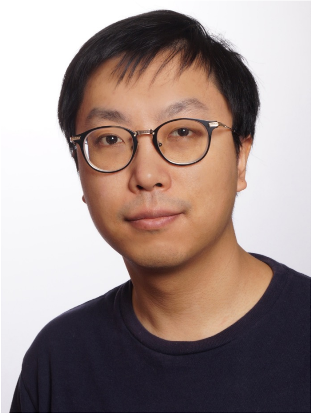
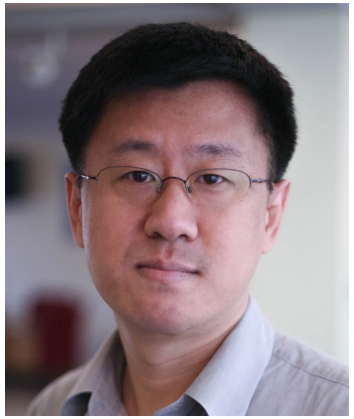
Precise and Generalized Robustness Certification for Neural Networks

Yuanyuan Yuan, Shuai Wang, Zhendong Su

HKUST, ETH Zurich





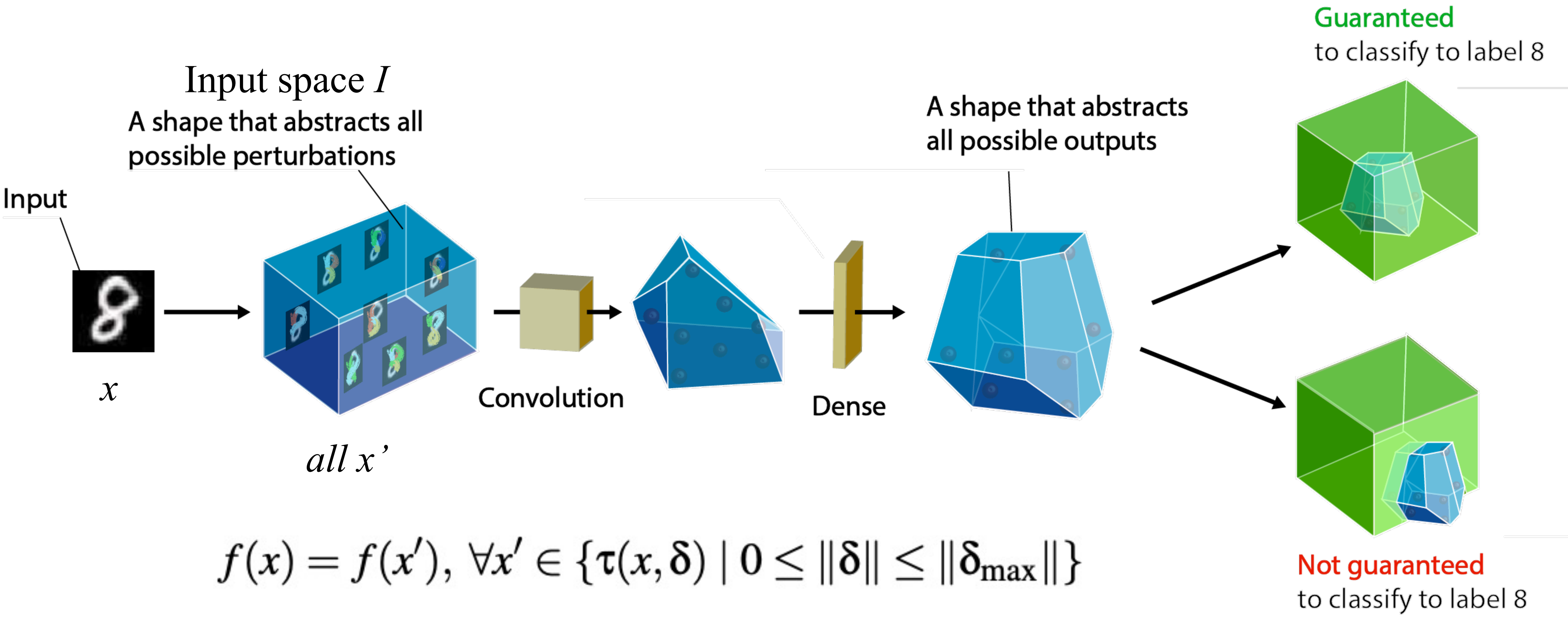
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

ETH zürich

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Robustness Certification

$$\forall x', f(x) = f(x')$$



$$f(x) = f(x'),\ \forall x' \in \{\tau(x,\delta) \mid 0 \leq \|\delta\| \leq \|\delta_{max}\|\}$$

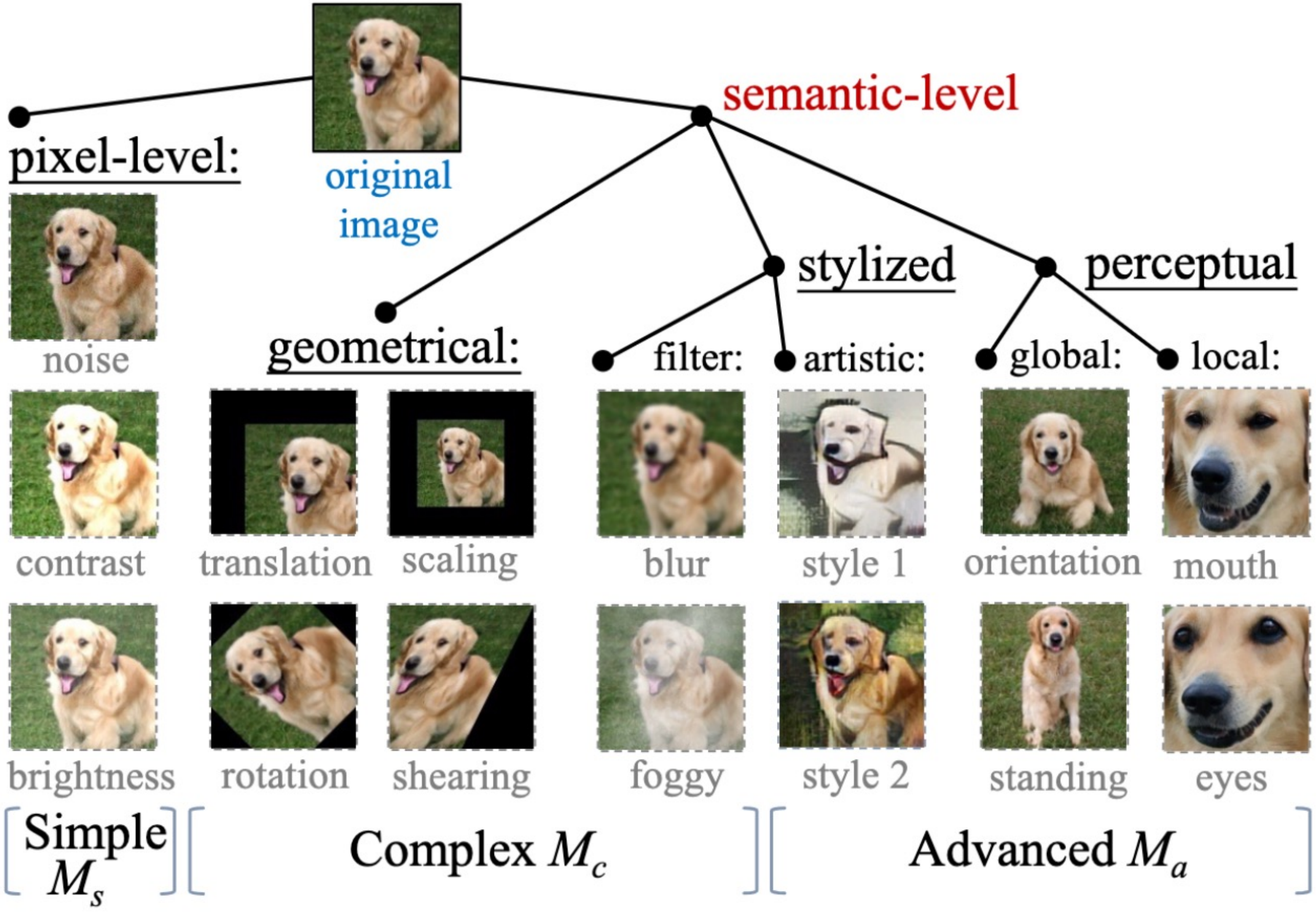
**Figures are from the ERAN project: https://github.com/eth-sri/eran*

Robustness Certification

	$f(x) = f(x'), \forall x' \in \{\tau(x,\delta) \mid 0 \leq \|\delta\| \leq \|\delta_{\max}\|\}$	$\forall x', f(x) = f(x')$
Sound	✓	✓

	$f(x) = f(x'), \forall x' \in \{\tau(x,\delta) \mid 0 \leq \|\delta\| \leq \|\delta_{\max}\|\}$	$\exists x', f(x) \neq f(x')$
Incomplete	✗	?
Complete	✗	✓

Input Mutation



Diverse input mutations

Input Mutation

Simple mutations:
- Explicit math forms
- Linear I

Can directly get precise input space representation

pixel-level

The focus of previous works

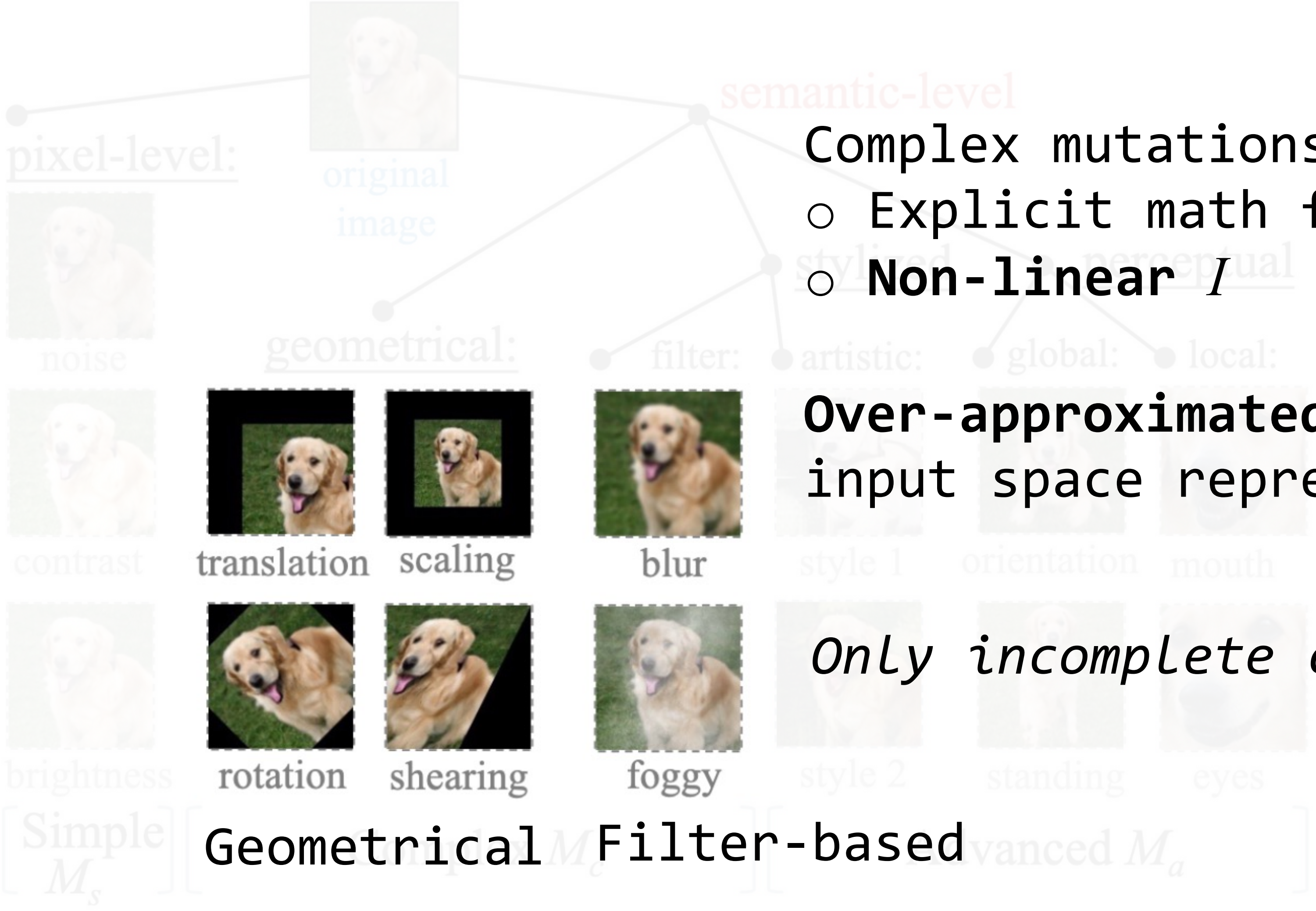
Sound & complete certification

Input Mutation

Complex mutations:
- Explicit math forms
- **Non-linear** I

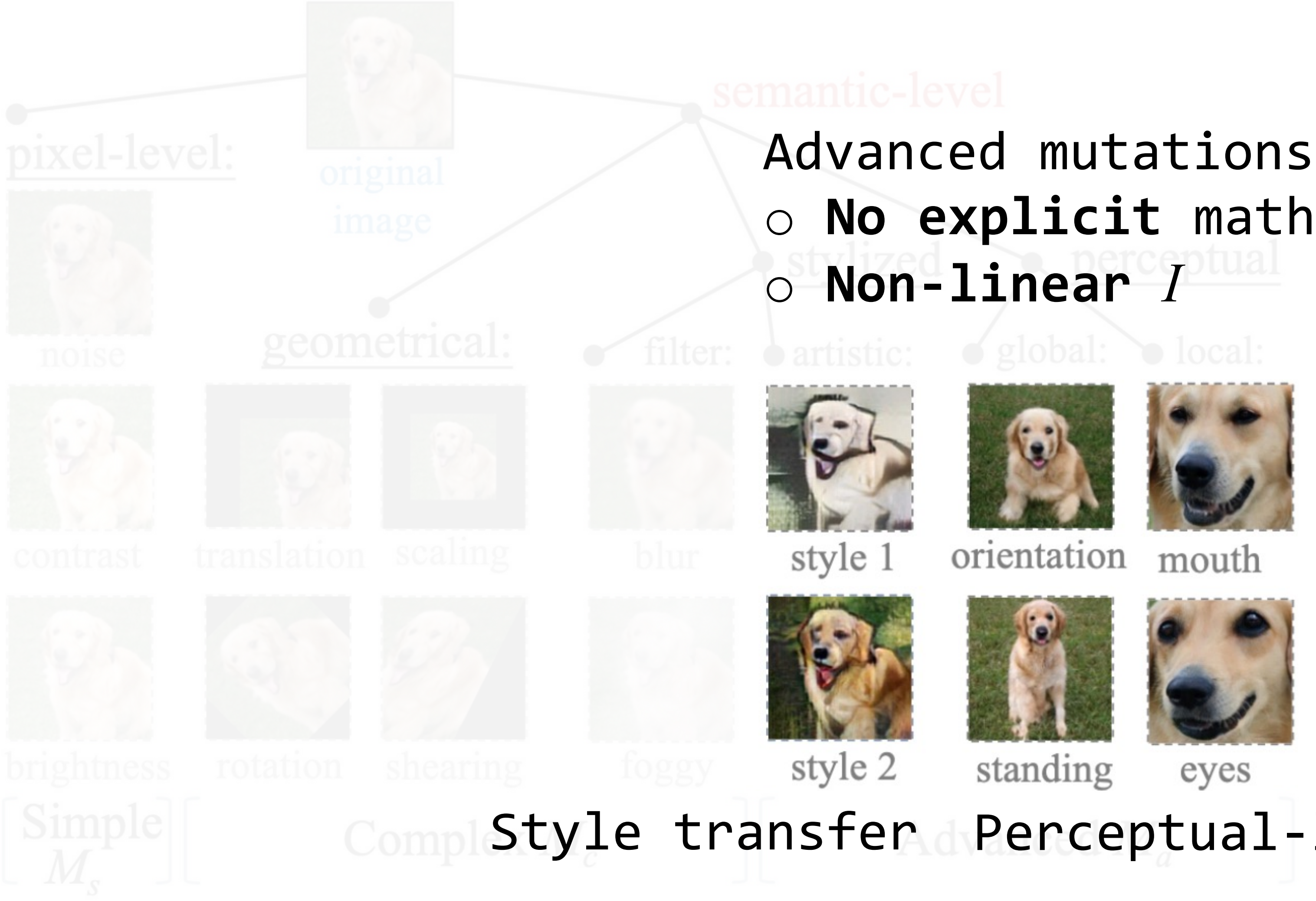
Over-approximated
input space representation

Only incomplete certification



Geometrical Filter-based

Input Mutation



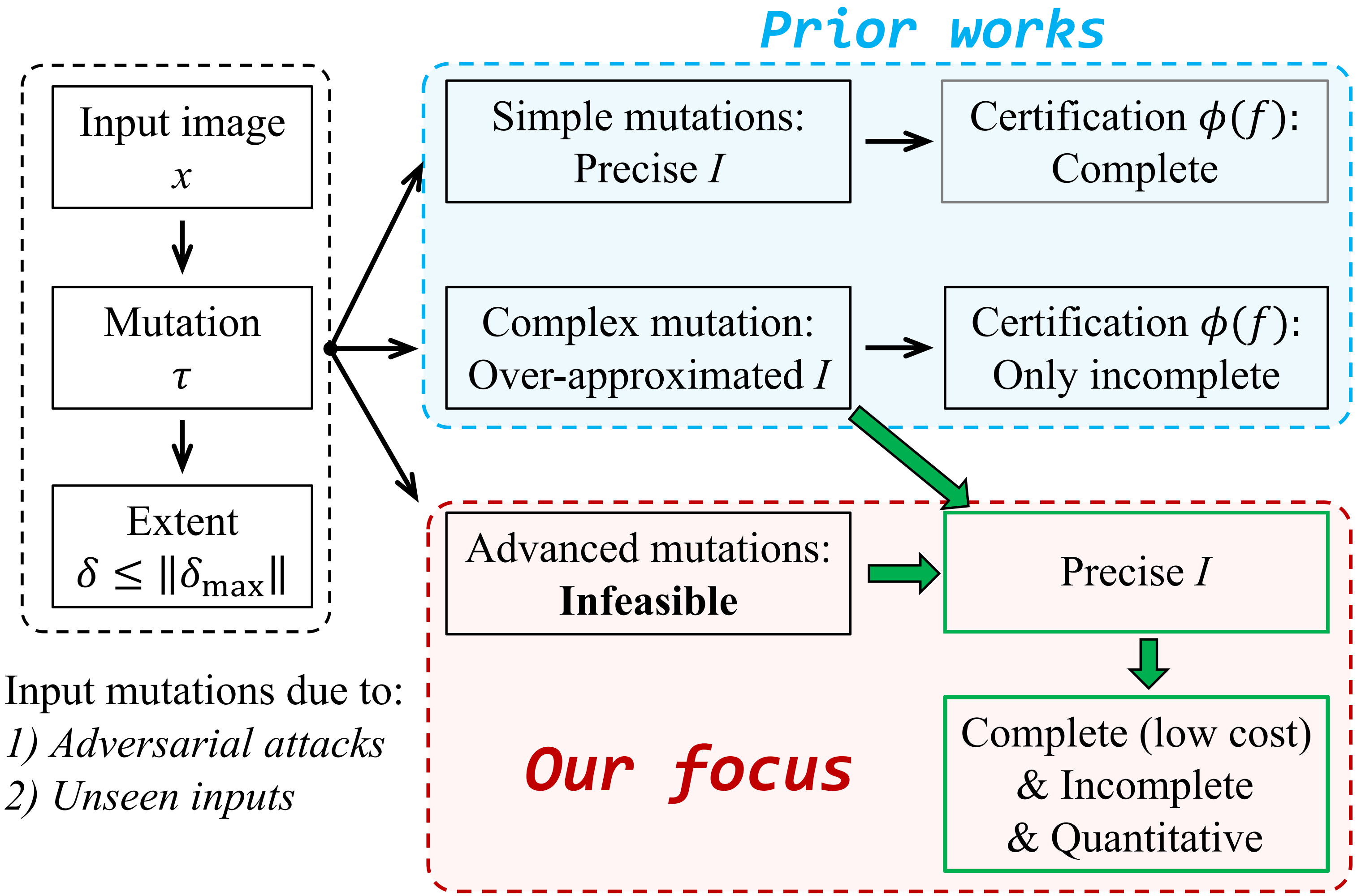
Advanced mutations:

- **No explicit** math form
- **Non-linear** I

Style transfer Perceptual-level

Never studied!

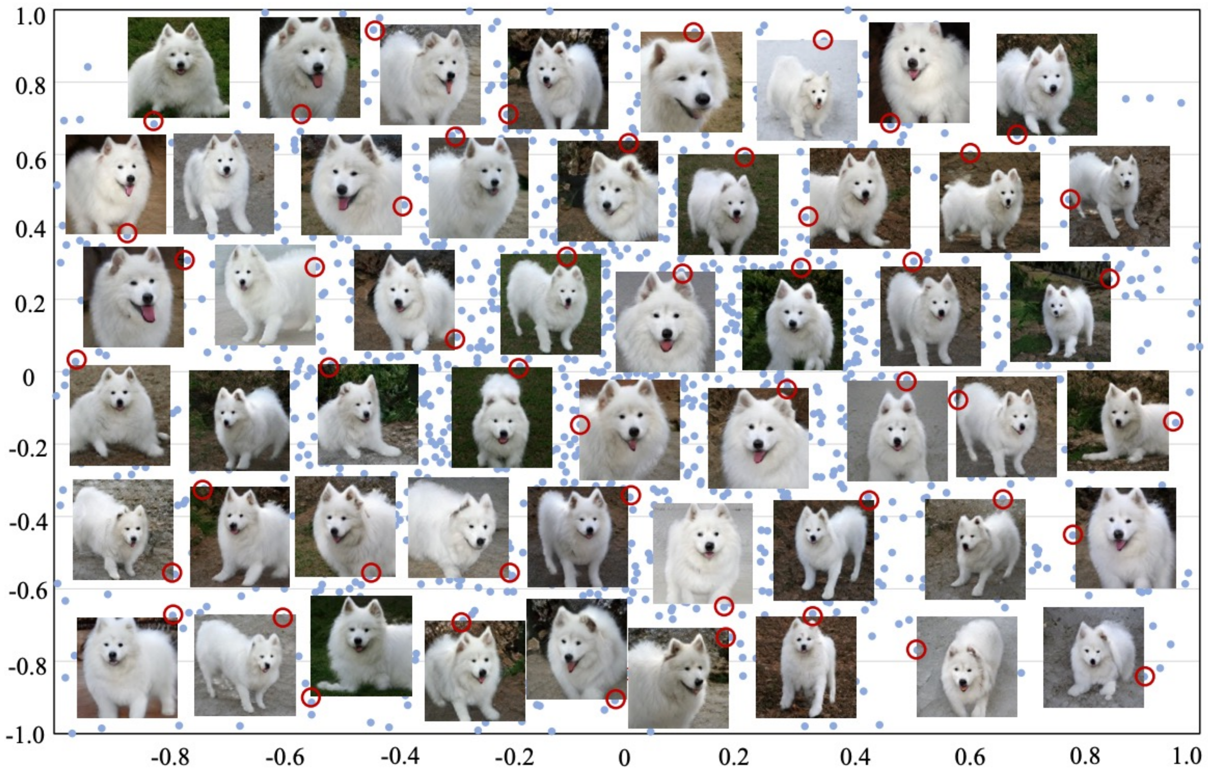
Overview



Precise:
- Deliver precise I

Generalized:
- Support advanced mutation
- Unified implementation
- Support conventional certification frameworks (complete/quantitative)

Motivation: Generative Model

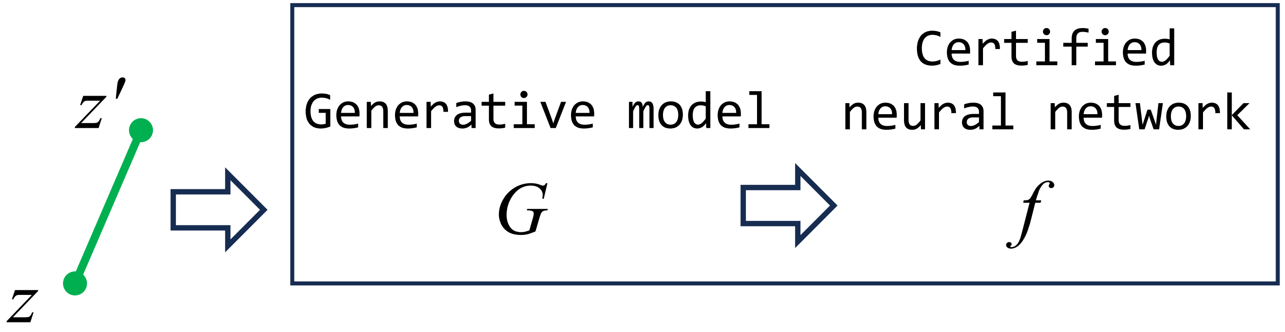


A collection of images → Latent space → **Infinite** images by (inter)extrapolation

Data-driven mutations:

1) *Extract mutations from diverse images*
2) *Represent mutations as moving directions in latent space*

Motivation and Problems

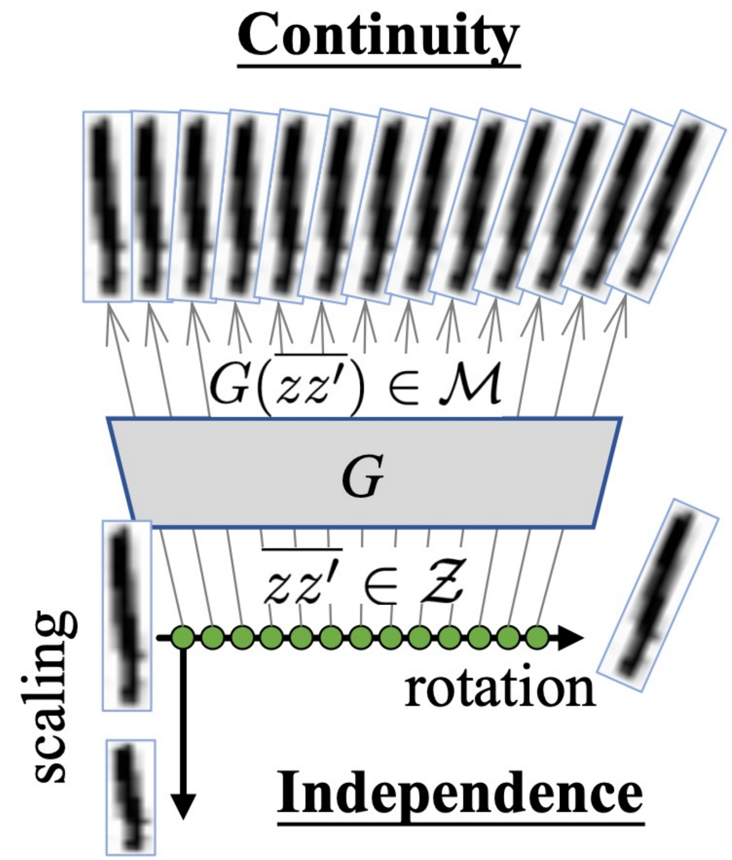


Certify $f \circ G$

$G(z)$: original input
$G(z')$: maximumly mutated inputs
$\overline{zz'}$: corresponds to all mutated inputs
$z \rightarrow z'$: mutating direction

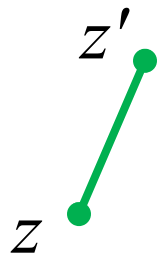
The problem: $G(z)$ *changes arbitrarily with* z*!*

Two Requirements



Continuity: when performing mutations, $G(z)$ changes continuously with z.

Independency: when mutating $G(z)$ into $G(z')$, $z \rightarrow z'$ should only correspond to the expected mutation.



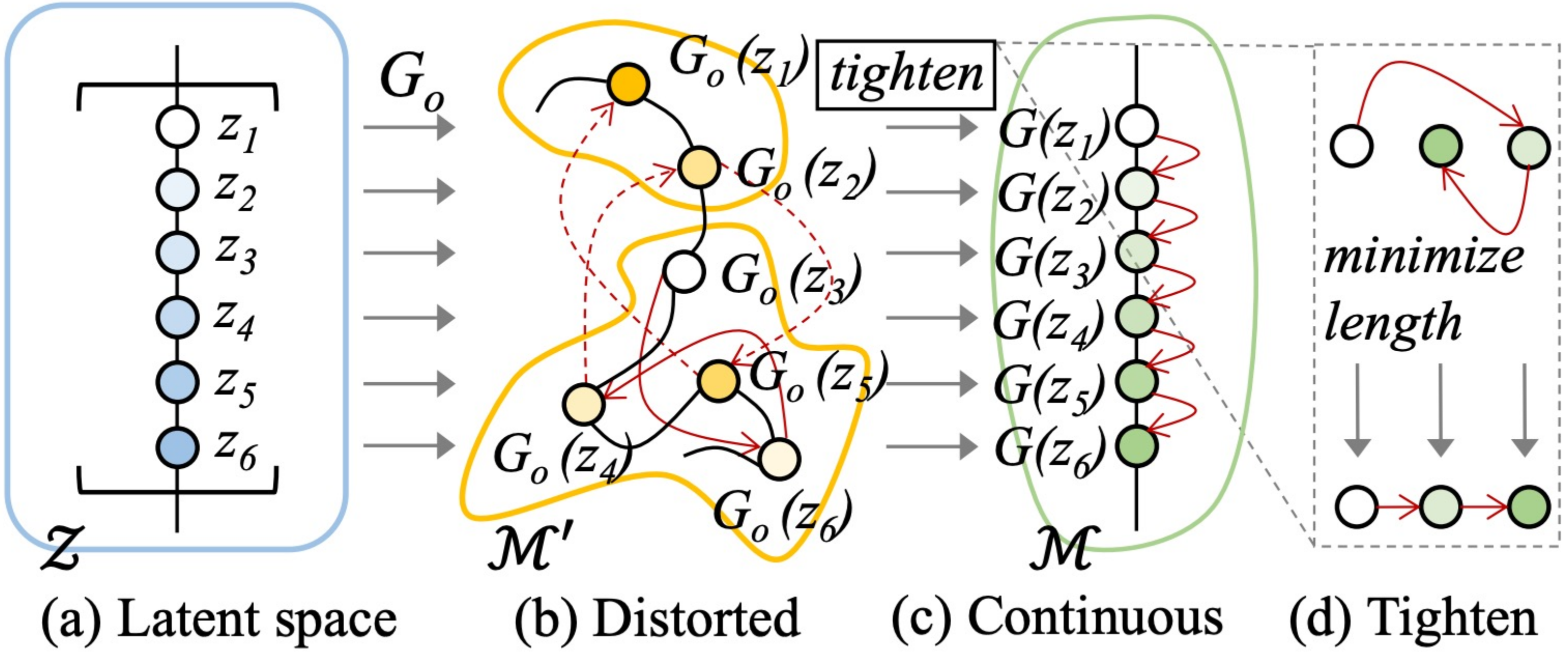
$\overline{zz'}$ will exclusively correspond to all mutated inputs between $G(z)$ and $G(z')$.

Continuity

$$\forall z, z' : \frac{1}{C} d_1(z, z') \leq d_2(G(z), G(z')) \leq C d_1(z, z')$$

d_1: distance metric over z

d_2: distance metric over $G(z)$



Bound the Jacobian norm of G!

Independency

When extracting mutations, different mutations are represented as **orthogonal** directions.

When performing local mutations, projecting the mutating direction into the **non-mutating direction** of the remaining region.

Evaluation: Mutations

Findings:

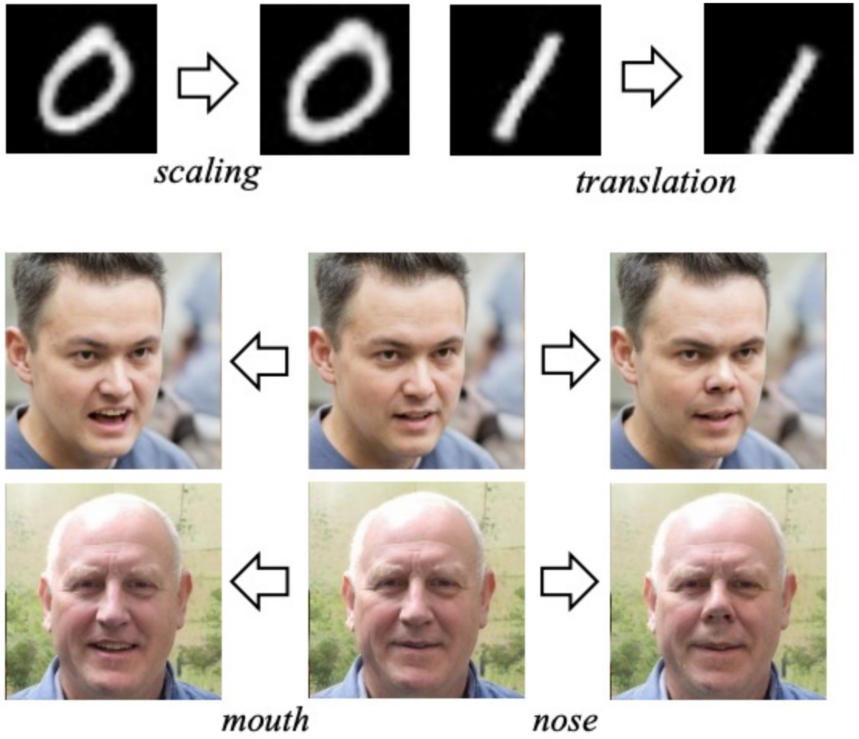
The resolution of G's training data affects the number of enabled (perceptual) mutations.

- *Use higher resolution training data for the generative model.*

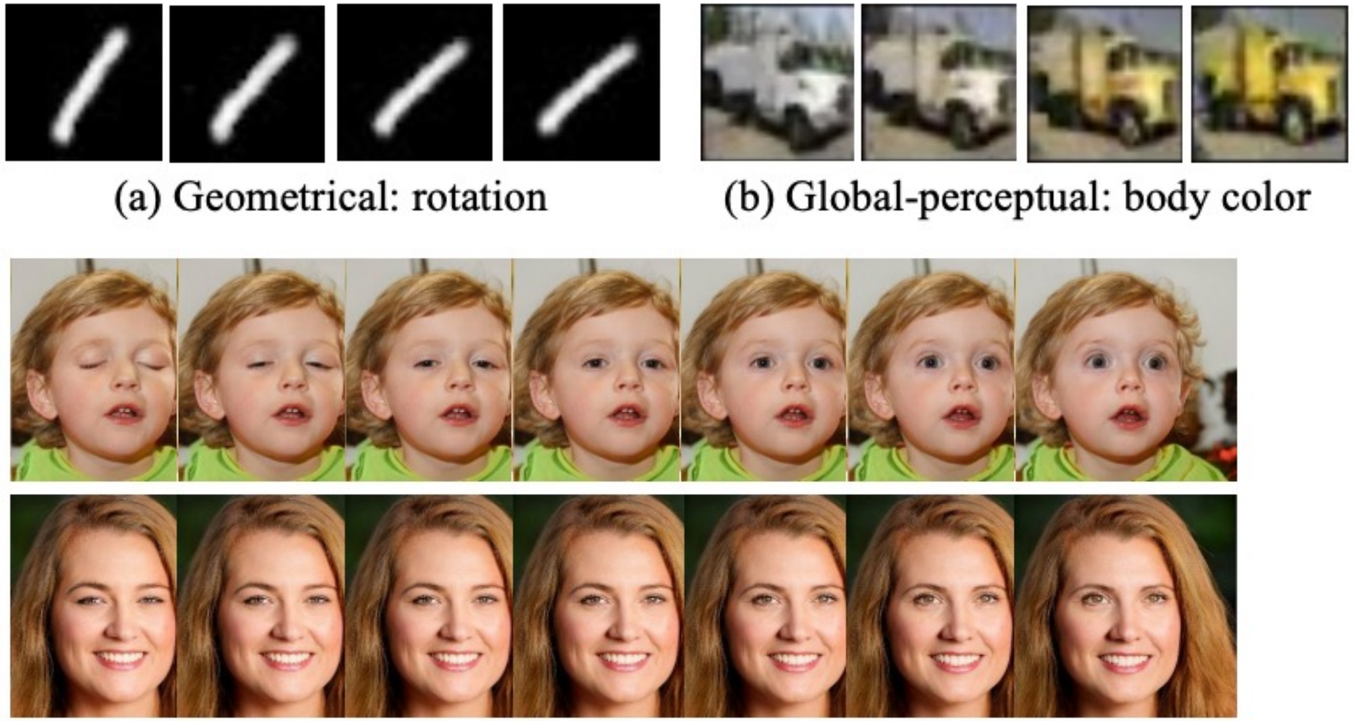
Training data decide the enabled mutations and the maximal extent of mutations.

- *E.g., To enable rotation* $30°$, *augment the training data by rotating them* $30°$. *But it's unnecessary to cover all* $[0, 30°]$ *to enable all rotation within* $[0, 30°]$ *due to continuity.*

Evaluation: Mutations



Independency



(c) Local-perceptual: opening eyes

Continuity

Evaluation: Certification

Complete certification over geometrical mutations

Cost: $O((2^N)^L) \longrightarrow O((N^2)^L)$ Input to $f \circ G$ *is a segment*

N: #maximal neurons in one layer
L: #layers

Findings on different neural networks:

Conv vs. FC: convolution layer can enhance the robustness
Depth: deeper neural network has better robustness
Data augmentation: can also enhance the robustness

Evaluation: Certification

Quantitative certification over perceptual mutations

1) Quantifies the robustness with lower/upper bounds
2) Requires inputs are represented via segments

Quantitative certification for face recognition.

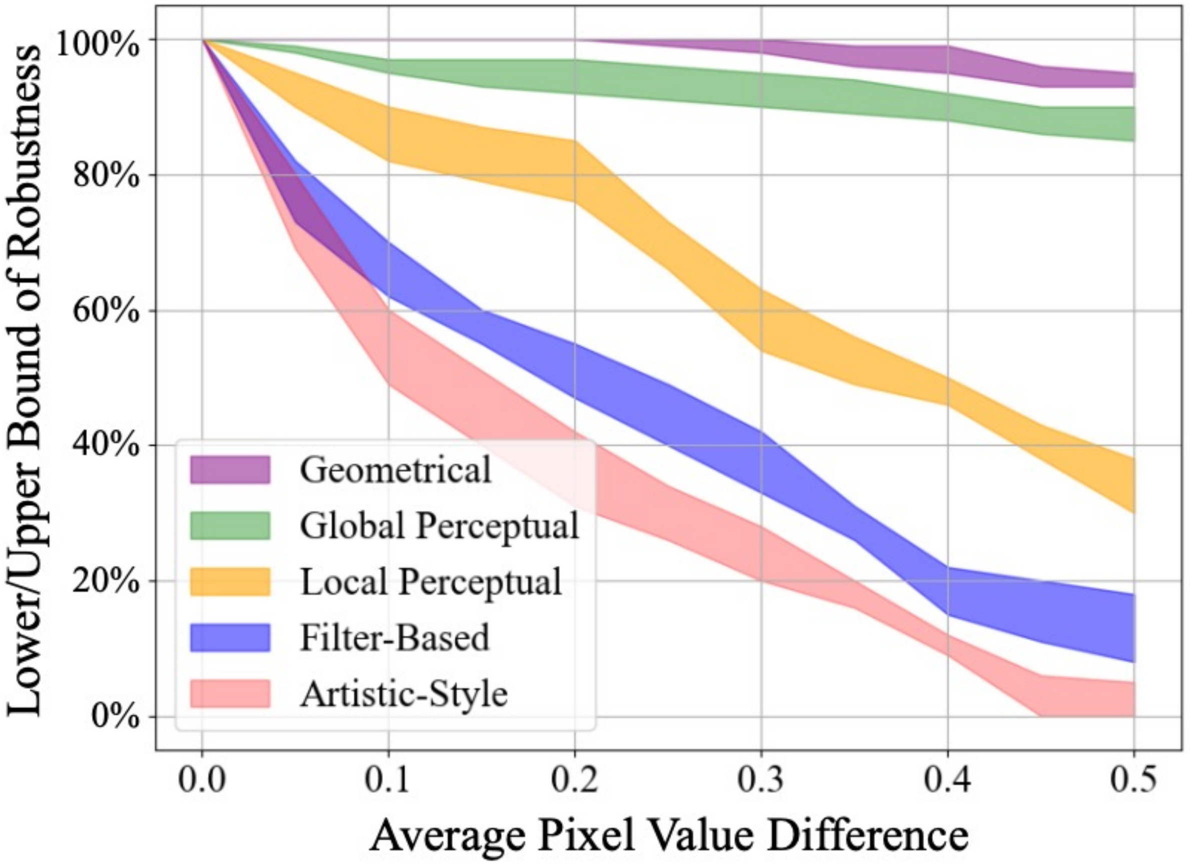
	Global	Local		
	Orientation	Hair	Eye	Nose
Upper Bound	100%	98.1%	69.7%	95.2%
Lower Bound	97.6%	95.0%	60.3%	90.3%

1. Orientation: change face orientation. 2. Hair: change hair color.
3. Eye: open/close eyes, or add glasses. 4. Nose: change nose size.

More sensitive to mutating eyes

Evaluation: Certification

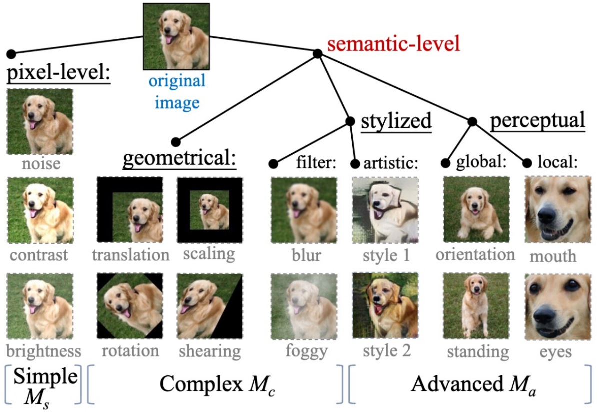
Quantitative certification over different mutations



1. Geometrical mutation is not a major concern;
2. Artistic-style and filter-based mutations are more effective (consistent to the texture-bias);
3. Local perceptual (may mutate key attributes) is also effective.

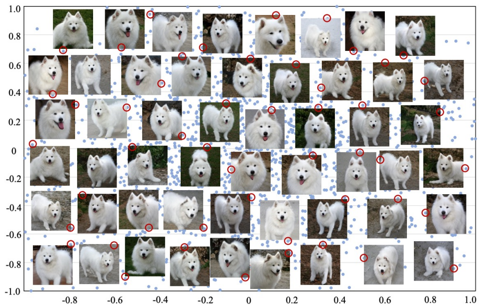
Summary

Input Mutation



Diverse input mutations

Motivation: Generative Model

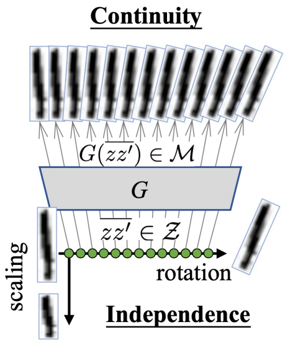


A collection of images → Latent space → **Infinite** images by (inter)extrapolation

Data-driven mutations:

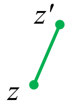
1) *Extract mutations from diverse images*
2) *Represent mutations as moving directions in latent space*

Two Requirements



Continuity: when performing mutations, $G(z)$ changes continuously with z.

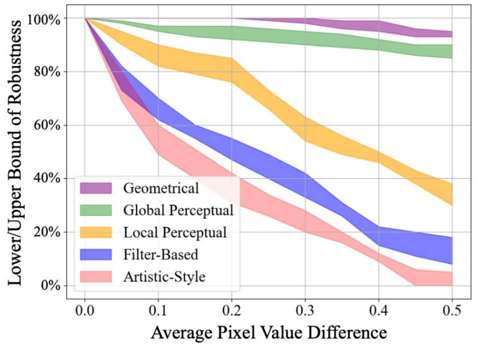
Independency: when mutating $G(z)$ into $G(z')$, $z \to z'$ should only correspond to the expected mutation.



$\overline{zz'}$ will exclusively correspond to all mutated inputs between $G(z)$ and $G(z')$.

Evaluation: Certification

Quantitative certification over different mutations



1. Geometrical mutation is not a major concern;
2. Artistic-style and filter-based mutations are more effective (consistent to the texture-bias);
3. Local perceptual (may mutate key attributes) is also effective.

Thanks!

Contact Yuanyuan for more information.

https://yuanyuan-yuan.github.io

Paper

arxiv.org/pdf/2306.06747.pdf

Code

github.com/Yuanyuan-Yuan/GCert