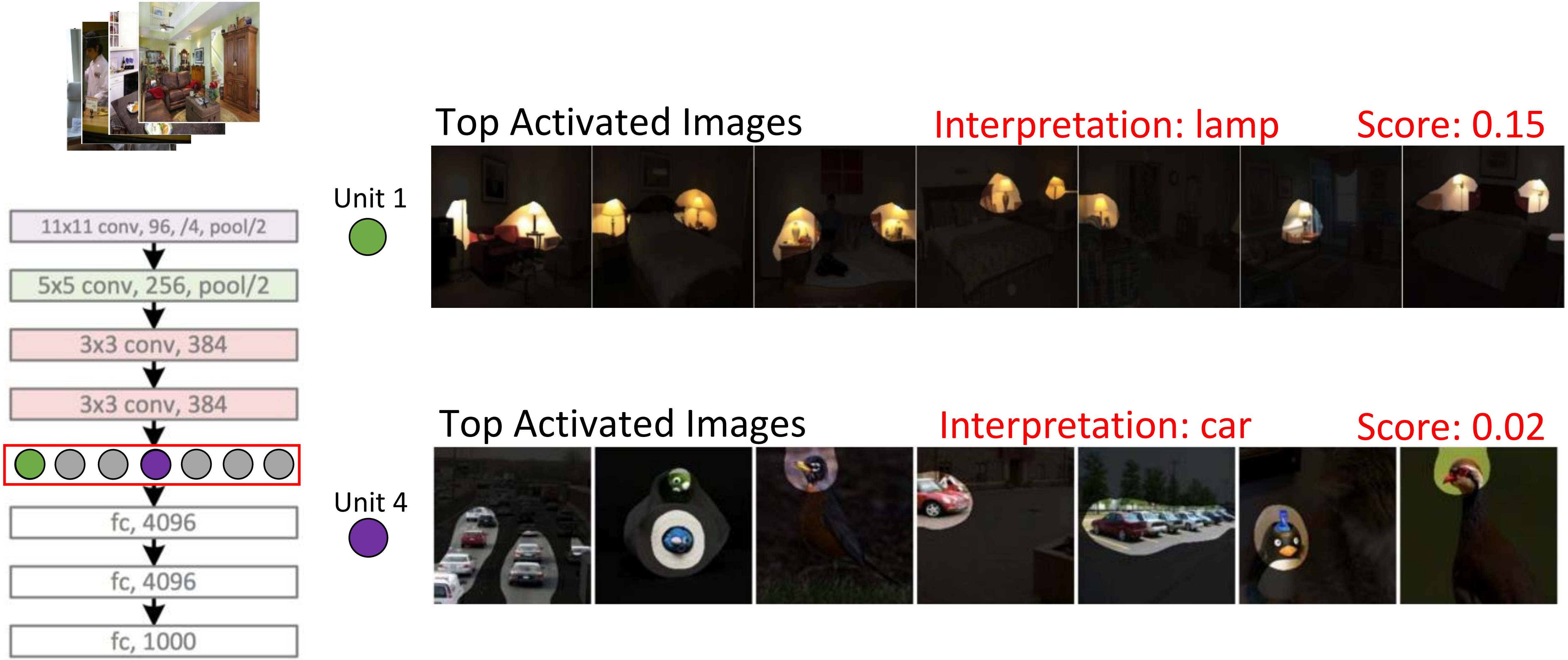
Input Attributions Methods for Deep Neural Networks – 35 min

Interpreting Inside of Deep Neural Networks – 50 min

- Network Dissection
- GAN Dissection
- Explorative Generative Boundary Award Sampling

[10 min break]

Explainable Models for Time Series Data – 50 min



Goal: From Visualization to Interpretation

David Bau et. al., 2017

Unit 1 Top activated images

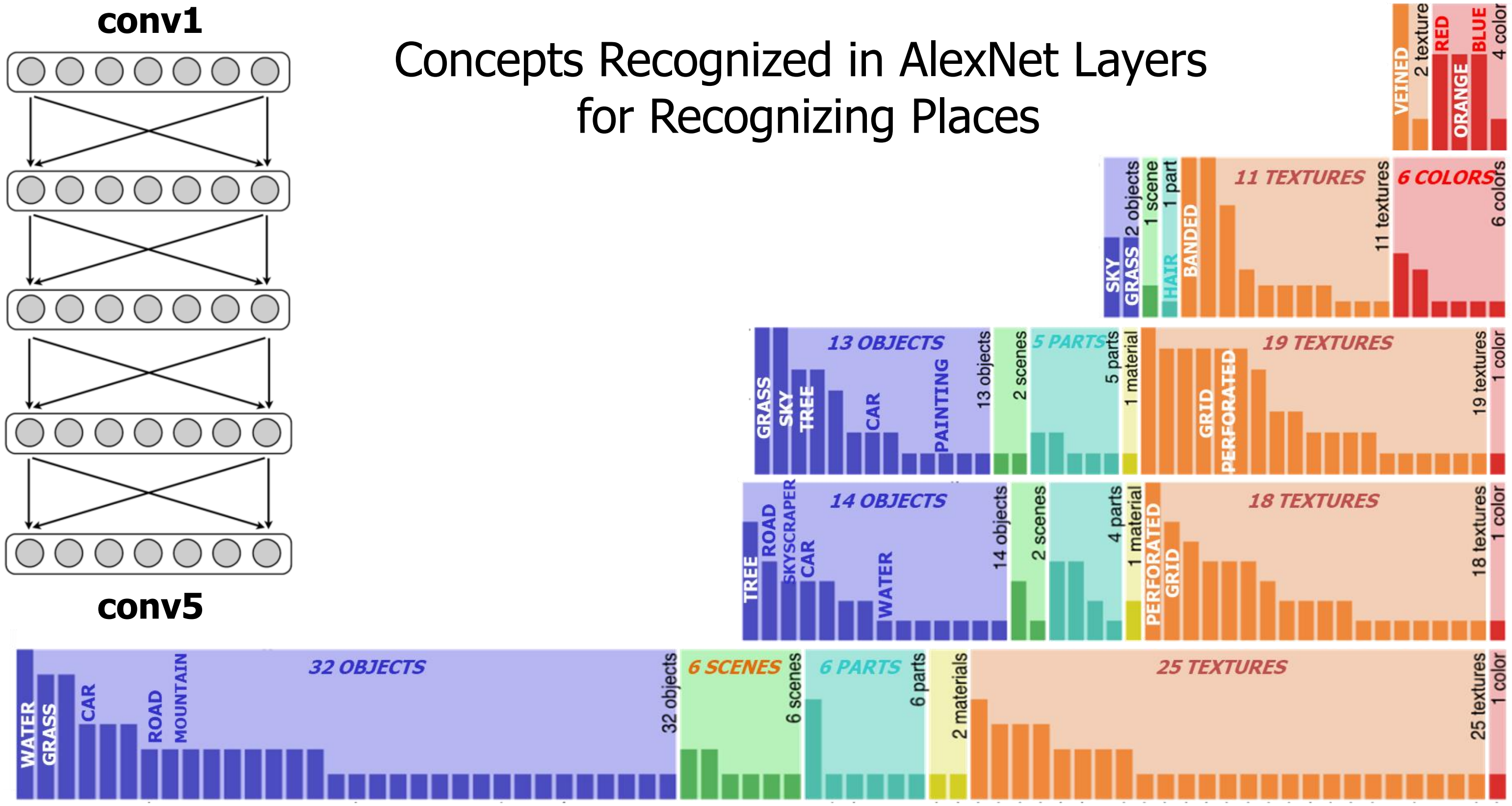
Lamp Intersection over Union (IoU)= 0.12

Approach: Test units for semantic segmentation

David Bau et. al., 2017

Network Dissection David Bau et. al., 2017

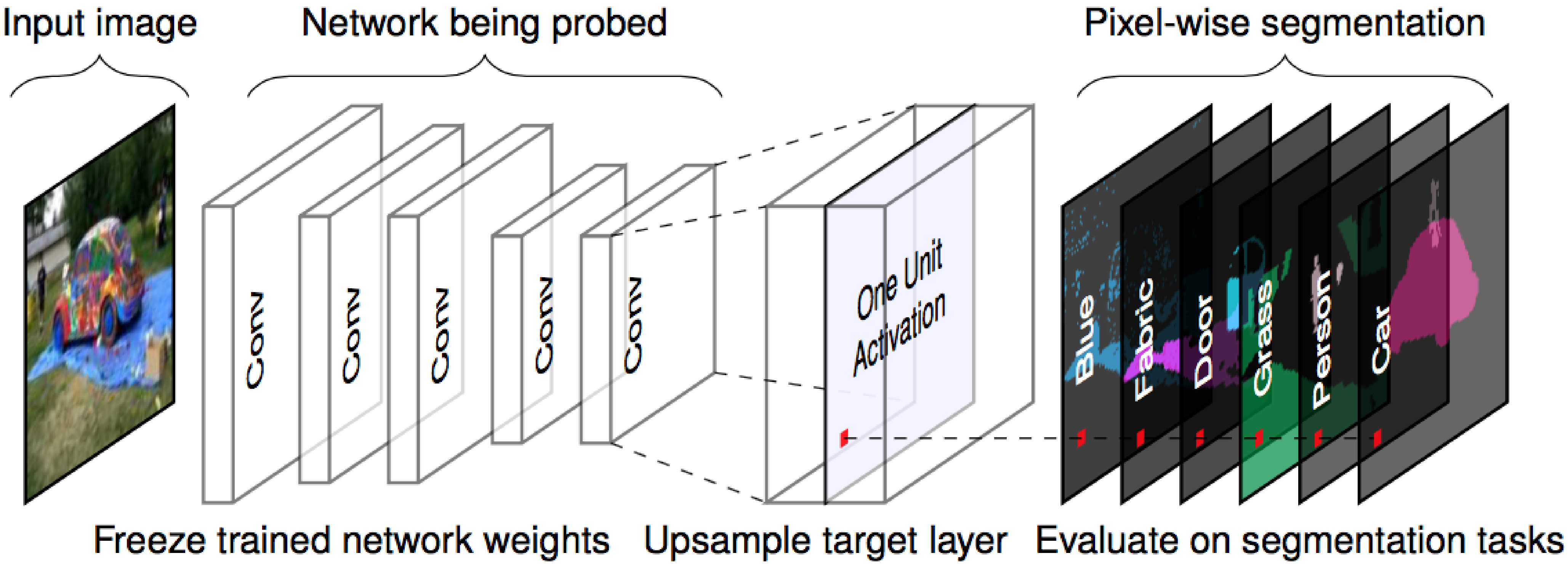
Concepts Recognized in AlexNet Layers
for Recognizing Places



Dissecting Deep Neural Networks

David Bau et. al., 2017

Network Dissection David Bau et. al., 2017

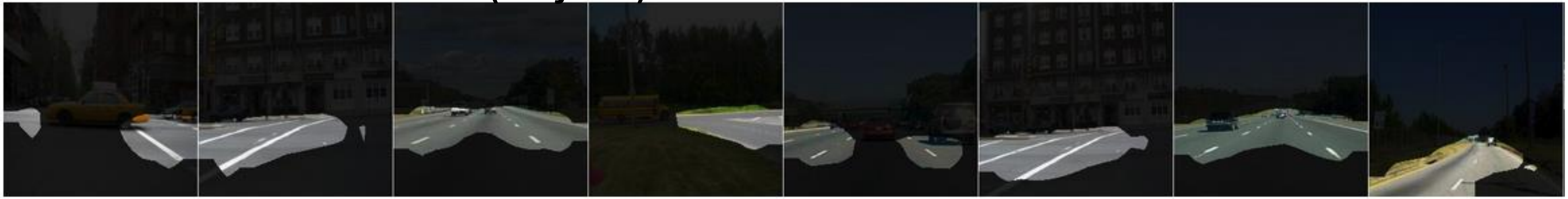


Dissecting Deep Neural Networks

David Bau et. al., 2017

conv5 unit 79 car (object) IoU=0.13

conv5 unit 107 road (object) IoU=0.15



AlexNet trained on place dataset

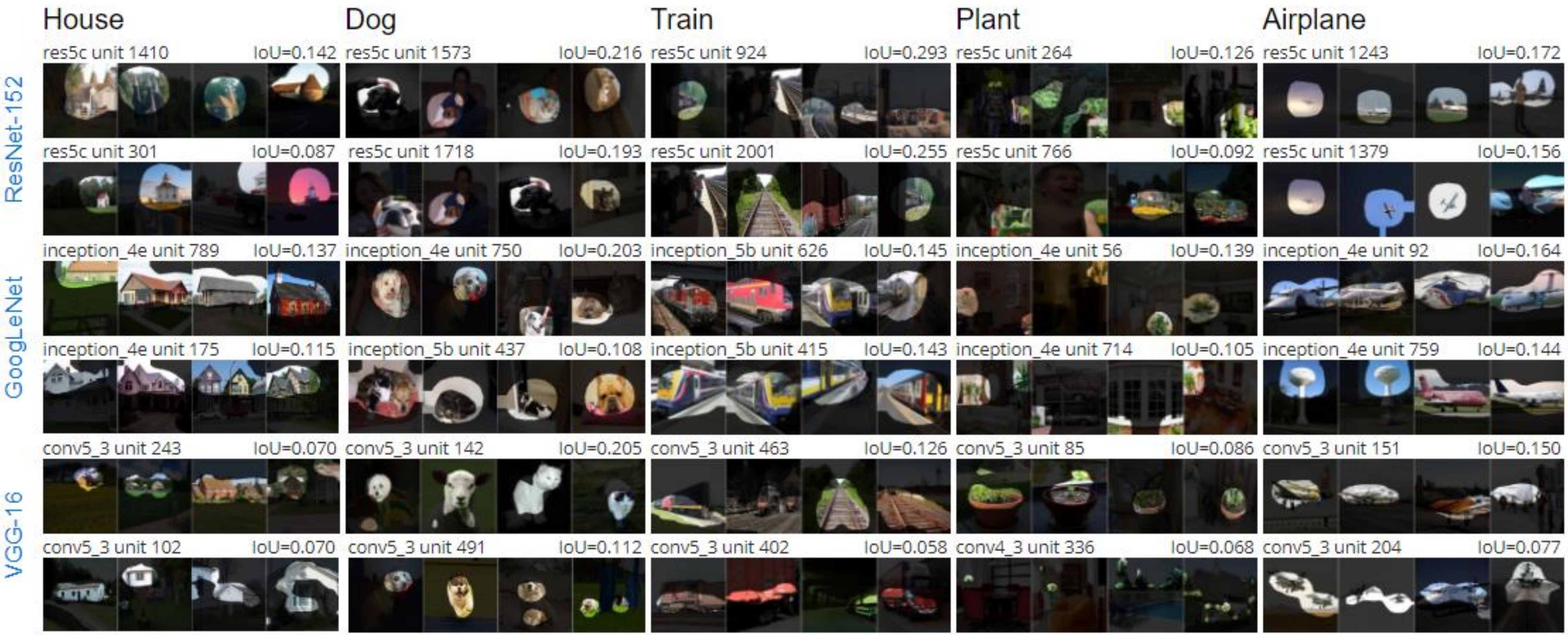
David Bau et. al., 2017

conv5 unit 144 mountain (object) IoU=0.13

conv5 unit 200 mountain (object) IoU=0.11

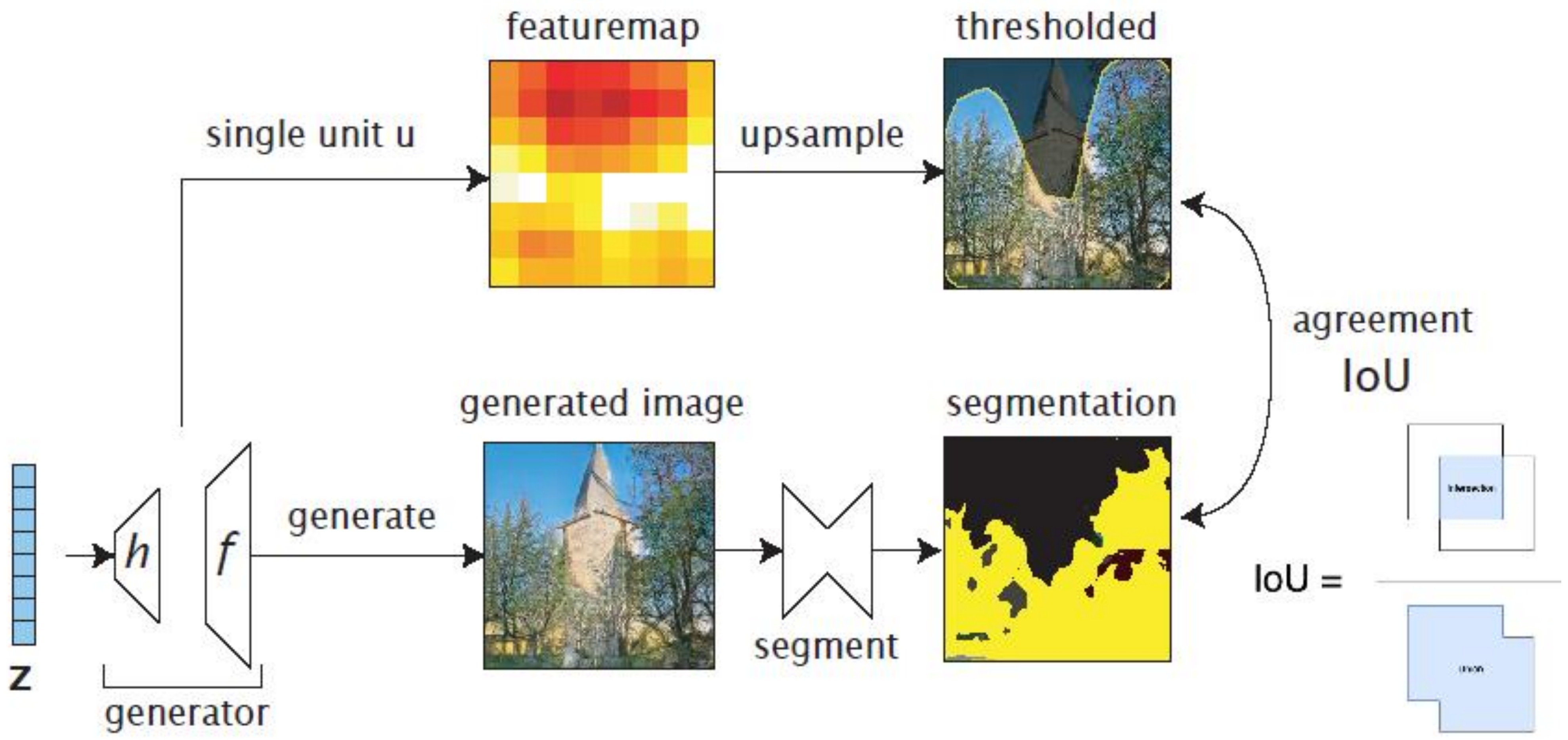
AlexNet trained on place dataset

David Bau et. al., 2017



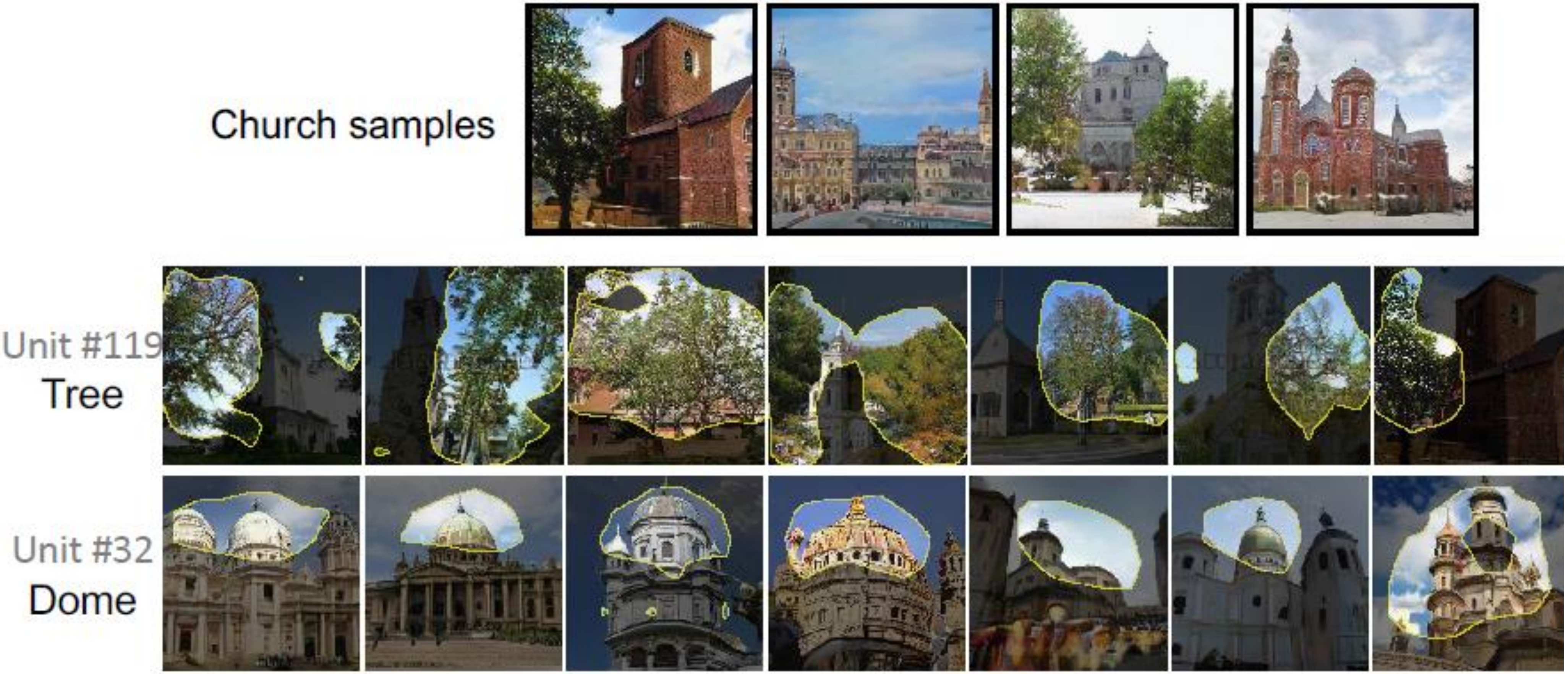
Dissecting Deep Neural Networks

David Bau et. al., 2017



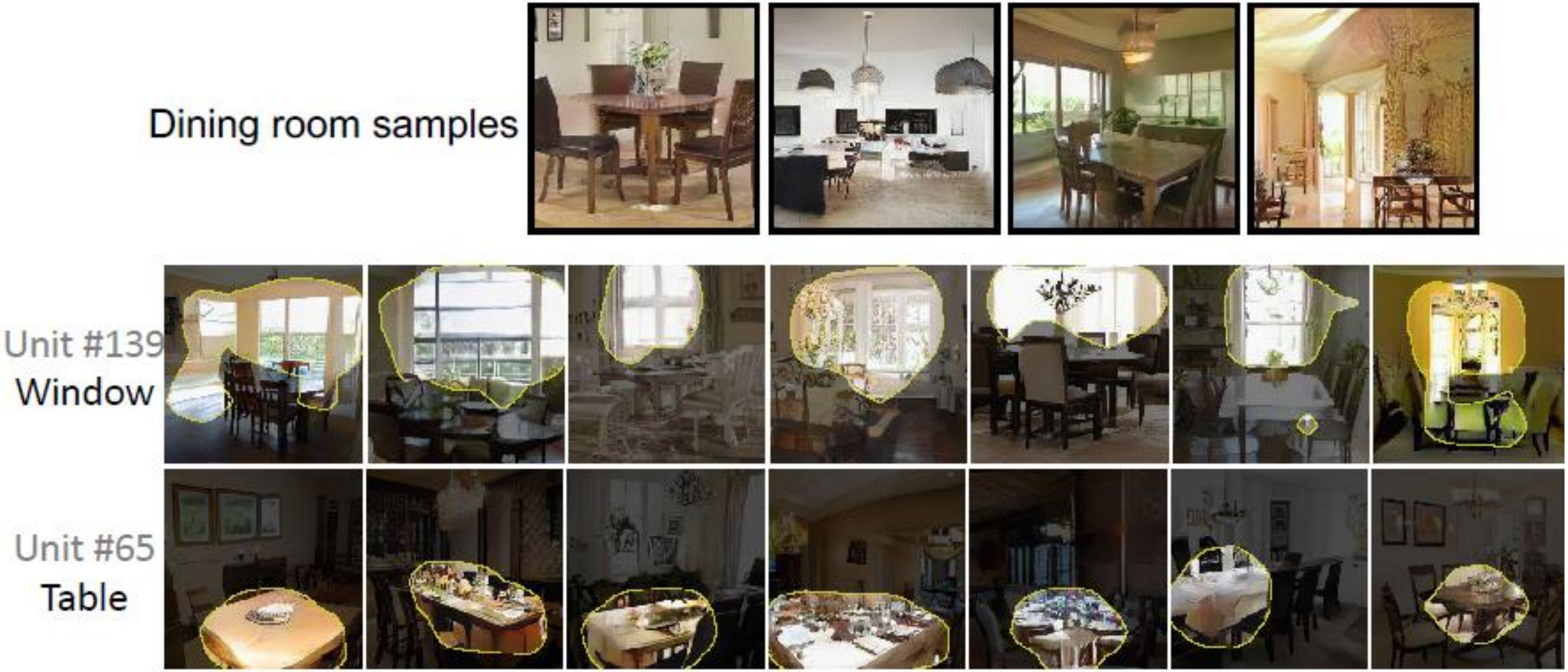
GAN Dissection – Dissecting explainable units in a GAN

David Bau et. al., 2019



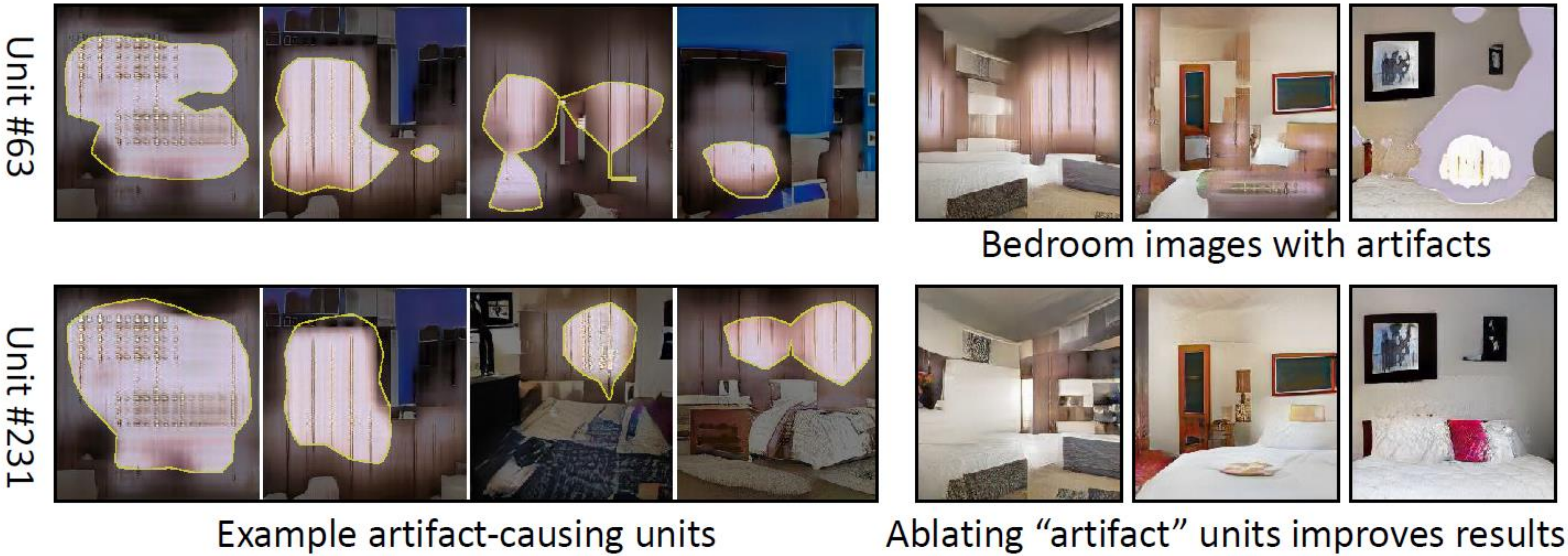
GAN Dissection – Do units correlate to an object class?

David Bau et. al., 2019



GAN Dissection – Do units correlate to an object class?

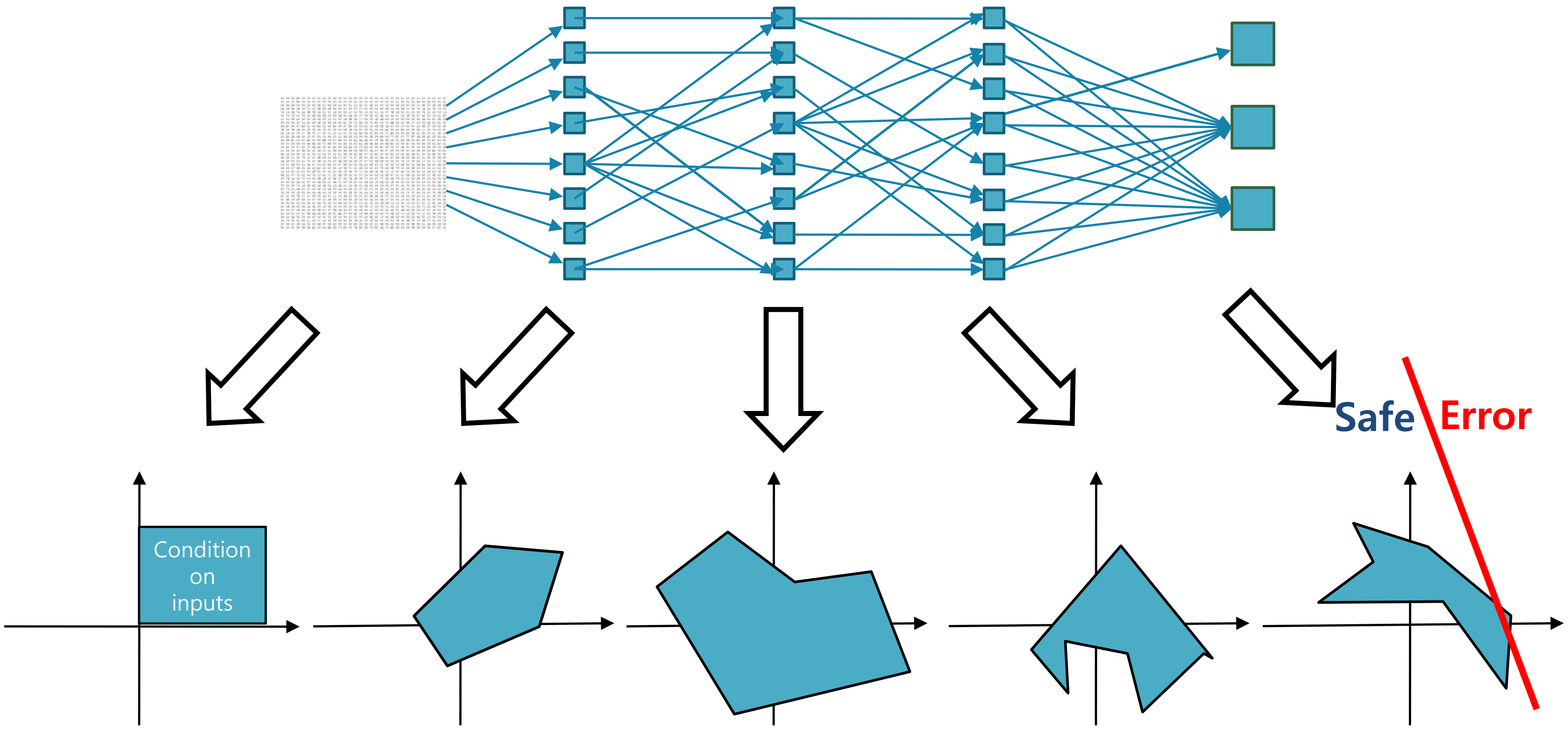
David Bau et. al., 2019



GAN Dissection – Debugging and Improving GANs

David Bau et. al., 2019

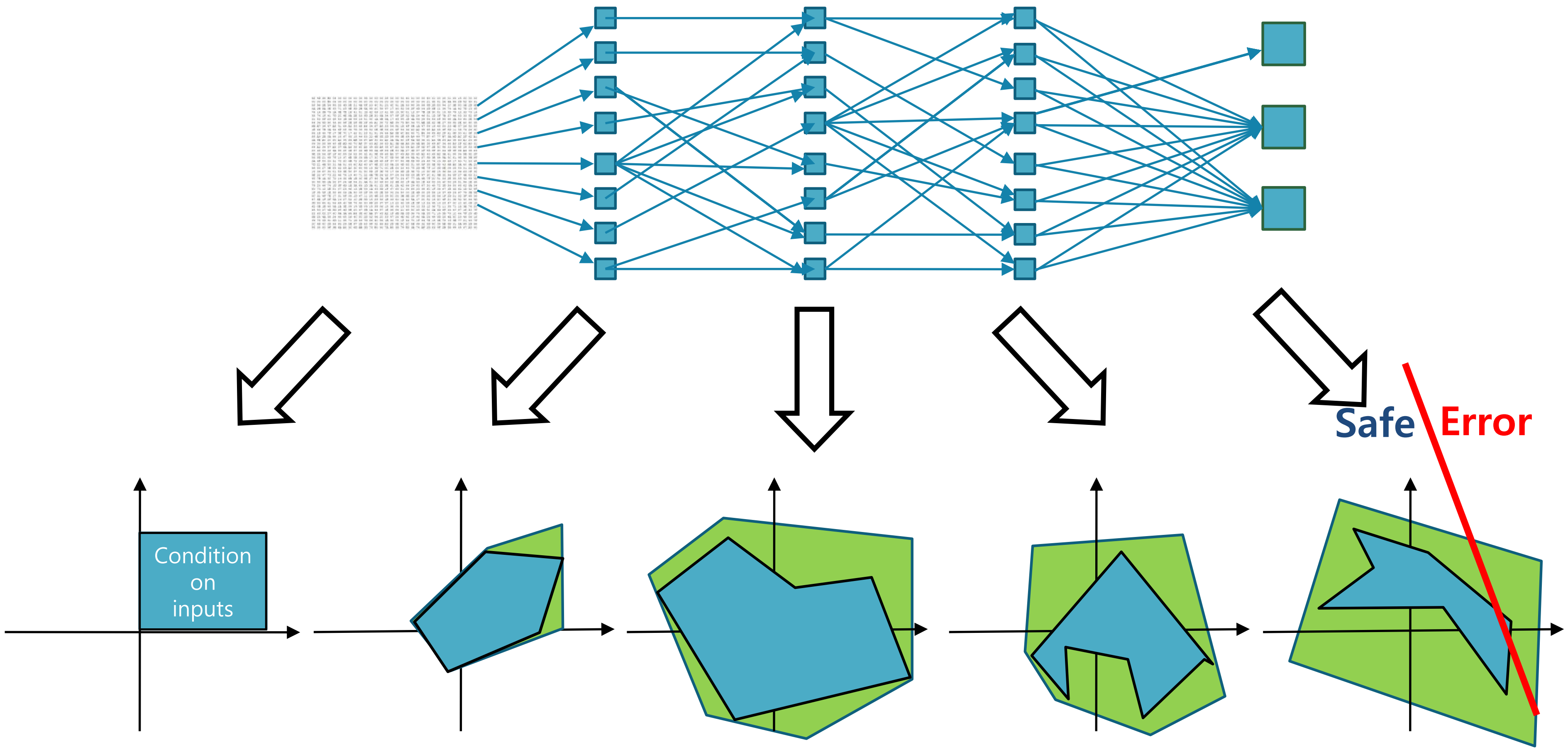
Is there an erroneous output?



Non-convexity makes the problem NP-hard

[Slide courtesy of M. Pawan Kumar]

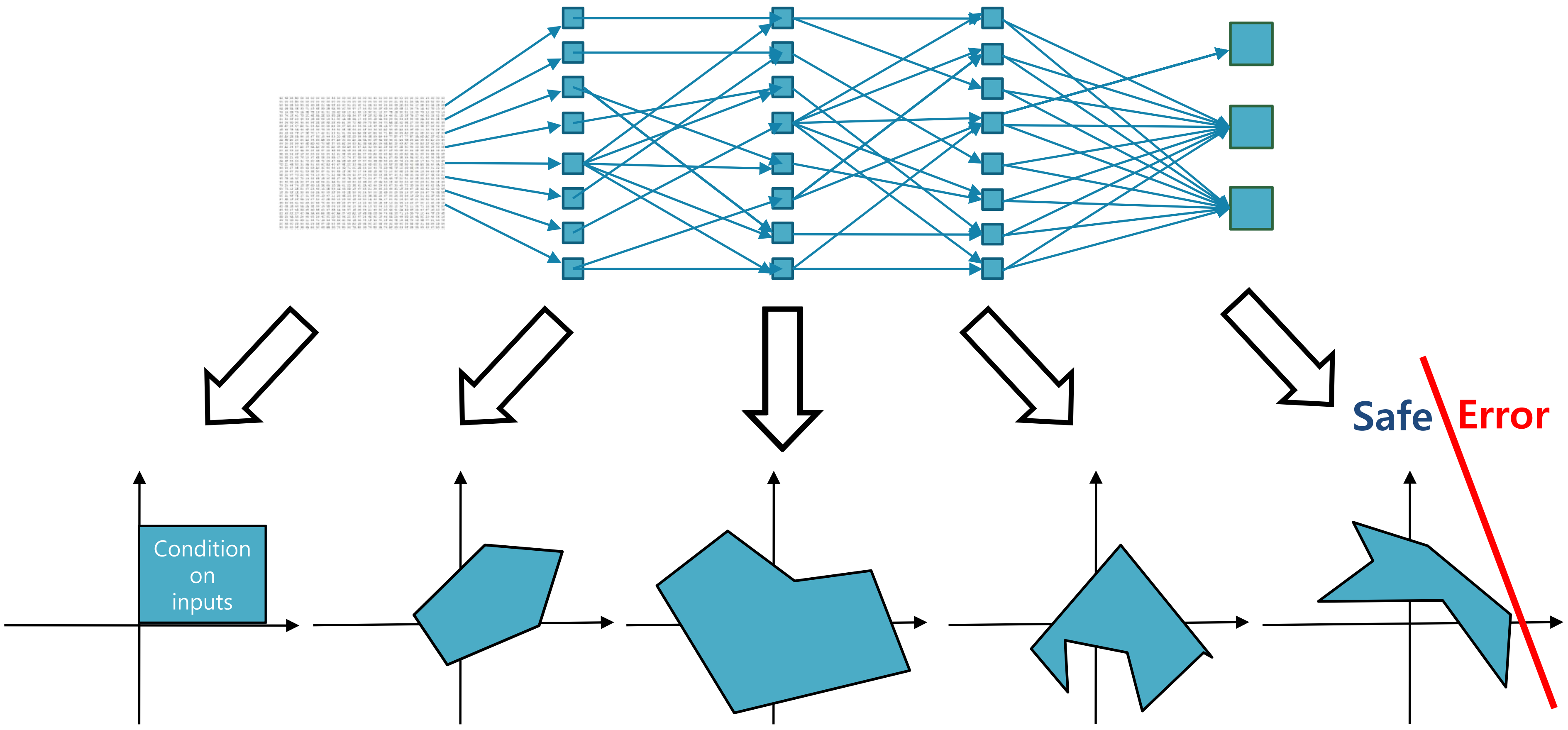
Is there an erroneous output?



Replace by a convex superset

[Slide courtesy of M. Pawan Kumar]

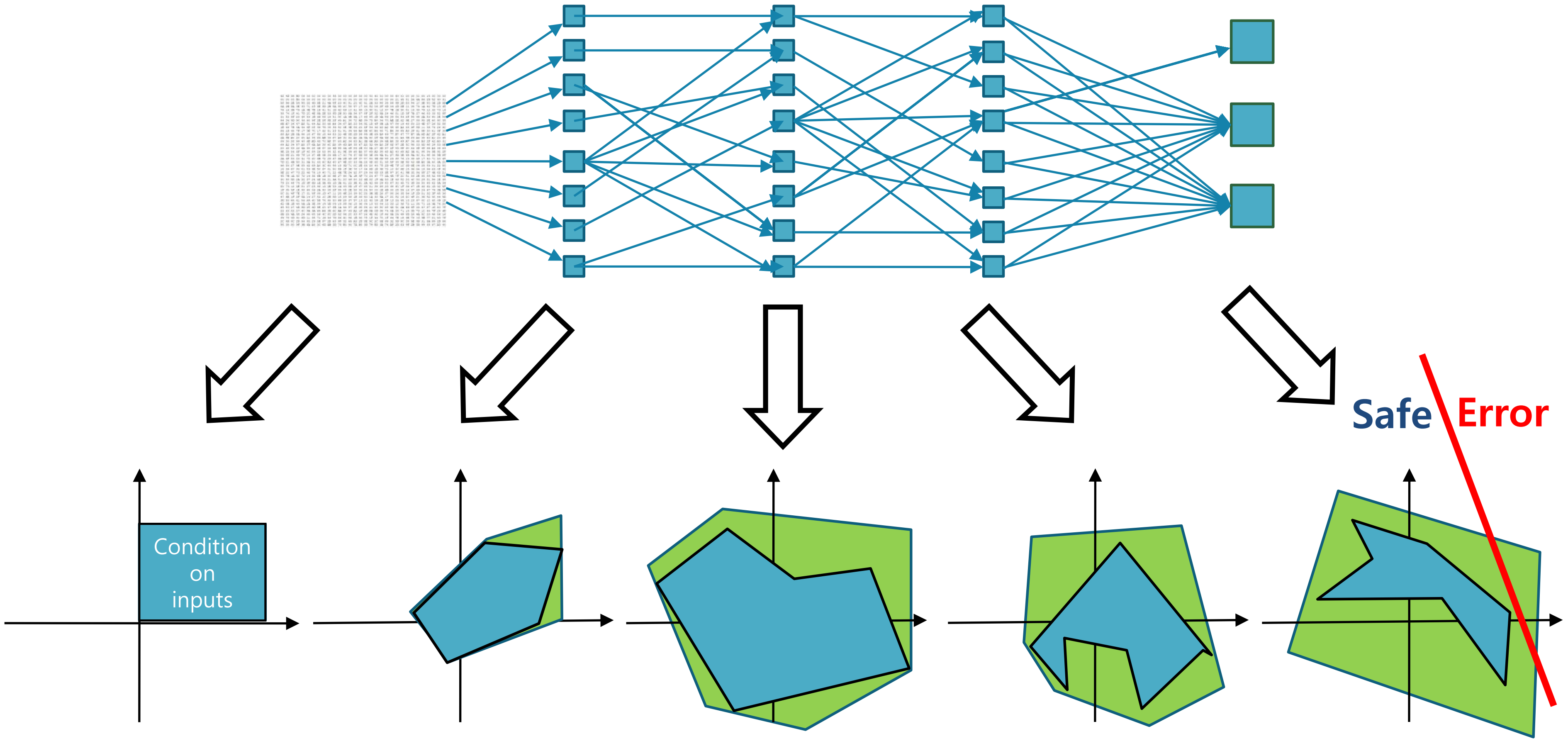
Is there an erroneous output?



Supose, non-convex set has no erroneous output

[Slide courtesy of M. Pawan Kumar]

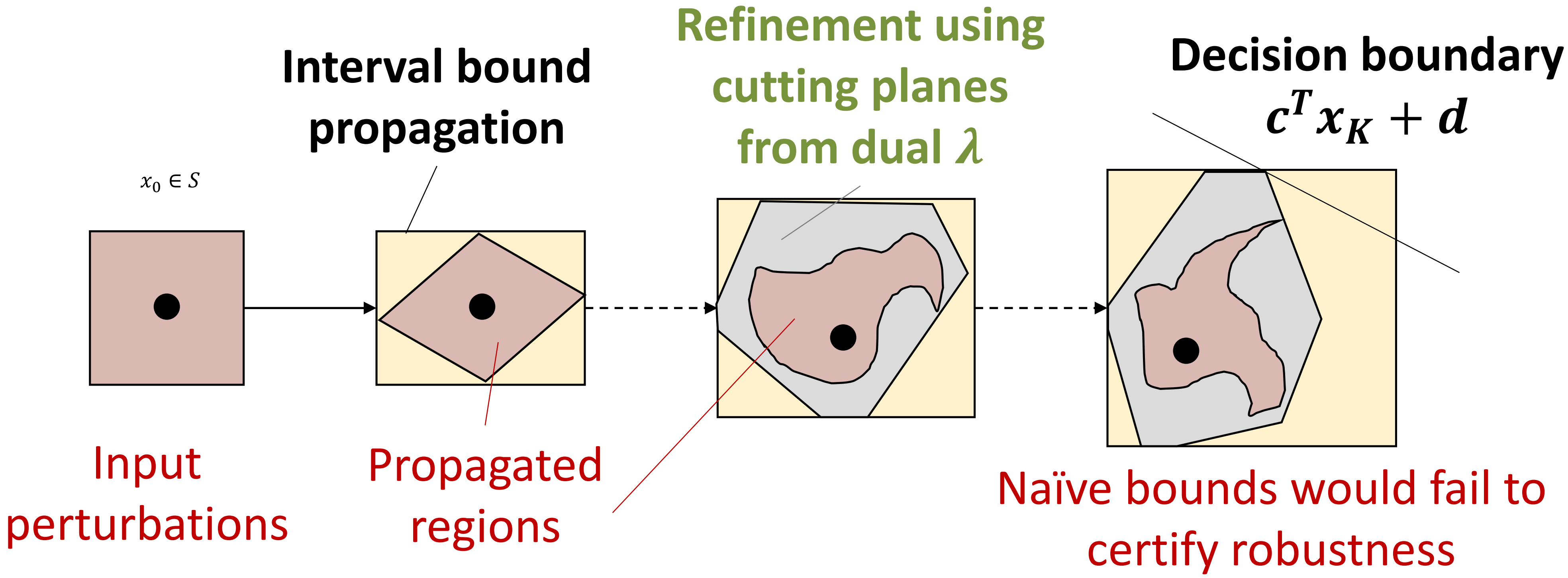
Is there an erroneous output?



Convex superset might give incorrect answer

[Slide courtesy of M. Pawan Kumar]

Finding a tight convex bound with Lagrangian relaxed decision boundary



[Dvijotham et al., 2018]

Neural Networks Verification – An Incomplete Method

Notations:

$x^{in}(= x^0)$: the input to the neural network
z^l: pre-activations of neurons at layer l
x^l: the vector of neural activations after application of the activation to z^{l-1}
h^l: the activation function at layer l

$\underline{x^l}$, $\overline{x^l}$: upper and lower bounds of x^l

[Dvijotham et al., 2018]

Notations:

$\mathrm{x}^{\mathrm{in}}(=\mathrm{x}^{0})$: the input to the neural network
z^{l}: pre-activations of neurons at layer l
x^{l}: the vector of neural activations after application of the activation to $\mathrm{z}^{\mathrm{l}-1}$

- $\mathrm{x}^{\mathrm{l}}(\mathrm{x}^{\mathrm{in}}), \mathrm{z}^{\mathrm{l}}(\mathrm{x}^{\mathrm{in}})$: the activations at the l-th layer

h^{l}: the activation function at layer l

$\underline{\mathrm{x}^{\mathrm{l}}}, \overline{\mathrm{x}^{\mathrm{l}}}$: upper and lower bounds of x^{l}

[Dvijotham et al., 2018]

A verification problem

$\mathrm{x}^{\mathrm{nom}}$: a nominal input
$S_{\mathrm{in}}(\mathrm{x}^{\mathrm{nom}})$: constrained subset of inputs induced by the nominal input
S_{out}: the constraints on the output that we would like to verify are true

$$\forall x^{in} \in \mathcal{S}_{in}\left(x^{nom}\right) \quad x^{L}(x^{in}) \in \mathcal{S}_{out}$$

[Dvijotham et al., 2018]

A verification problem

When S_{out} is presented with a finite set of linear constraints as

$$\mathcal{S}_{out} = \cap_{i=1}^{m} \{x^L : \left(c^i\right)^T x^L + d^i \leq 0\}.$$

Solve the following problem efficiently,

$$\max_{x^{in} \in \mathcal{S}_{in}(x^{nom})} c^T x$$

A verification problem – Primal Problem

$$\max_{\substack{z^0,\ldots,z^{L-1}\\ x^0,\ldots,x^L}} c^T x^L + d$$

Upper bound of the constraints

$$\text{s.t } x^{l+1} = h^l\left(z^l\right), l = 0, 1, \ldots, L-1$$

Non-linear constraints

$$z^l = W^l x^l + b^l, l = 0, 1, \ldots, L-1$$

Linear layer models

$$x^0 = x^{in}, x^{in} \in \mathcal{S}_{in}\left(x^{nom}\right)$$

Inputs

[Dvijotham et al., 2018]

A verification problem – Dual Problem

$$\max_{\substack{z^0,\ldots,z^{L-1}\\ x^0,x^1,\ldots,x^{L-1}}} c^T\left(h^{L-1}\left(z^{L-1}\right)\right)+d$$

$$\text{s.t. } \underline{z}^l \le z^l \le \overline{z}^l \quad ,l=0,1,\ldots,L-1$$

$$\underline{x}^l \le x^l \le \overline{x}^l \quad ,l=0,1,\ldots,L-1$$

$$x^0 \in \mathcal{S}_{in}\left(x^{nom}\right)$$

$$+\sum_{l=0}^{L-1}\left(\mu^l\right)^T\left(z^l-W^lx^l-b^l\right)$$ Linear layer models (Lagrangian)

$$+\sum_{l=0}^{L-2}\left(\lambda^l\right)^T\left(x^{l+1}-h^l\left(z^l\right)\right)$$ Non-linear constraints (Lagrangian)

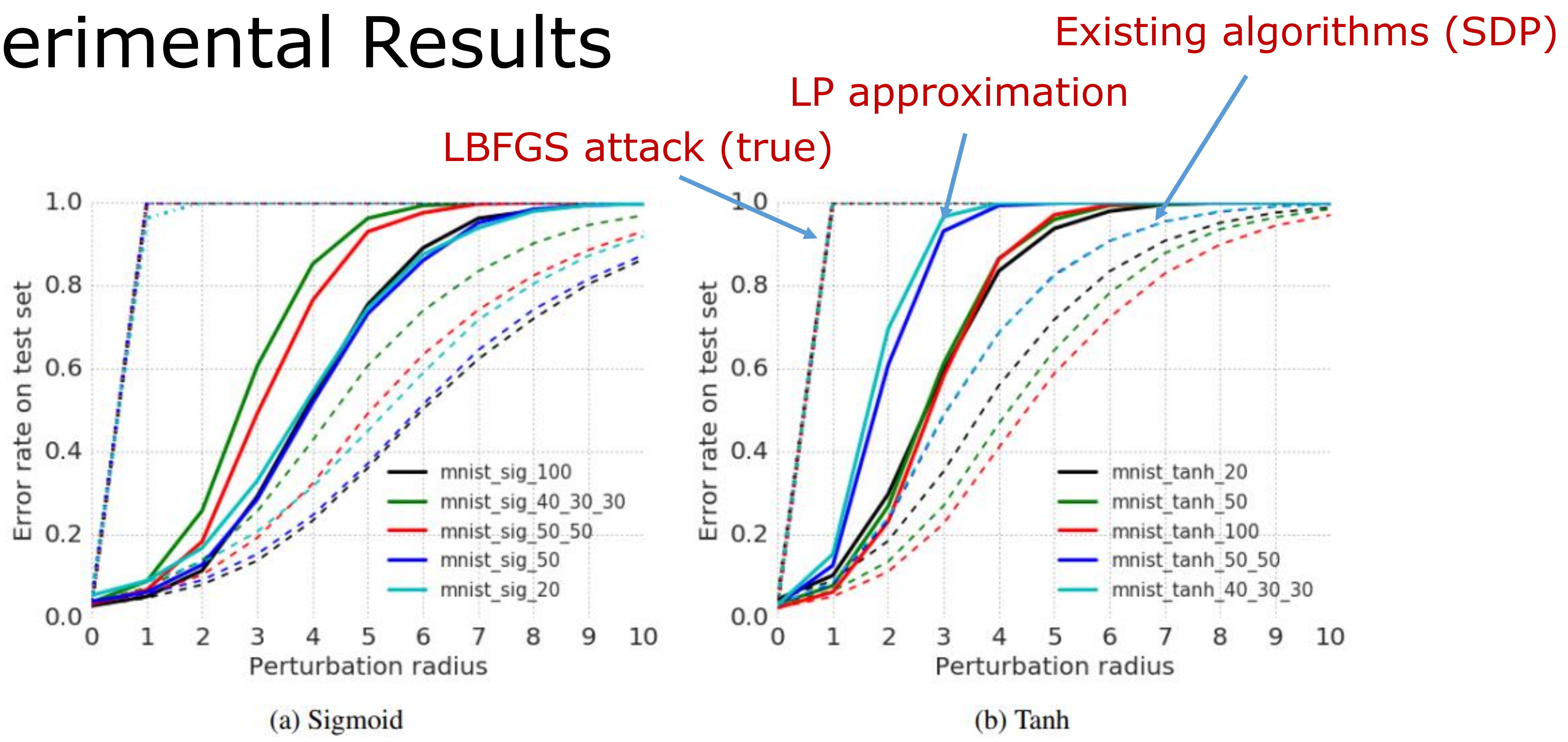
[Dvijotham et al., 2018]

A verification problem

Theorem: For any values of λ, μ, the objective of is an upper bound on the optimal value of the primal form. Hence, the optimal value of the dual form is also an upper bound. Further, the dual form is a convex optimization problem in (λ, μ).

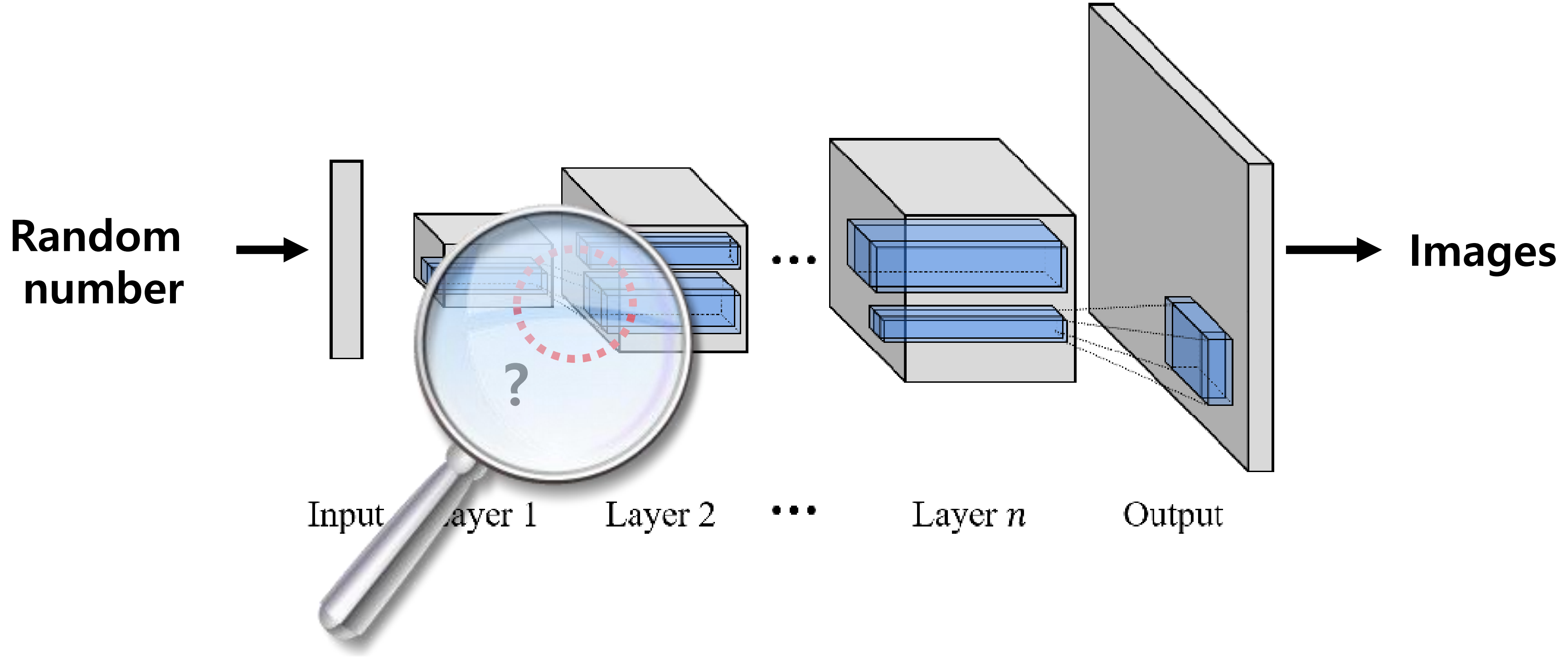
[Dvijotham et al., 2018]

Experimental Results



[Dvijotham et al., 2018]

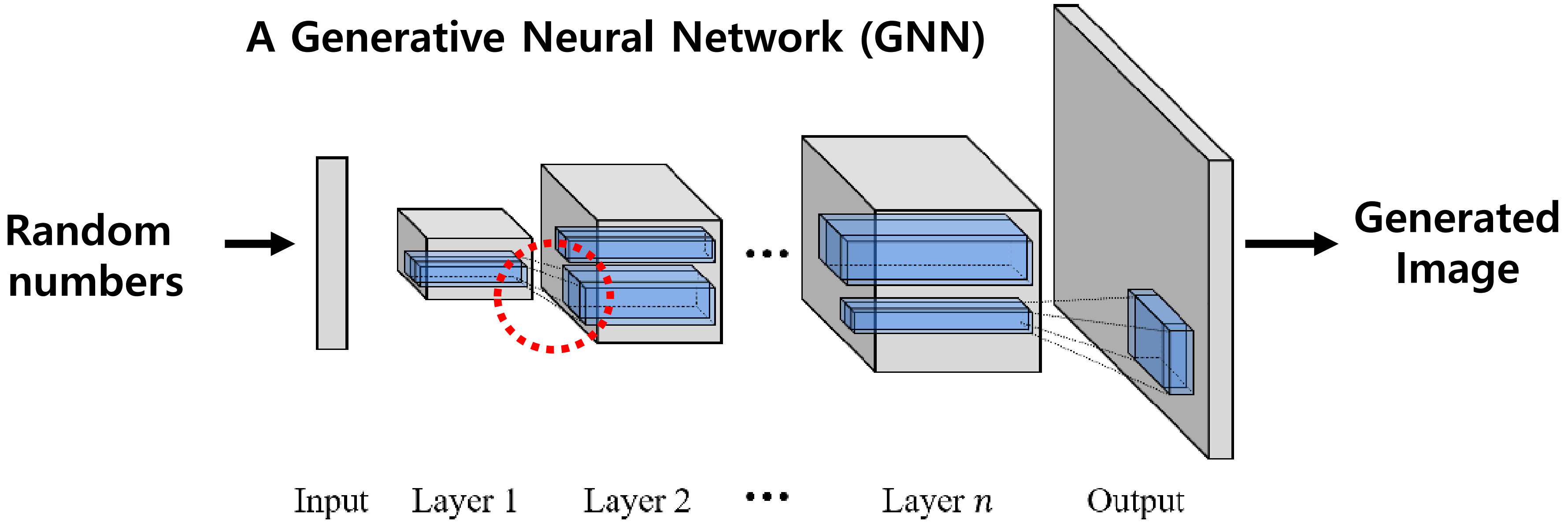
A Generative Neural Network (GNN)

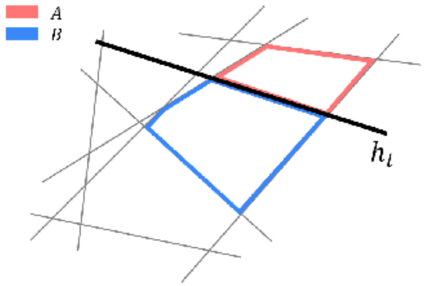


Generative Boundary Aware Sampling

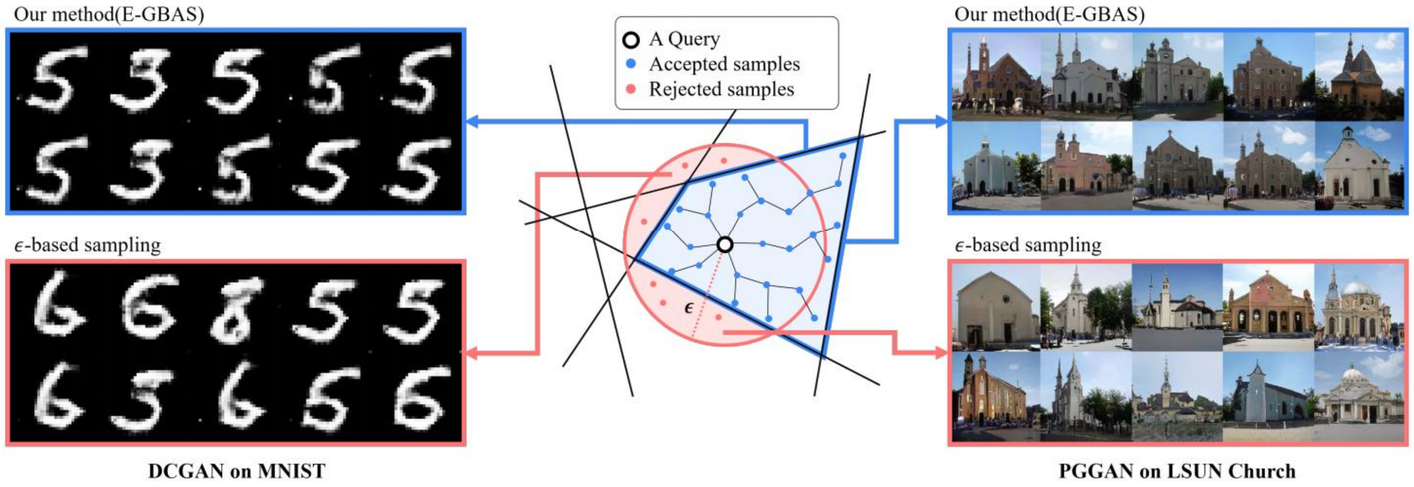
Giyoung Jeon et. al., 2020

A Generative Neural Network (GNN)





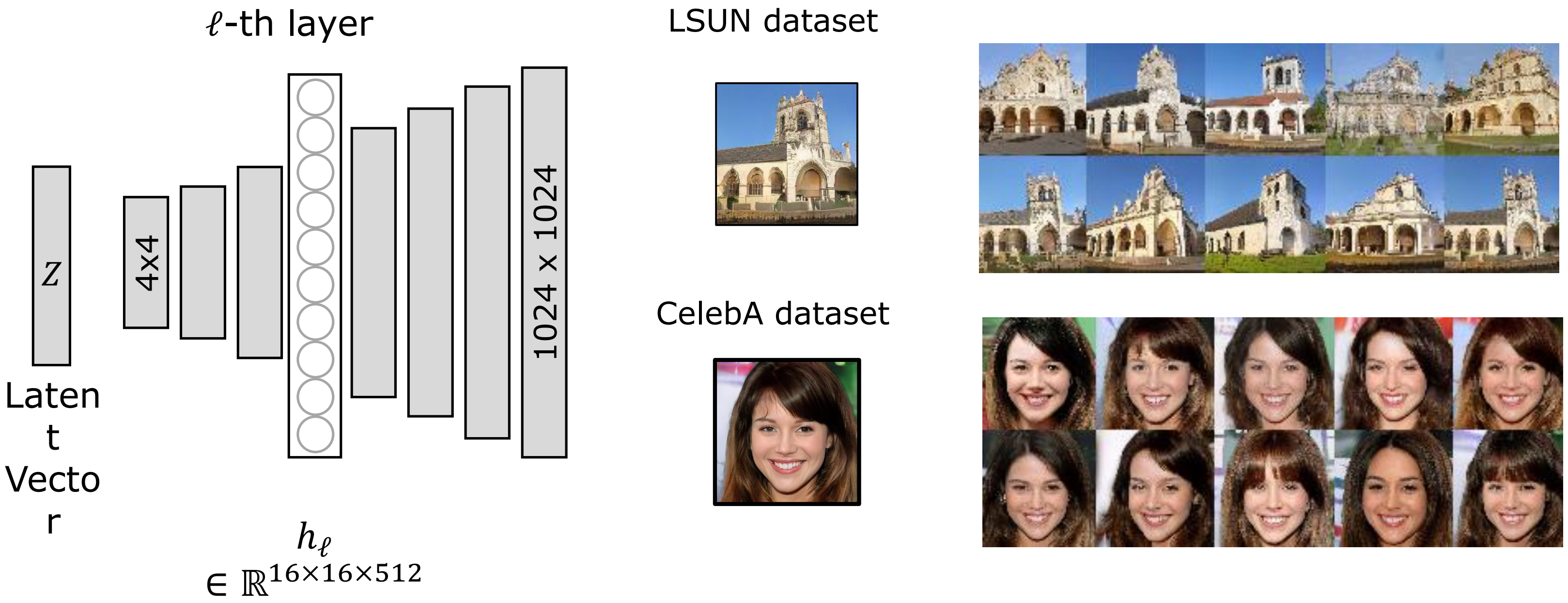
Generative boundaries



Generative Boundary Aware Sampling

Giyoung Jeon et. al., 2020

- The generative process is not well understood yet.
- We wish to give example-based explanation on the generative process.

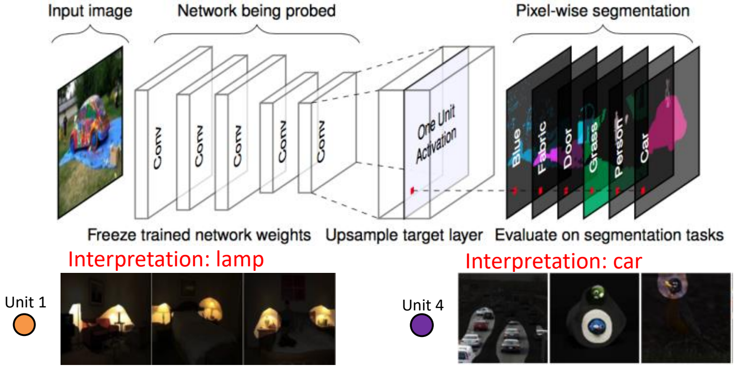


Generative Boundary Aware Sampling: Motivation

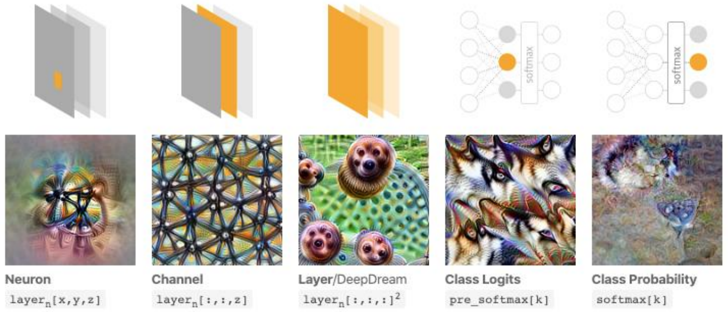
Giyoung Jeon et. al., 2020

Previous work: analyzing the inside of deep neural networks

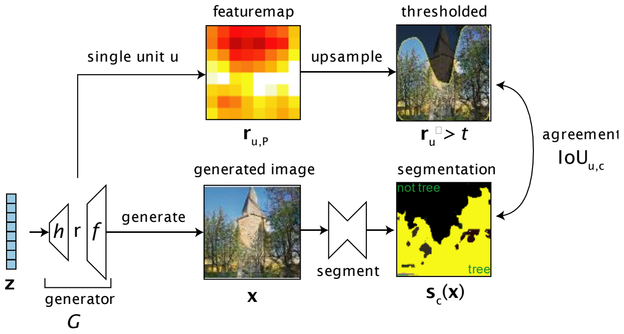
Network dissection (Bau et al., 2017)



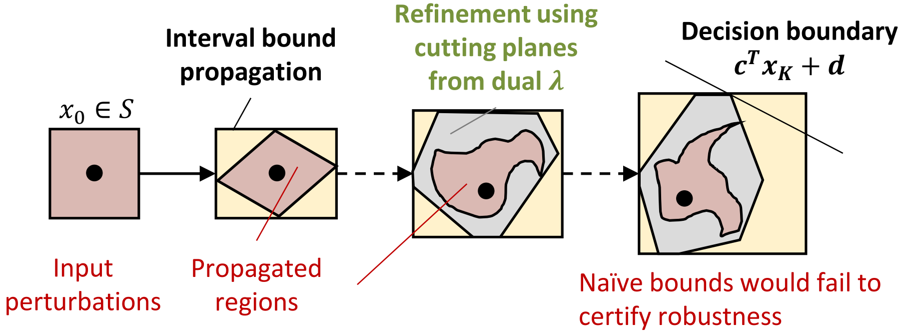
Google Deep Dream (Mordvintsev et al., 2015)



GAN dissection (Bau et al., 2019)



Lagrangian relaxed decision boundary (Dvijotham et al., 2018)

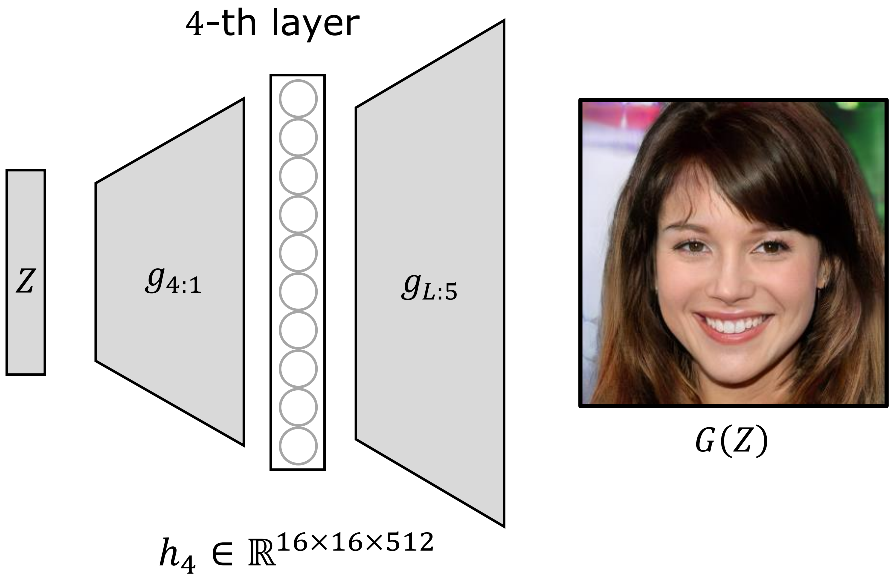


Generative Boundary Aware Sampling: Related Work

Giyoung Jeon et. al., 2020

Definitions

- Generator $G(Z)$: a generated image from Z
- Hidden nodes h_ℓ: a neural representation of ℓ-th layer
- Partial generation $g_{\mathrm{j:i}}: \mathbb{R}^{|\mathrm{h_i}|} \to \mathbb{R}^{|\mathrm{h_j}|}$: a generative function from layer i to layer j

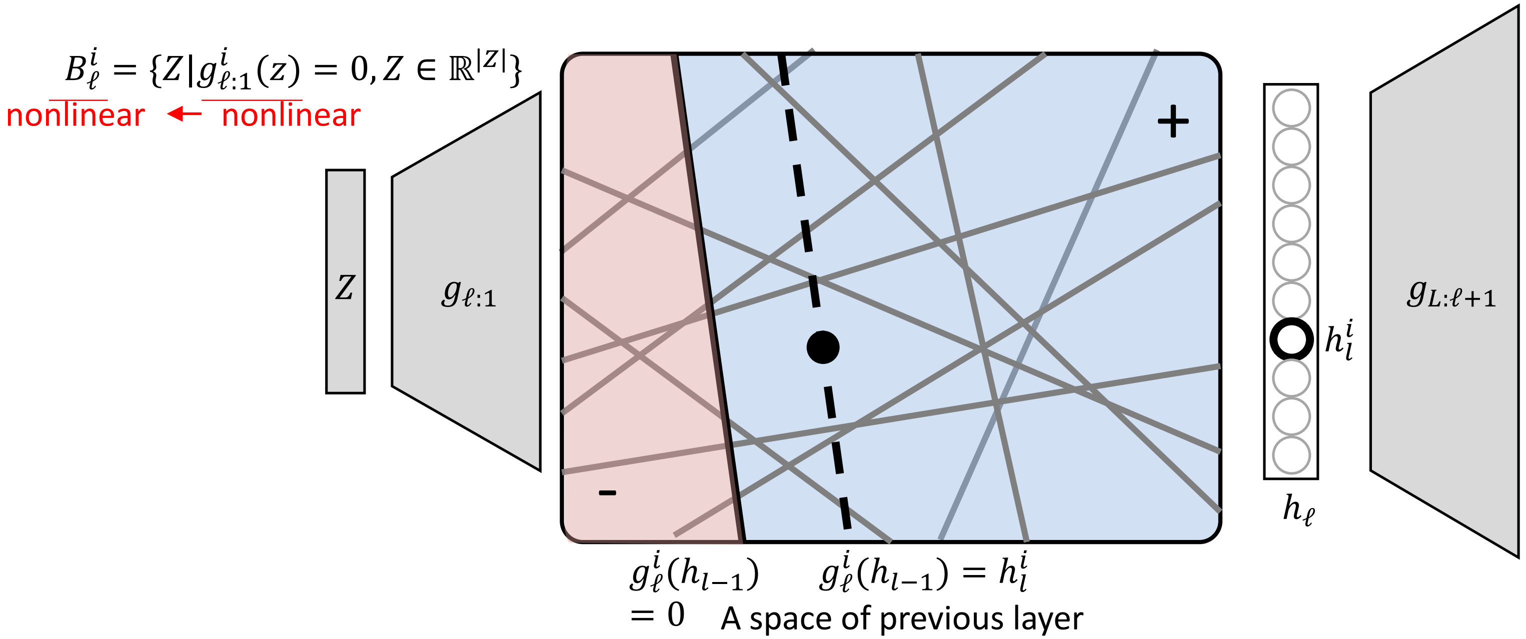


Generative Boundary Aware Sampling: Definitions

Giyoung Jeon et. al., 2020

Generative Boundary

- A value of h_ℓ is determined by the linear hyperplane in the space of the previous layer, $h_{\ell-1}$
- Stacking of layers toward input makes highly non-linear and non-convex shape
  - We want to see only feasible regions which constructed from the input to the target.
- Trained to fool the discriminator in GANs

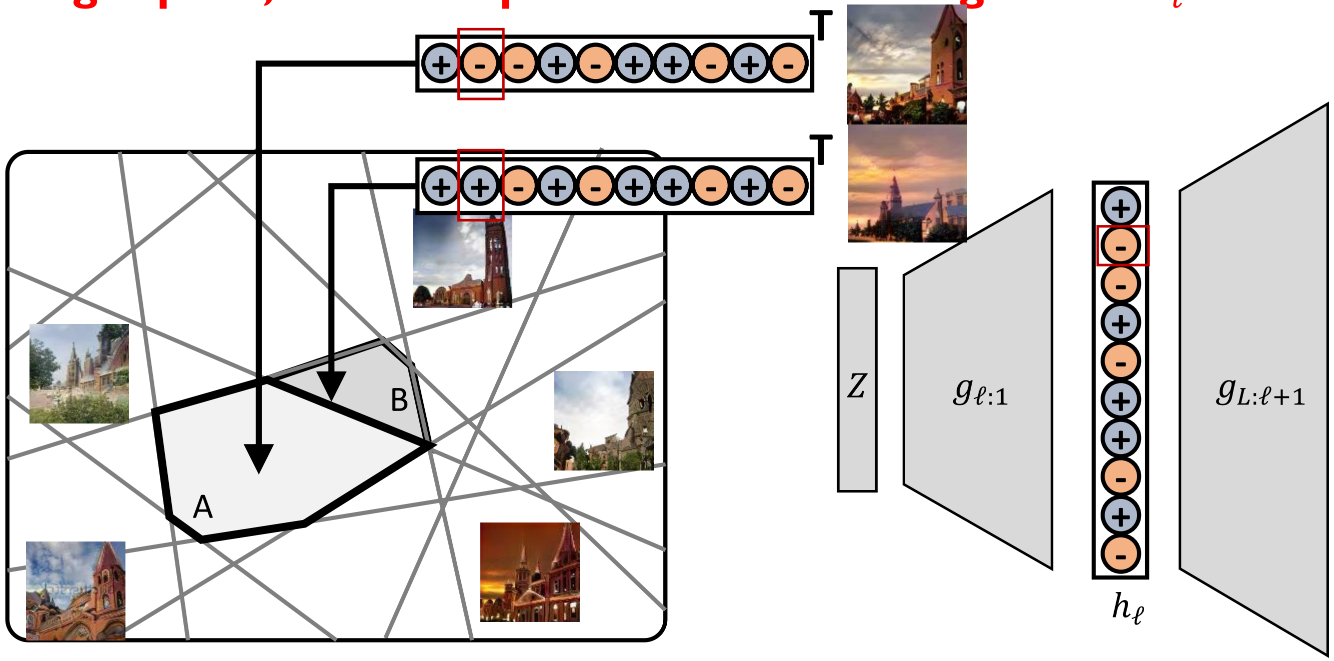


Generative Boundary Aware Sampling: Generative Boundary

Giyoung Jeon et. al., 2020

Generative Region

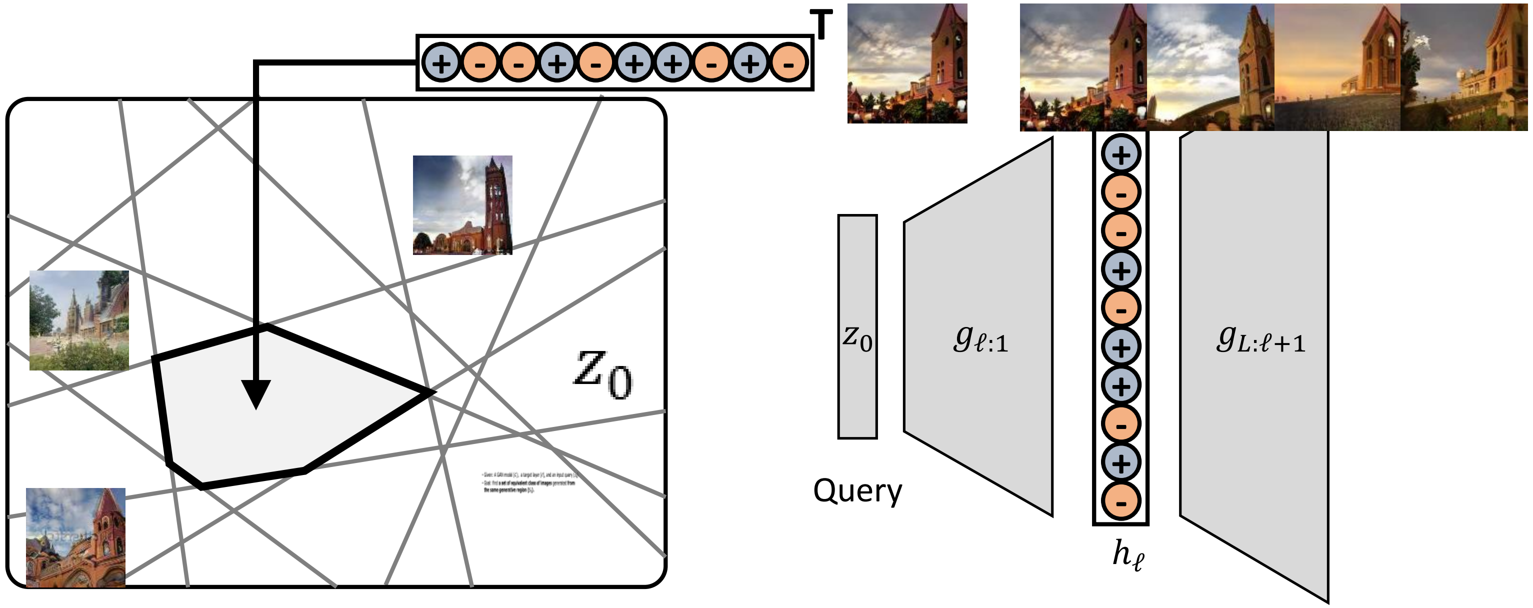
- In the ℓ-th layer, a space (S_ℓ) which is surrounded by a set of generative boundaries.
- **In the input space, a set of equivalent class of Z w.r.t S_ℓ.**
- **In the image space, a set of equivalent class of image w.r.t. S_ℓ.**



A space of previous layer

Problem Definition: Explorative sampling in a generative region

- Given: A GAN model (G), a target layer (ℓ), and an input query (z_0)
- Goal: find **a set of equivalent class of images** generated **from the same generative region (S_ℓ).**



A space of previous layer

Generative Boundary Aware Sampling: Problem Definition

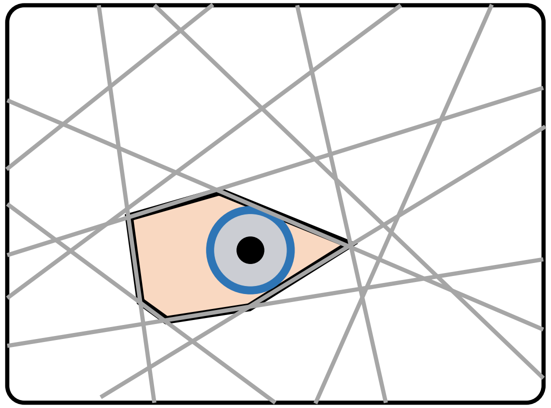
Giyoung Jeon et. al., 2020

Challenges of Sampling in a Generative Region

- The **dimension** of latent space and **a lot of hyperplanes** are hard to handle in practice. (E.g., 4th layer in PGGAN: $\mathbb{R}^{512} \rightarrow \mathbb{R}^{8192}$)
- Typically generative region is **nonconvex** in higher layer due to nonlinear activations.

Small ϵ-based sampling

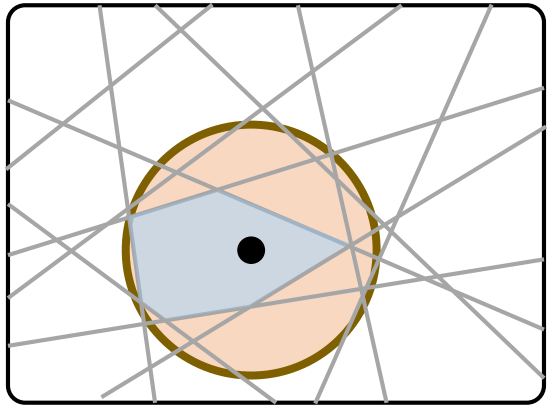
- Every samples inside the region
- Exists blind regions



Latent Space

Large ϵ-based spherical sampling

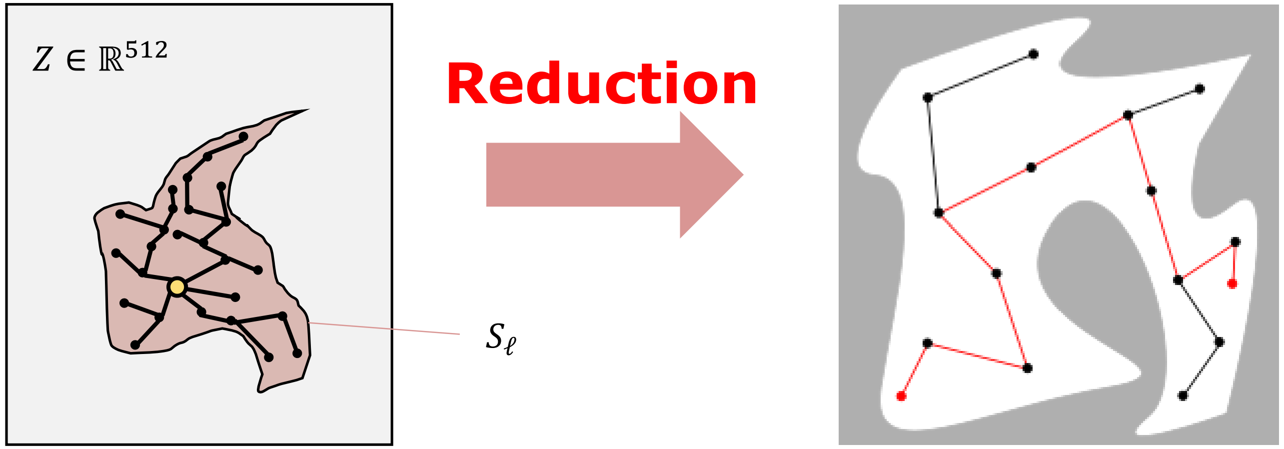
- Cover the region
- Might have out-of-region samples



Latent Space

Reduction to the Robot Planning Problem

Exploring a Generative Region Problem | **Robot Planning Problem**



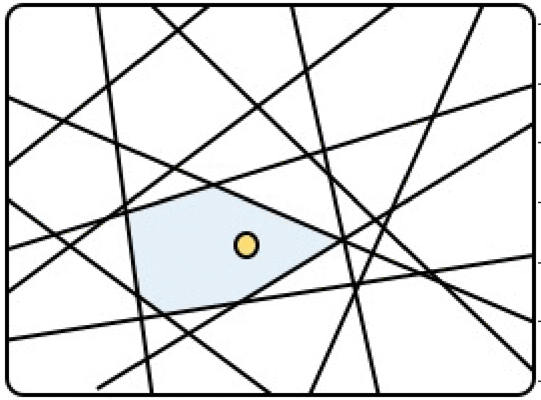
- Searching samples in non-convex space
- High dimensional explorative space

- Searching a path in non-convex space
- High degree of freedom of robot joint

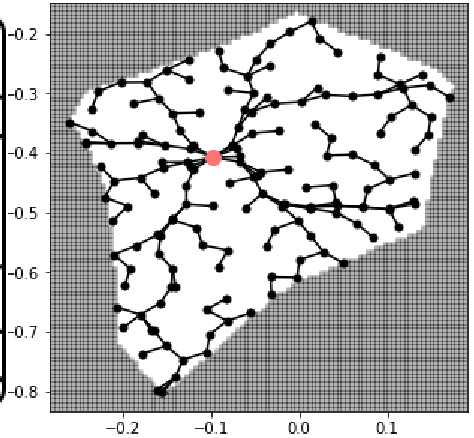
We reduce our sampling problem into robot-planning problem.

Generative Boundary constrained Rapidly-exploring Random Tree (RRT)

- Given generative boundary as constraints,
  RRT is gives solution to search over the generative region.
- This explorative sampling always guarantee acceptance inside the region



Illustrative example



Example in nonconvex region

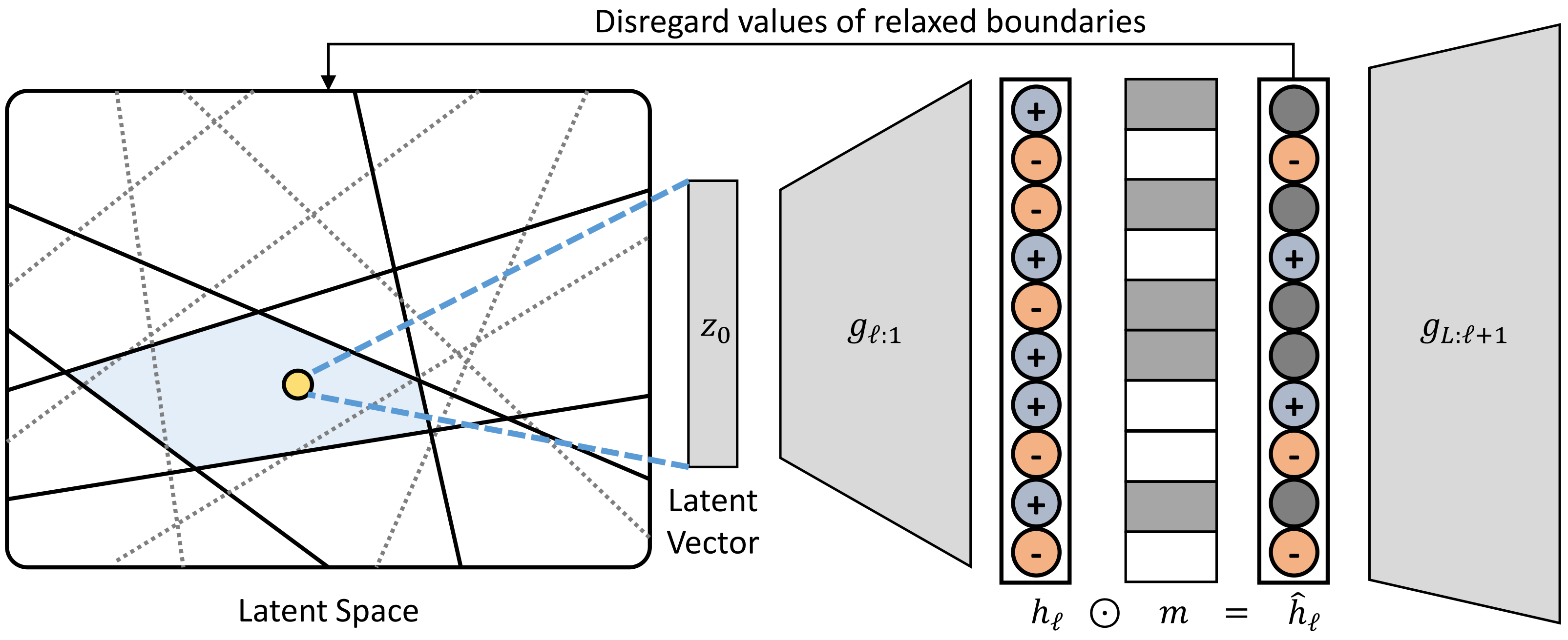
LaValle, Steven M. “Rapidly-exploring random trees: A new tool for path planning”. *Technical Report. Computer Science Department, Iowa State University.* 1998.

Generative Boundary Aware Sampling: Solution I

Giyoung Jeon et. al., 2020

Smallest Supporting Generative Boundary Set

- Using all the boundaries, constraints get **too tight** and **computationally expensive**.
- We observe not all the boundaries affects equally on the output.



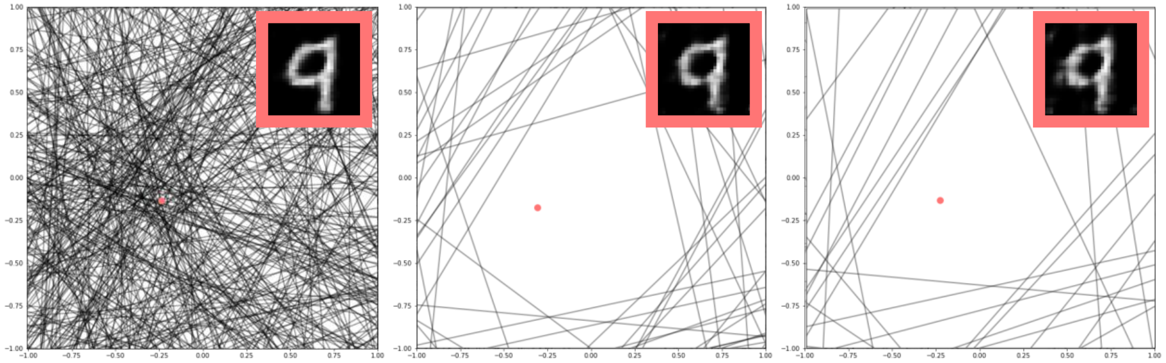
Generative Boundary Aware Sampling: Solution II

Giyoung Jeon et. al., 2020

Smallest Supporting Generative Boundary Set

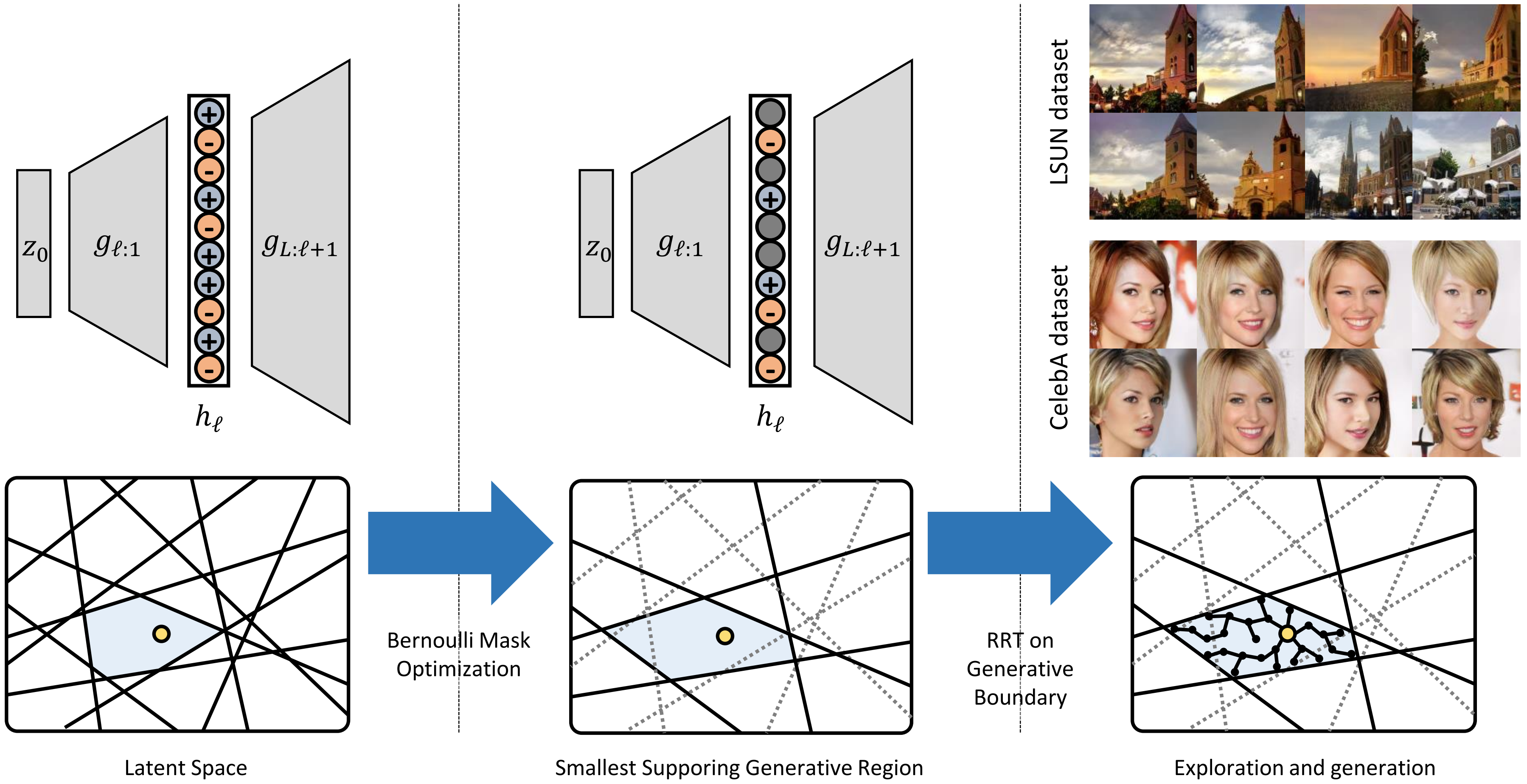
- Apply **Bernoulli mask optimization** to relax boundaries but maintain the output.

$$\begin{aligned}\theta^* &= \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(z_0, \ell, \theta) \\ &= \underset{\theta}{\operatorname{argmin}} \underbrace{\| g_{L:\ell+1}(g_{l:1}(z_0) \odot m) - G(z_0) \|}_{\text{Masked image reconstruction error}} + \underbrace{\lambda \|\theta\|_1}_{\text{Mask l1 regularizer}} \qquad \text{where } m \sim Ber(\theta)\end{aligned}$$



Entire boundaries | Using 10% | Using 5%

Chang, Chun-Hao, et al. "Explaining image classifiers by adaptive dropout and generative in-filling." *International Conference on Learning Representations (ICLR)*. 2018.



Generative Boundary Aware Sampling: Solution II

Giyoung Jeon et. al., 2020

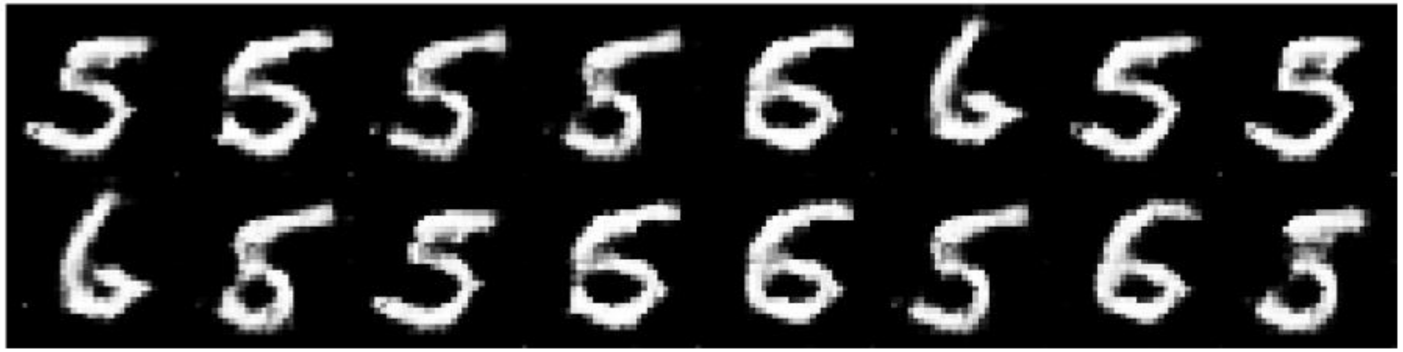
Explorative Generative Boundary Aware Sampling

- Accepted Cluster 1
- Accepted Cluster 2
- Accepted Cluster 3
- Rejected Sample

Experiment : DCGAN-MNIST

Query

ϵ-based sampling



E-GBAS

Generative Boundary Aware Sampling: Results

Giyoung Jeon et. al., 2020

Experiment : PGGAN-LSUN-church

Query

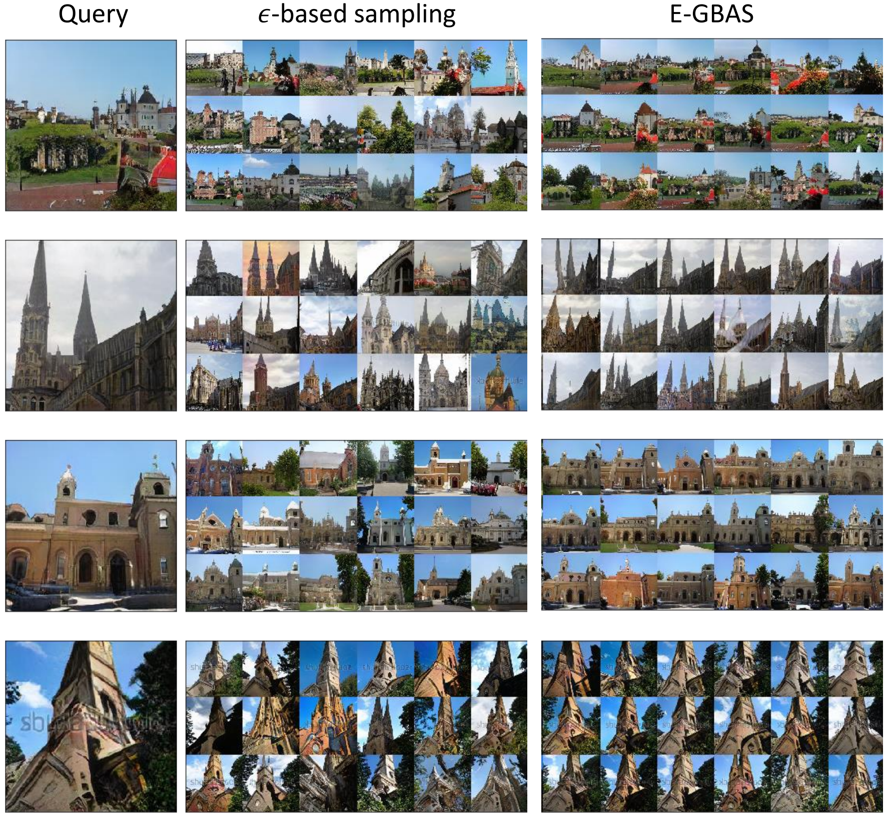
ϵ-based sampling

E-GBAS

Generative Boundary Aware Sampling: Results

Giyoung Jeon et. al., 2020

Experiment : PGGAN-LSUN-church



Generative Boundary Aware Sampling: Results

Giyoung Jeon et. al., 2020

Experiment : PGGAN-celebA

Query

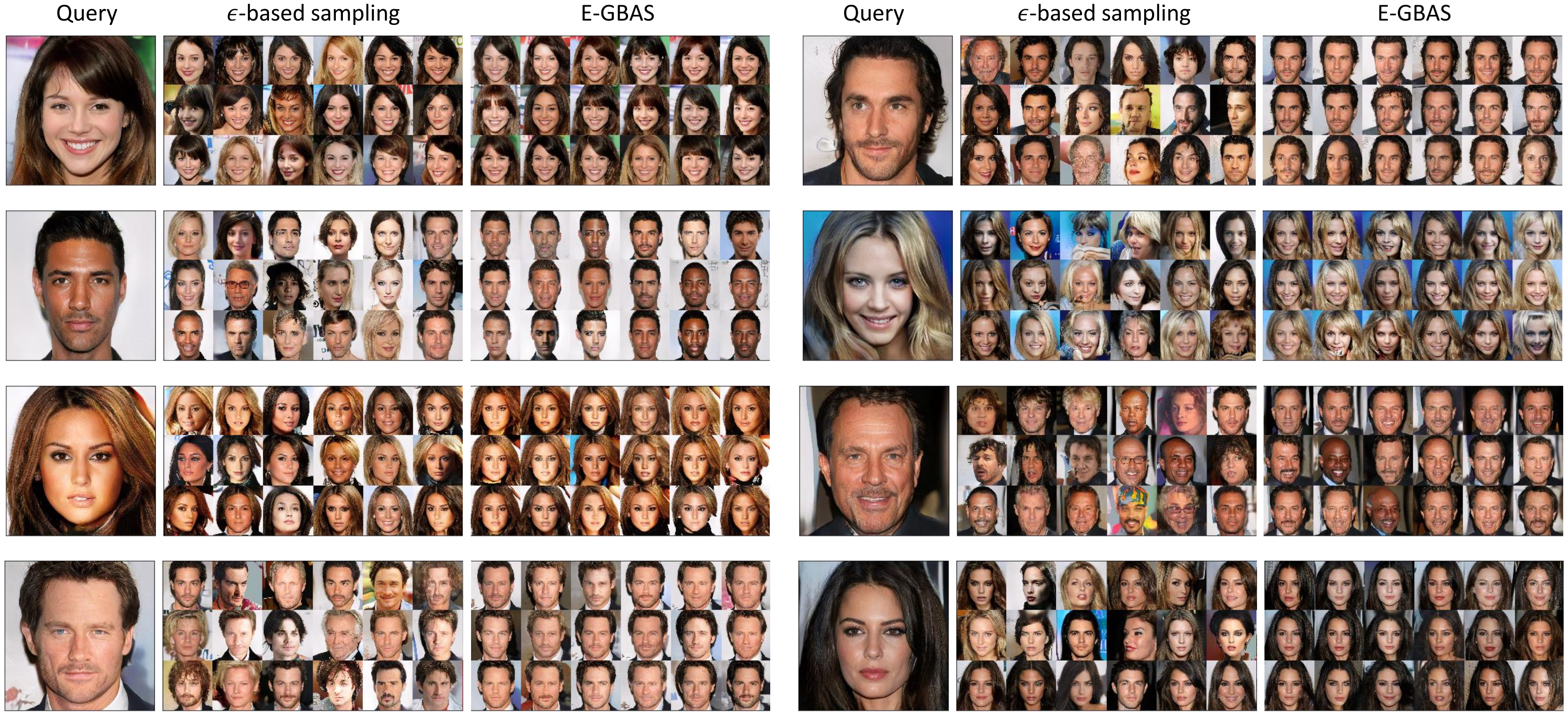
ϵ-based sampling

E-GBAS

Generative Boundary Aware Sampling: Results

Giyoung Jeon et. al., 2020

Experiment : PGGAN-celebA



Generative Boundary Aware Sampling: Results

Giyoung Jeon et. al., 2020

Experiment : According to the portion of activate mask

Generative Boundary Aware Sampling: Results

Giyoung Jeon et. al., 2020

- There are recent advances to analyze internal mechanisms of deep neural networks.

- Some deep neural networks models such as semantic segmentation and generative models make us to analyze internal nodes better.

- Thus, it would be possible to validate the correctness of individual decision/generative boundaries.

Conclusions of Part II

References

1. [Network Dissection] Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6541-6549).
2. [GAN Dissection] Bau, D., Zhu, J. Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. arXiv preprint arXiv:1811.10597.
3. [E-GBAS] Jeon, G., Jeong, H., Choi, J. (2020). An Efficient Explorative Sampling Considering the Generative Boundaries of Deep Generative Neural Networks. In Thirty-Third AAAI Conference on Artificial Intelligence.
4. Dvijotham, K., Stanforth, R., Gowal, S., Mann, T., Kohli, P (2018). A Dual Approach to Scalable Verification of Deep Networks, Uncertainty in Artificial Intelligence.
5. Mordvintsev, Alexander; Olah, Christopher; Tyka, Mike (2015). "DeepDream - a code example for visualizing Neural Networks". Google Research. Archived from the original on 2015-07-08.
6. Kumar, M. P. (2019) Neural Network Verification, VMCAI Winter School.