Semantic Manipulation of Visual Content

Sagie Benaim
Department of Computer Science, University of Copenhagen

About Me

- PhD at Tel Aviv University (04/2017 - 10/2021).
  Working with Prof. Lior Wolf.

- Postdoc at DIKU and a member of the Pioneer Center of AI (11/2021 -).
  Working with Prof. Serge Belongie.

Research Interests

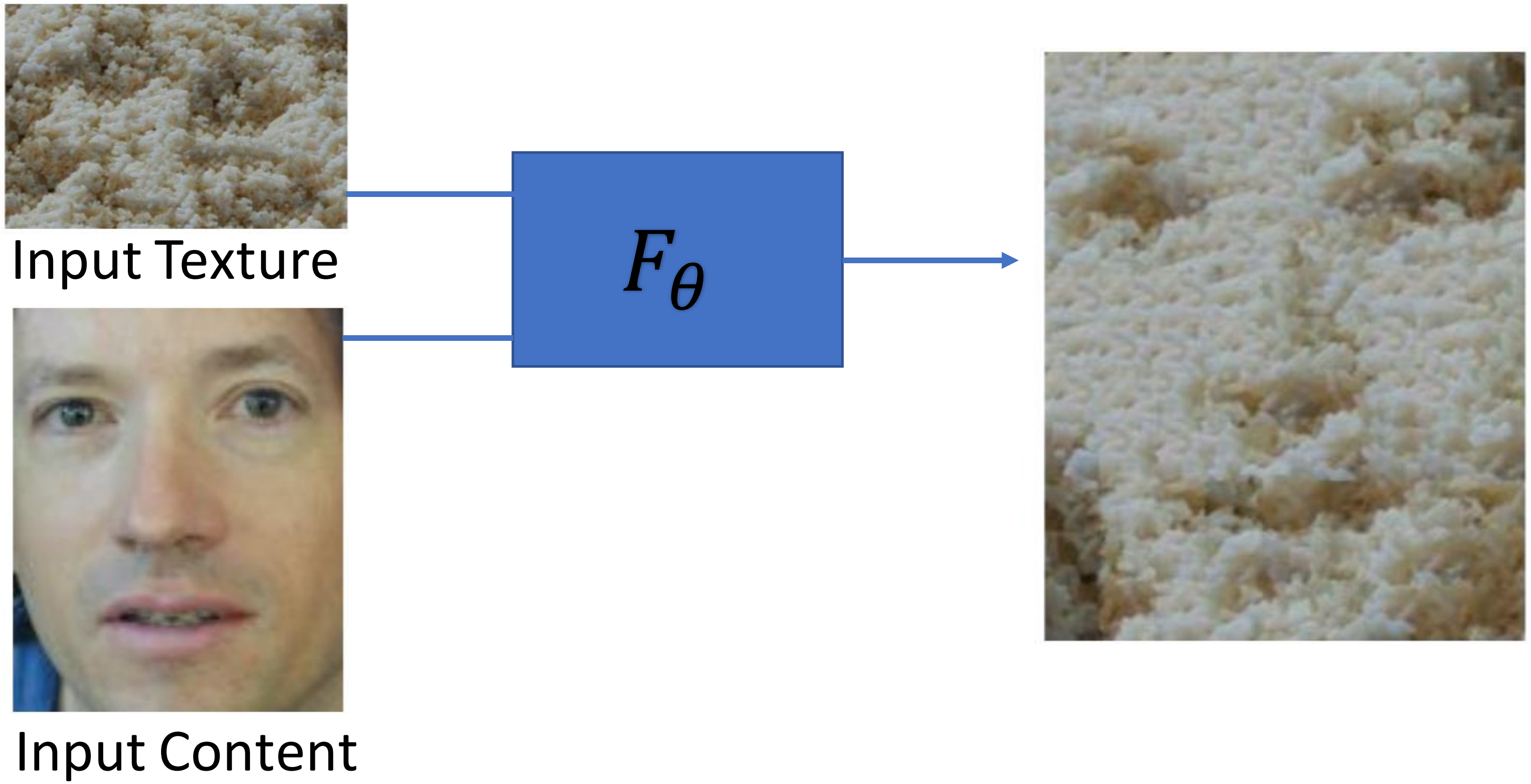
- Unsupervised, semi-supervised and self-supervised learning.
- Few-shot learning and domain adaptation. Emphasis on low-resource generative models.
- Content creation and manipulation.
- Computer vision for AR/VR.

What is a natural image?

Intelligent machines must **understand** perceived content

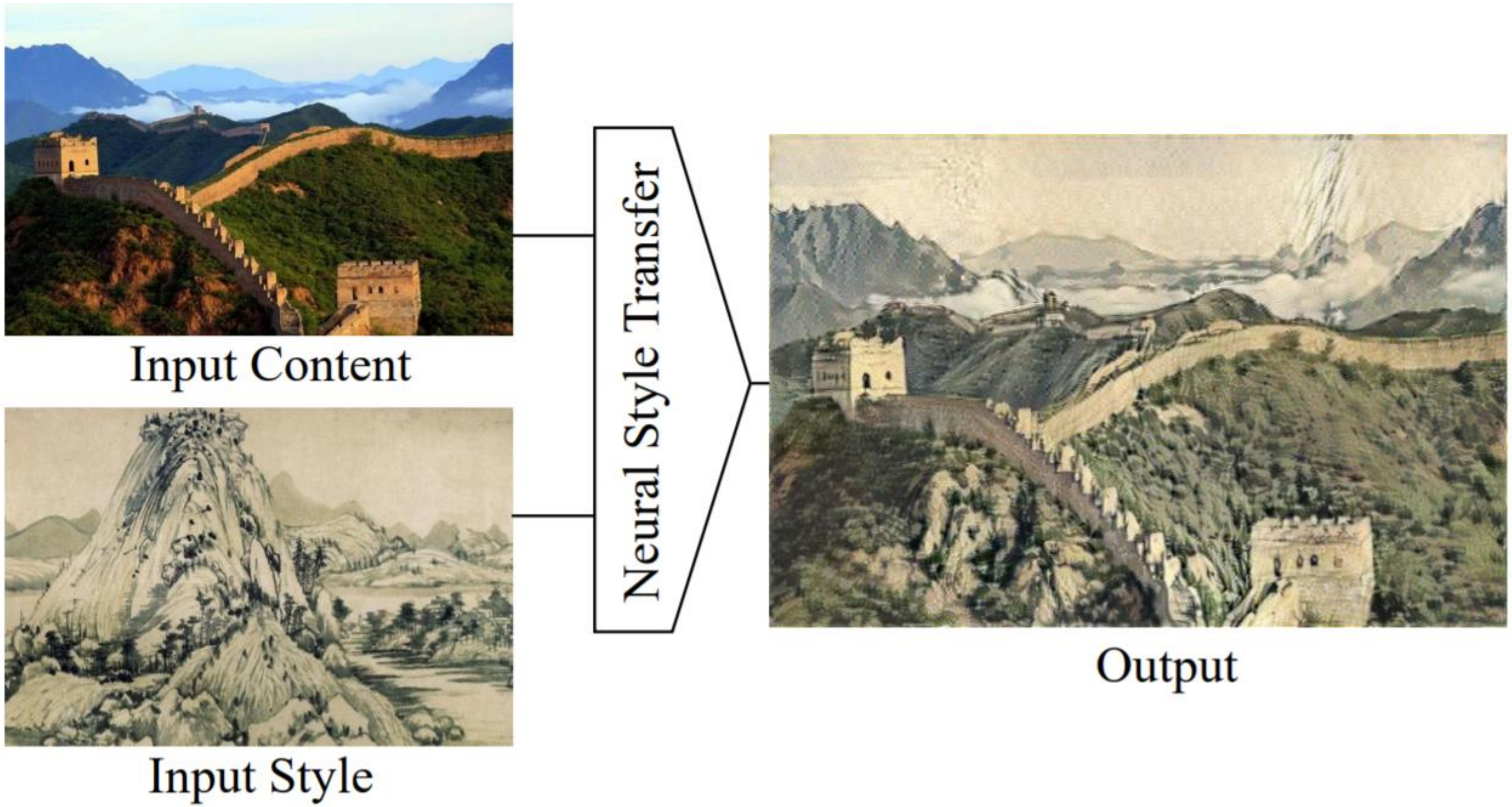
Understanding by creating/manipulating: “What I cannot create, I do not understand” (Richard Feynman)

Texture Manipulation



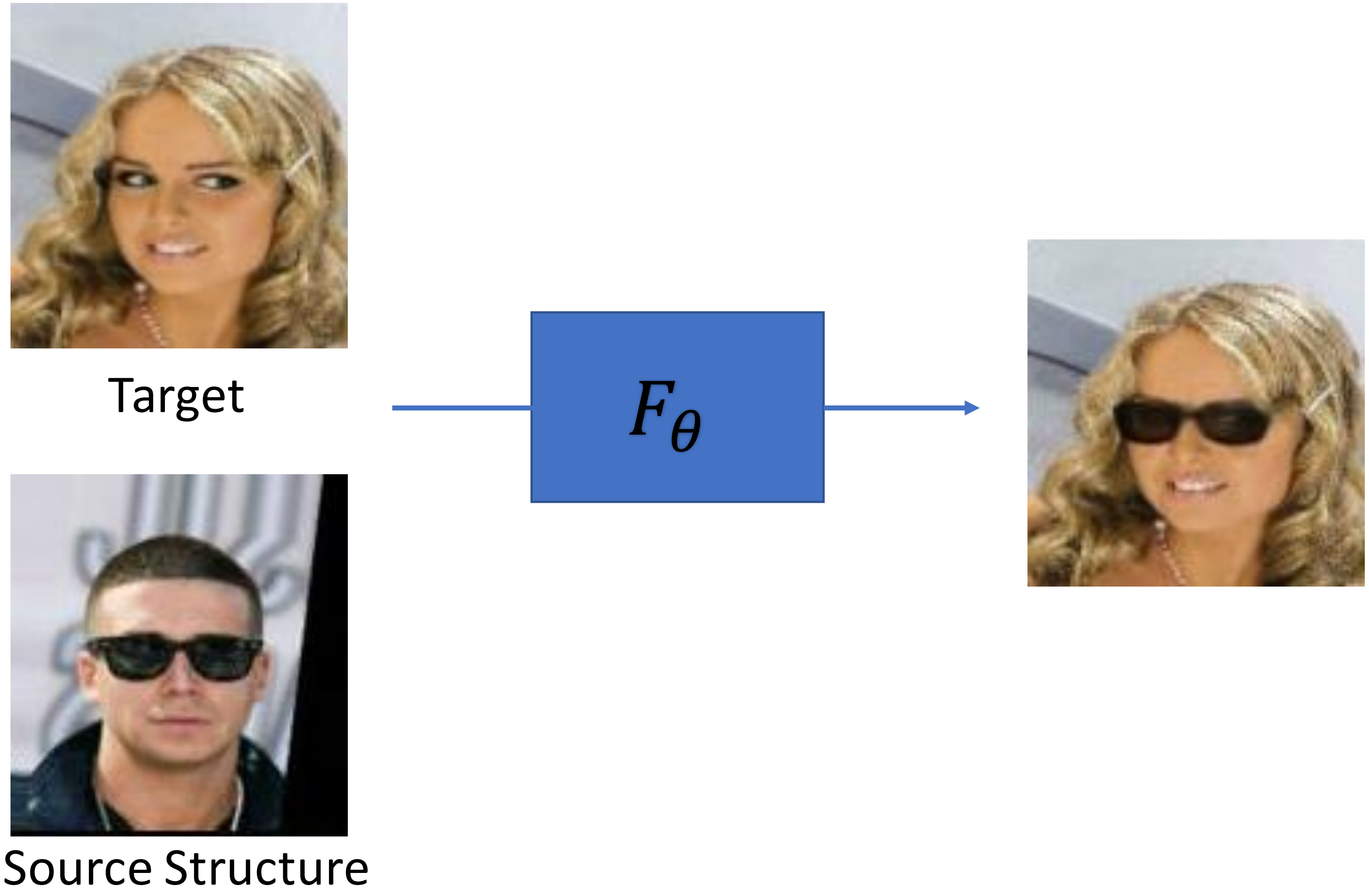
A.A.Efros, W.T.Freeman. “Image Quilting for Texture Synthesis and Transfer”. SIGGRAPH01

Style Manipulation

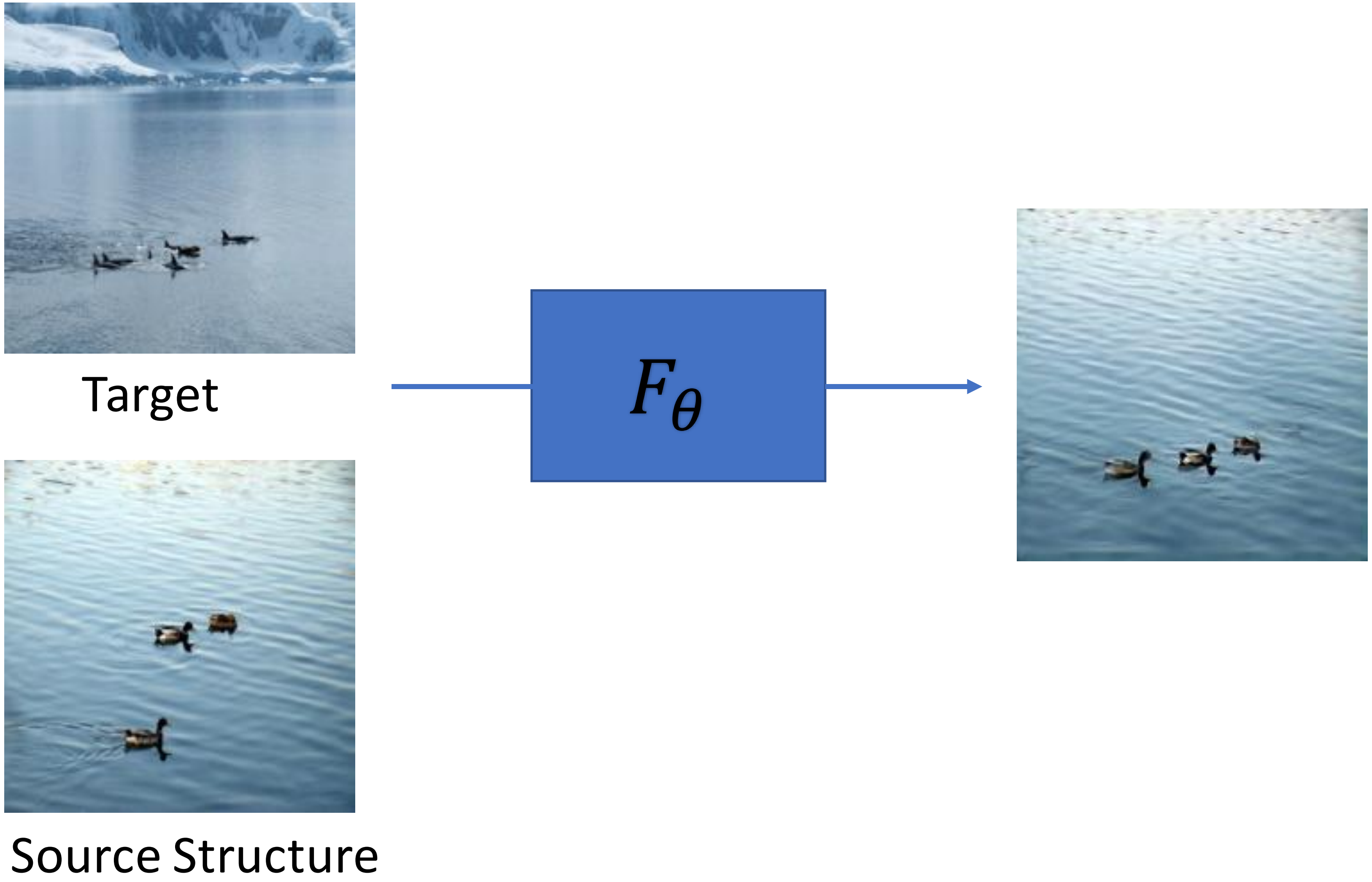


L. A. Gatys, A. S. Ecker, and M. Bethge. “A neural algorithm of artistic style”. 2015.

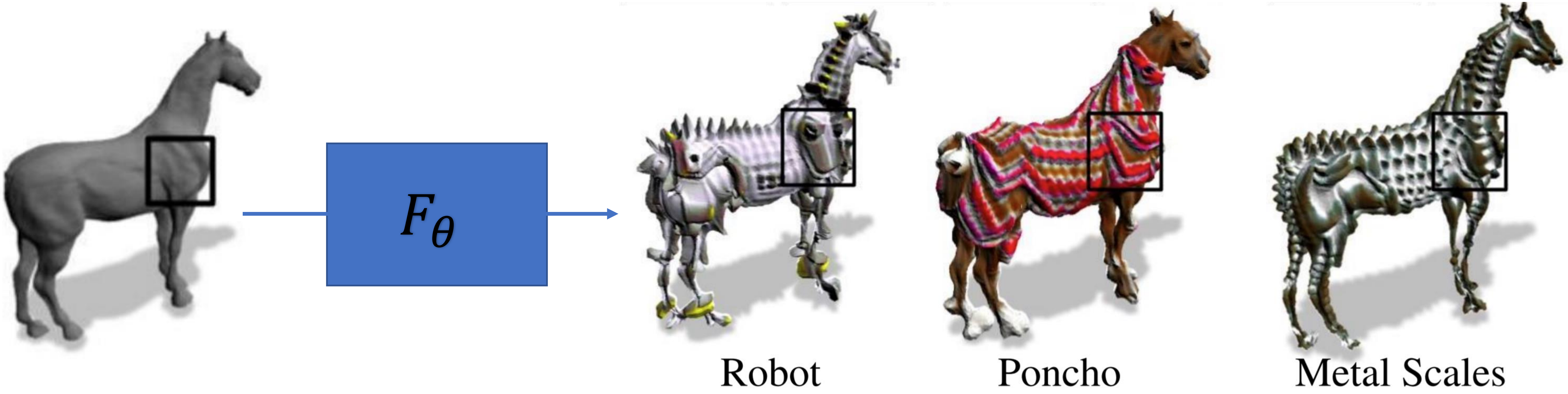
Semantic Manipulation



Semantic Manipulation



Semantic Manipulation



Applications

Architecture

Video games

Movies

Advertising

Autonomous Driving Simulations

AR/VR

Augmented Reality

Amazon rolls out a new AR shopping feature for viewing multiple items at once

Sarah Perez @sarahintampa / 2:00 PM GMT+2 • August 25, 2020

Comment

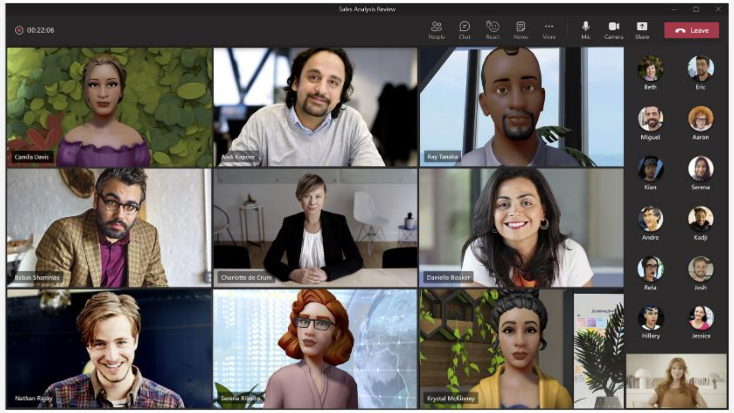
Image Credits: Amazon

Amazon is rolling out a new augmented reality shopping tool, Room Decorator, that will allow you to see furniture and other home décor in your own space. While the retailer had experimented with AR tools in the past, what makes Room Decorator different is that it's

Facebook gives a glimpse of metaverse, the planned virtual reality world – Guardian News

Mesh for Microsoft Teams aims to make collaboration in the 'metaverse' personal and fun

November 2, 2021 | John Roach



Sundar Pichai thinks of the metaverse as more immersive computing with AR

Apple AR glasses are nearly ready for your eyes, says key investment group

By Gerald Lynch last updated 2 days ago

Polishing up those specs

(Image credit: Martin Hajek/idropnews)

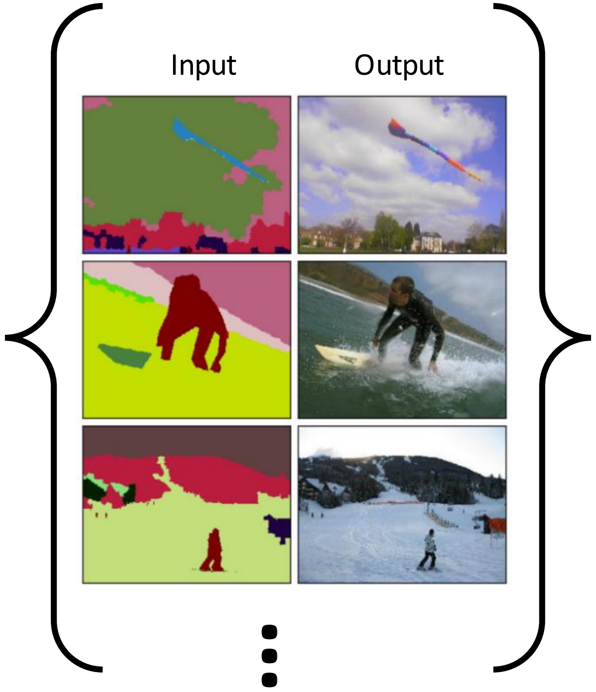
Apple's AR glasses may be approaching their big reveal, according to the tech investment analysts at megabank Morgan Stanley.

Part I: Semantic Manipulation of **Images**

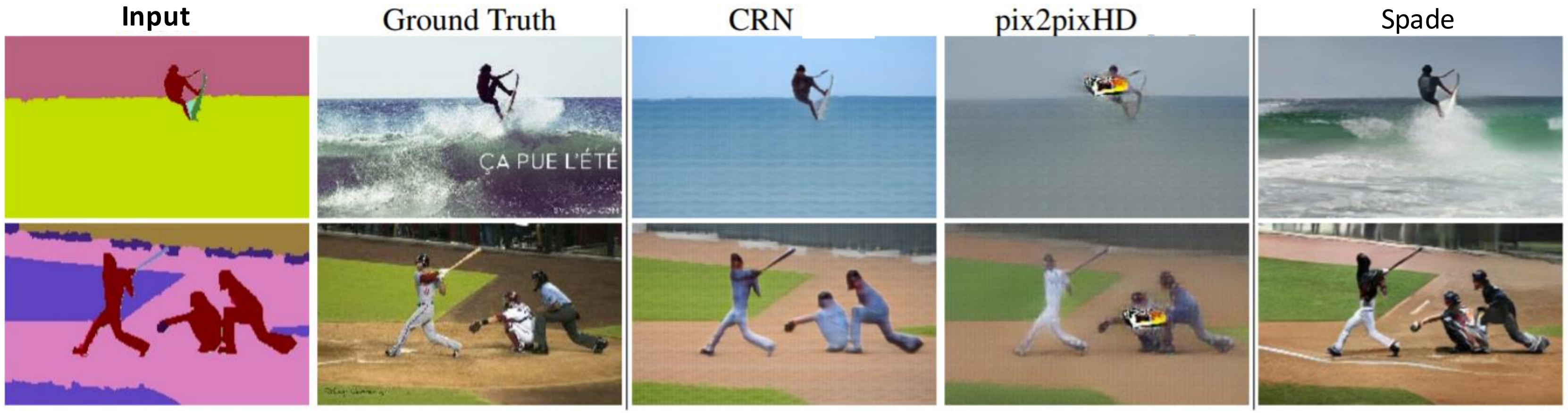
Multi-Image Approaches

Supervised (Paired) Setting

Train



Test



Unsupervised (Unpaired) Setting

A

Faces without glasses

B

Faces with glasses

Control Structure of Generated Faces (Transfer Glasses)

Unsupervised Approaches

O. Press, T. Galanti, **S. Benaim,** L. Wolf.
Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer.
In **ICLR 2019.**

S. Benaim, M. Khaitov, T. Galanti, L. Wolf.

In **ICCV, 2019.**

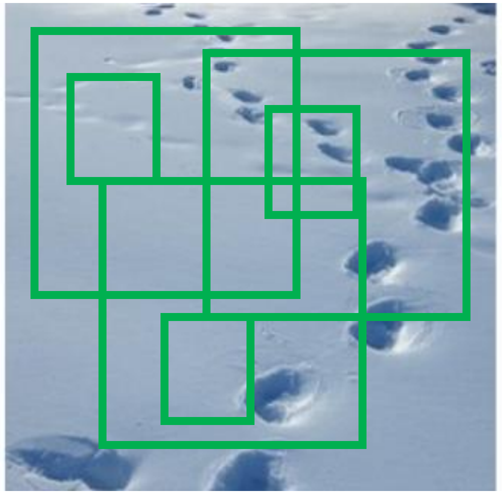
R. Mokady, **S. Benaim**, L. Wolf, A. Bermano.
Mask Based Unsupervised Content Transfer.
In **ICLR, 2020.**

Require a large collection of images from both domains

Patch-Based Approaches

Multi-Image Distribution

Multi-Scale Patch Distribution

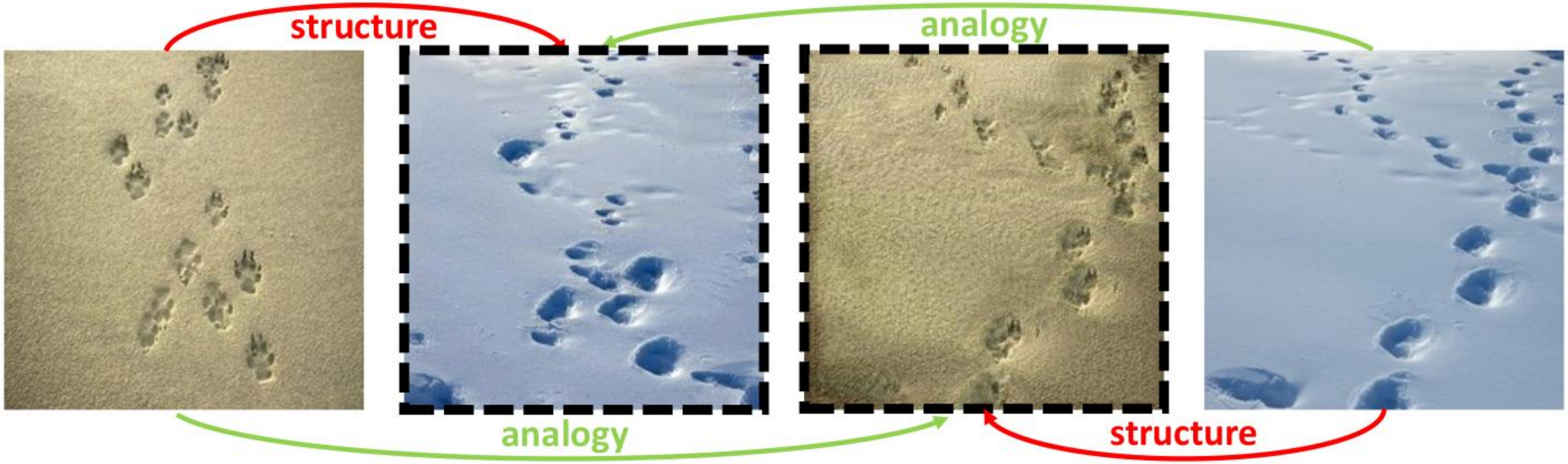


Karras et al., A Style-Based Generator Architecture for Generative Adversarial Networks. CVPR 2019

Structural-analogy from a **Single Image Pair**

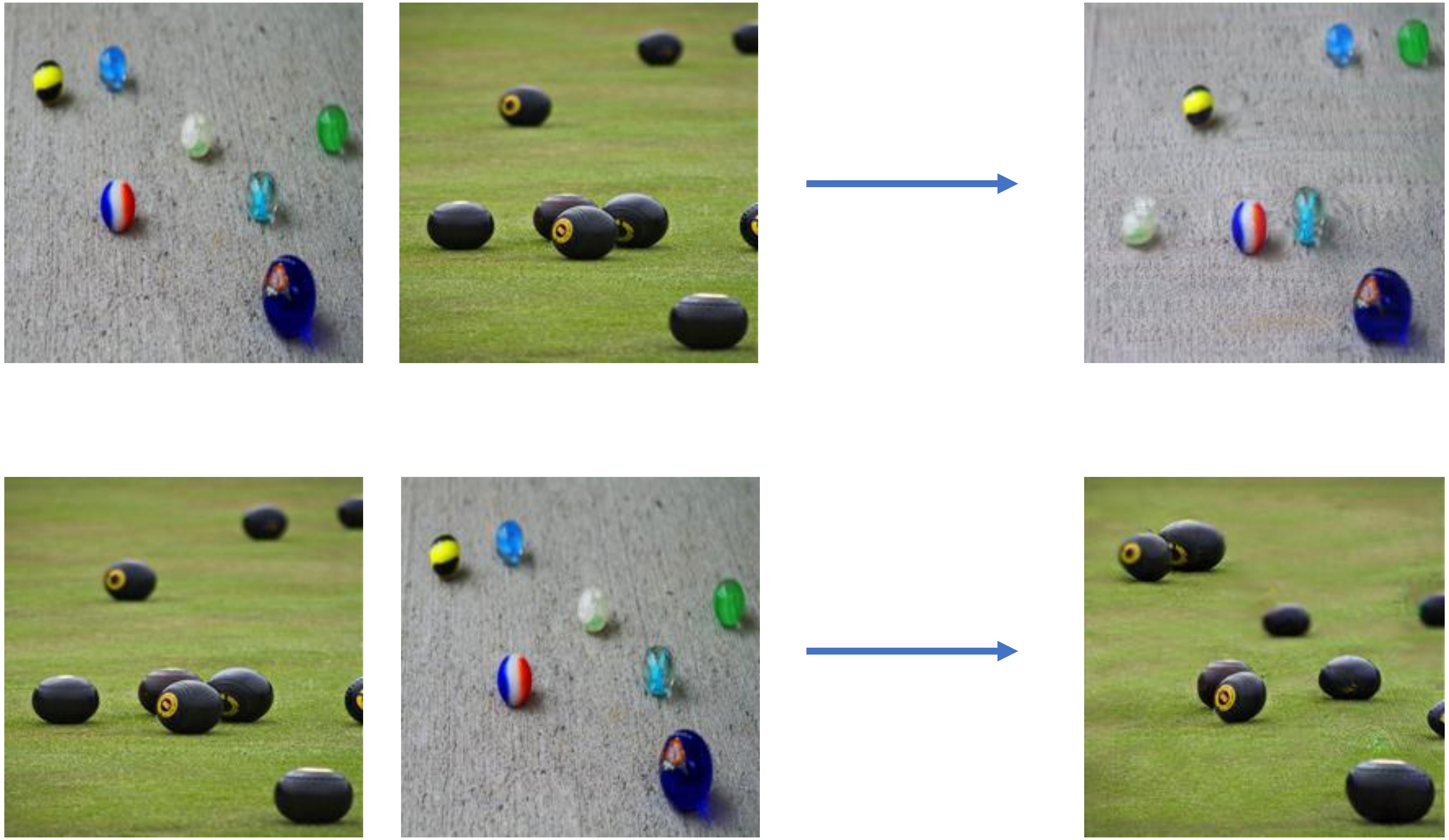
S. Benaim*, R. Mokady*, A. Bermano, D Cohen-Or, L. Wolf. CGF 2020.

Generate an image which is **aligned to the source image** but depicts **structure from a target image**



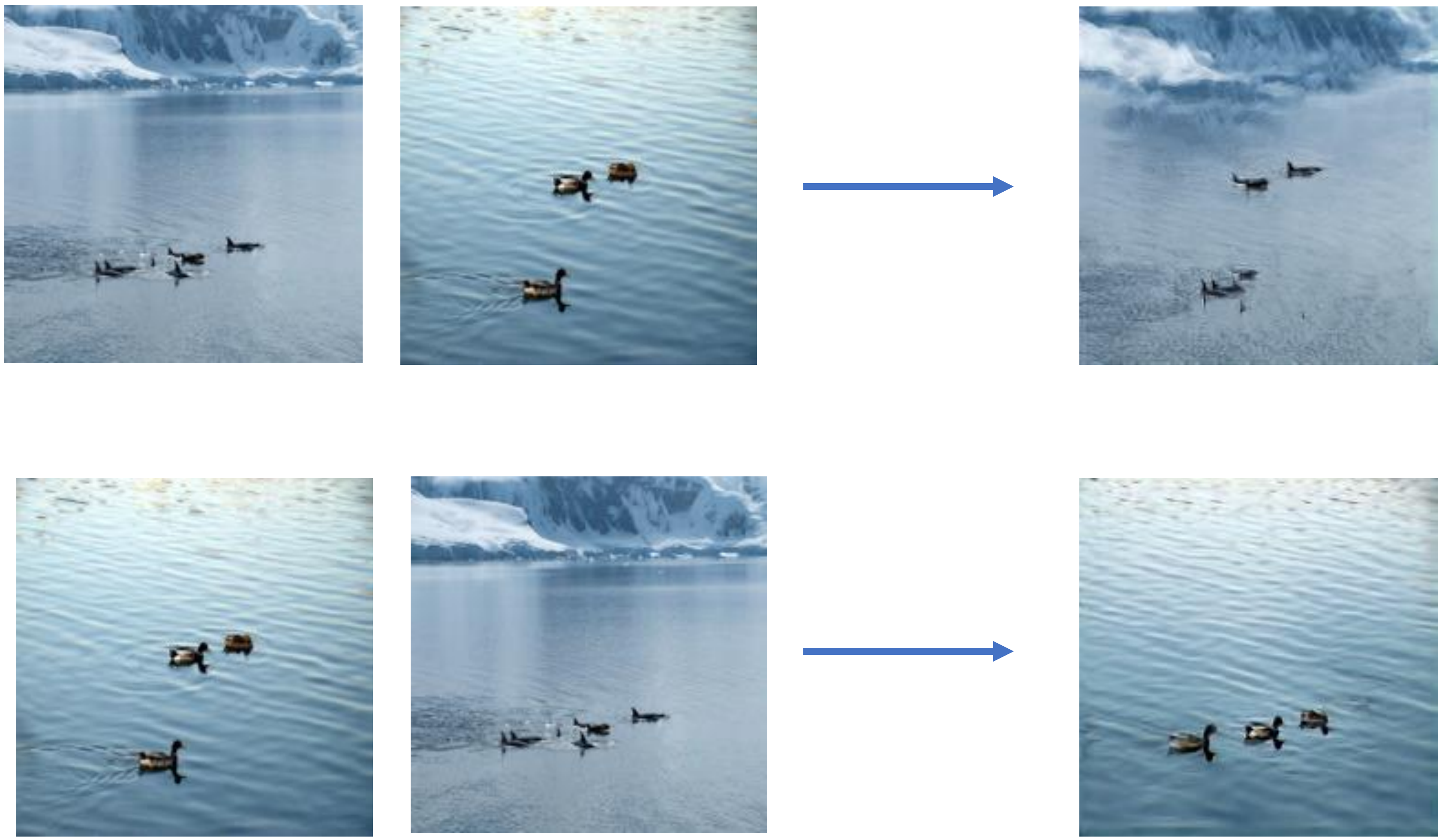
Structural Analogy

Target Source Output



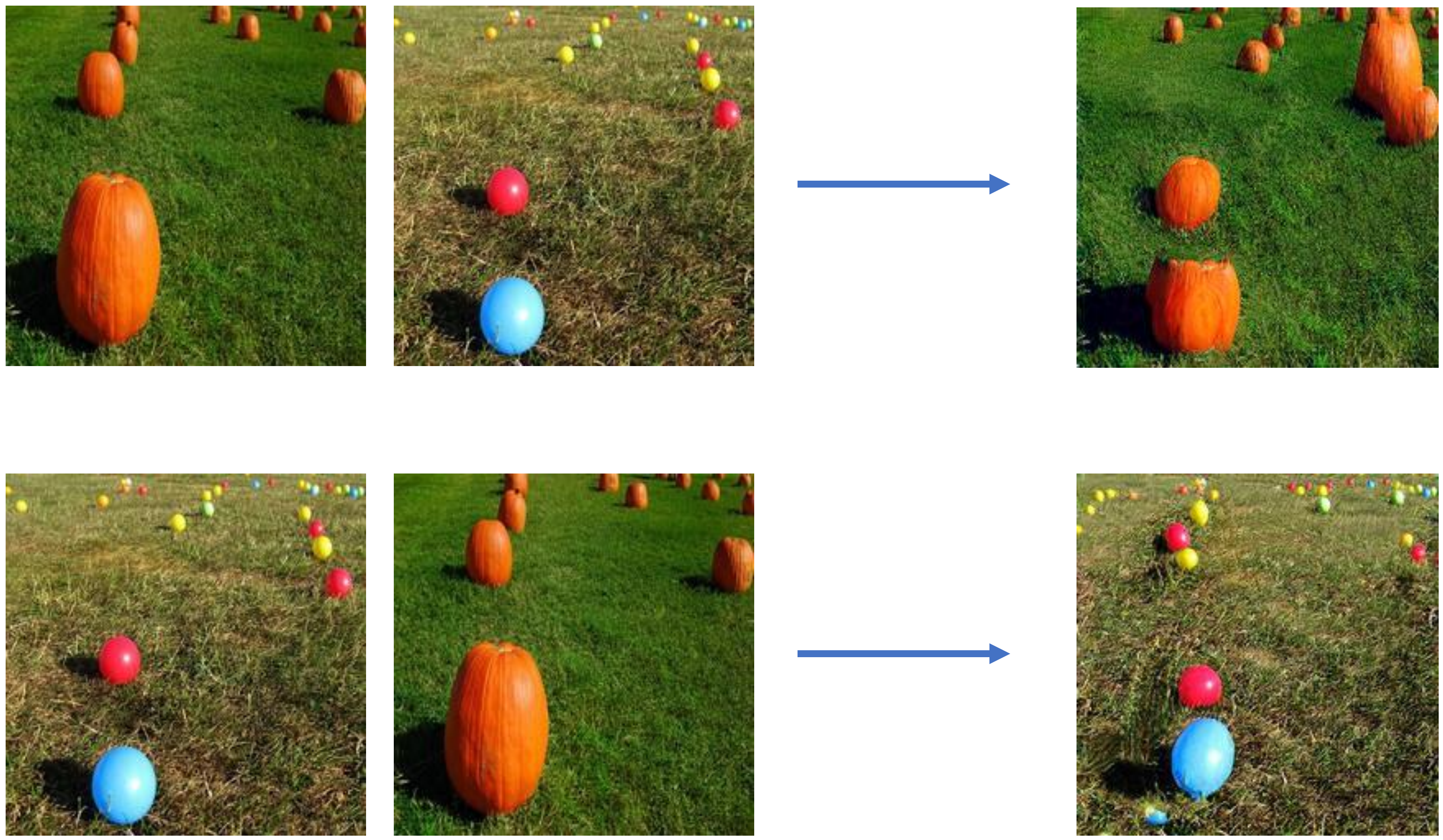
Structural Analogy

Target Source Output



Structural Analogy

Target Source Output



Style Transfer

Style

Content

Result

Deep Image Analogy

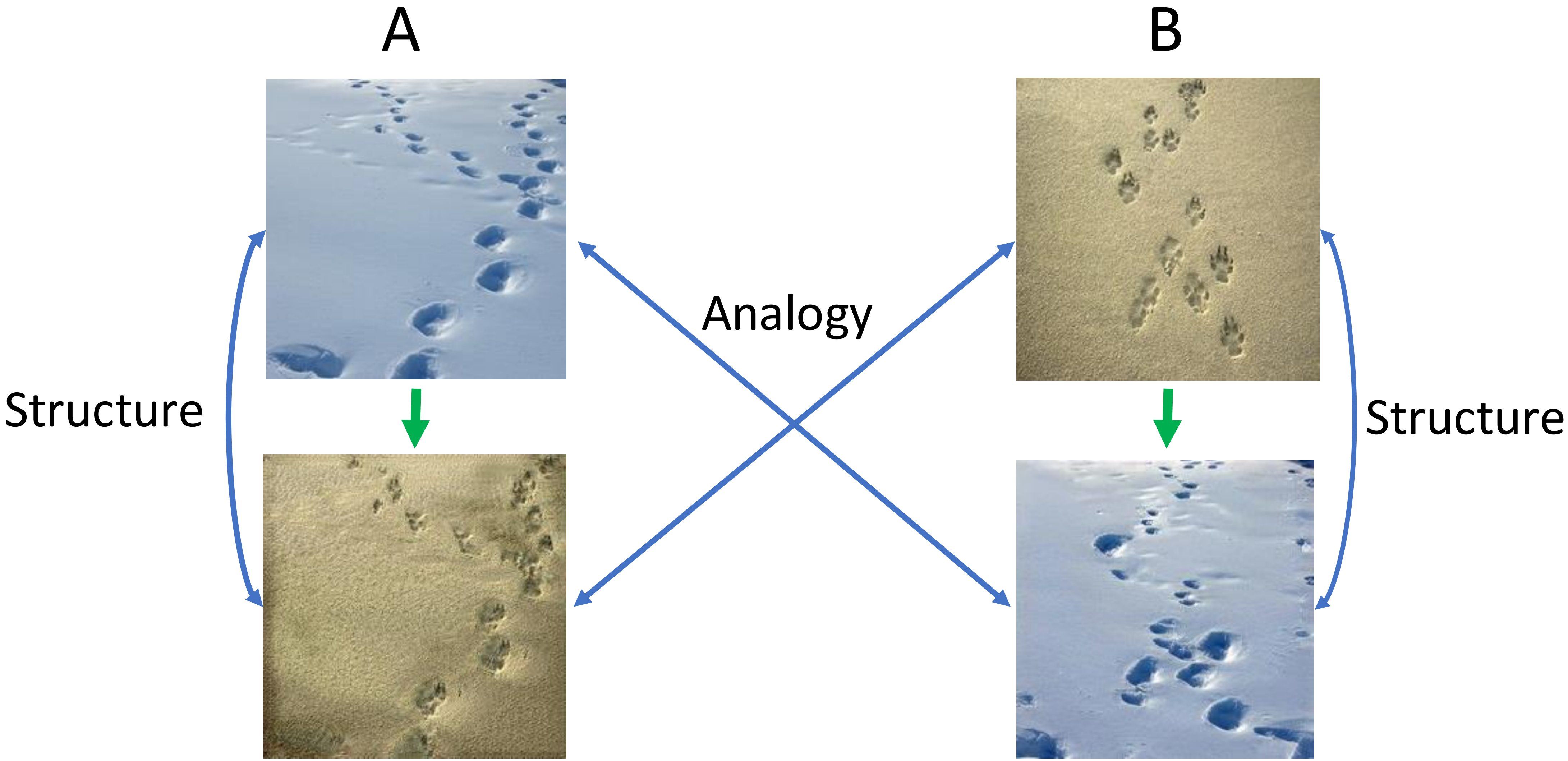
Style

Content

Result

Cannot Change Object Shape

Structural Analogy



Motivation

A

B

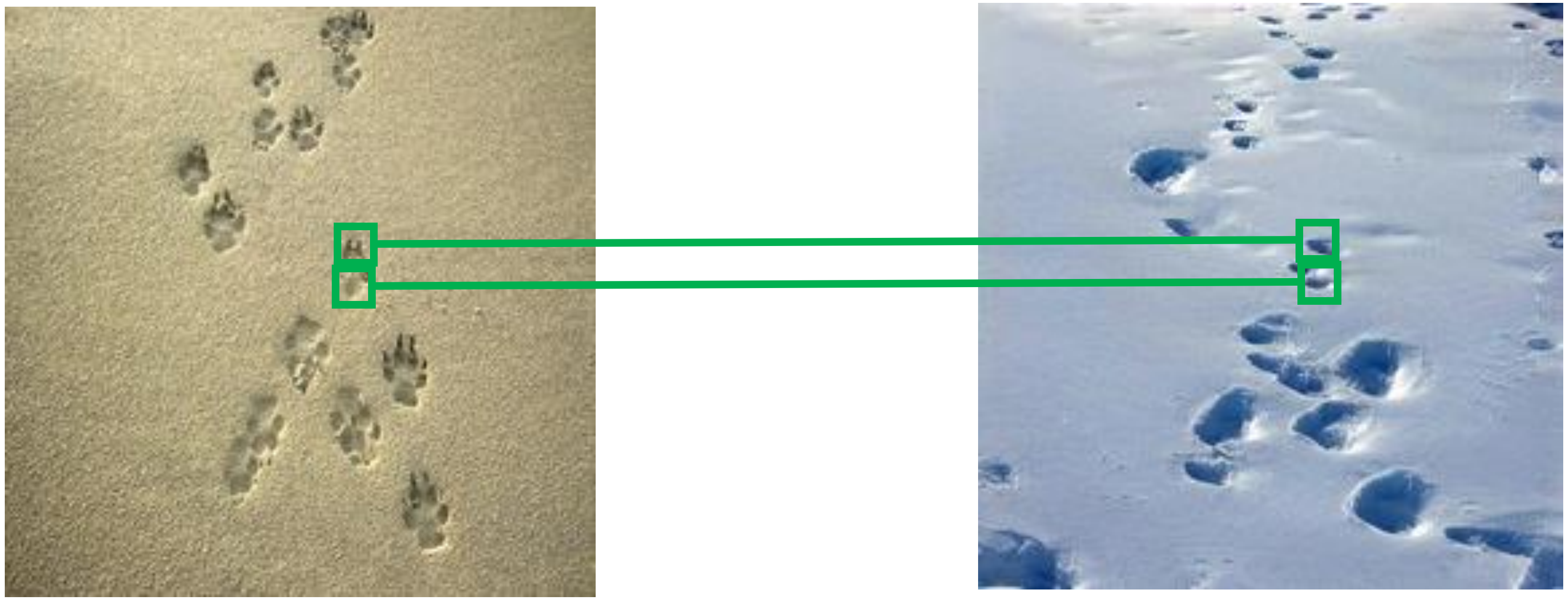
Motivation

A B

Motivation

A

B



Proposed Hierarchical Approach

Coarsest scale:
Large Patches

$\bar{a}^0$ (Unconditional)
$\overline{ab}^0$ (Conditional)

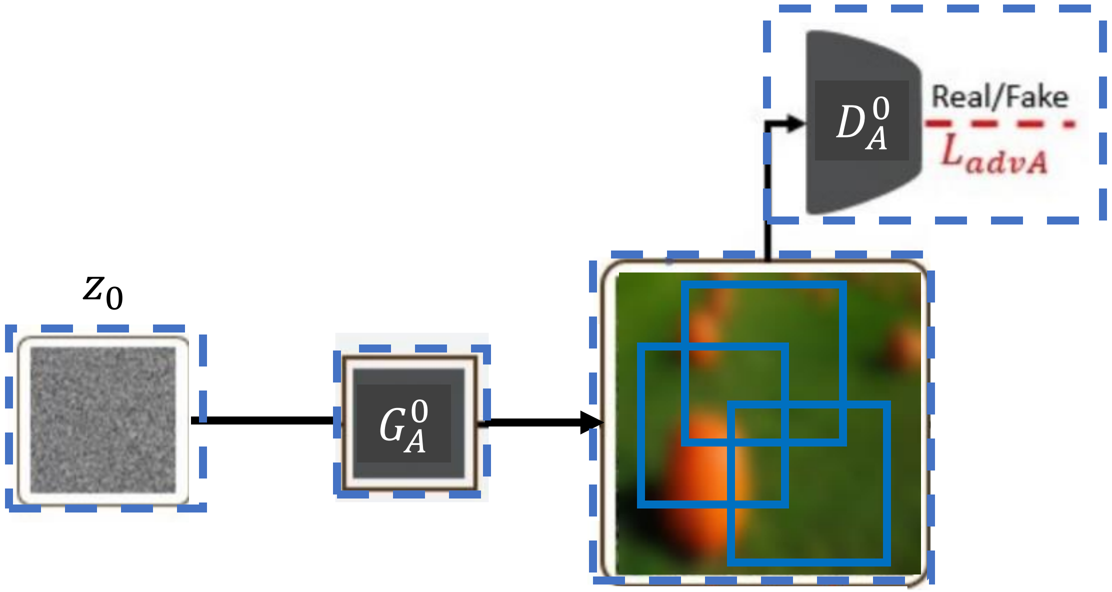
LEVEL $= 0$

Finest scale:
Small Patches

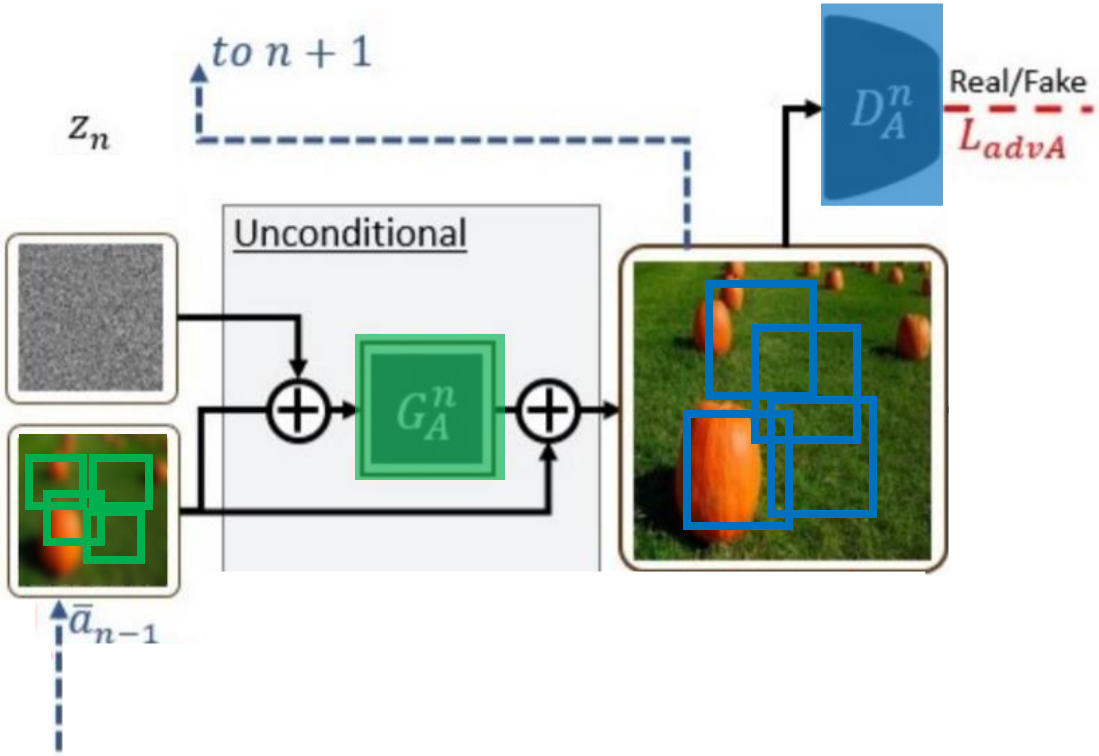
$\bar{a}^N$ (Unconditional)
$\overline{ab}^N$ (Conditional)

LEVEL $= N$

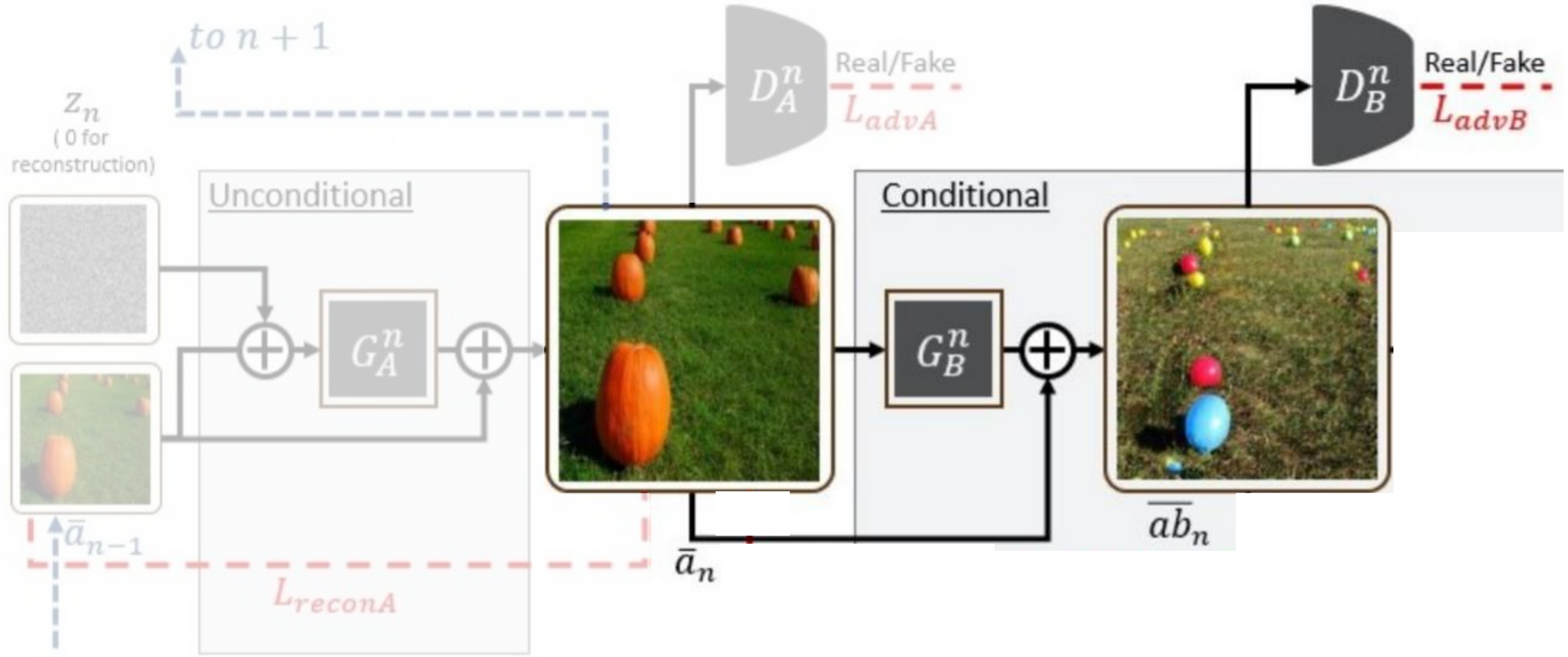
Unconditional Generation (Level 0)



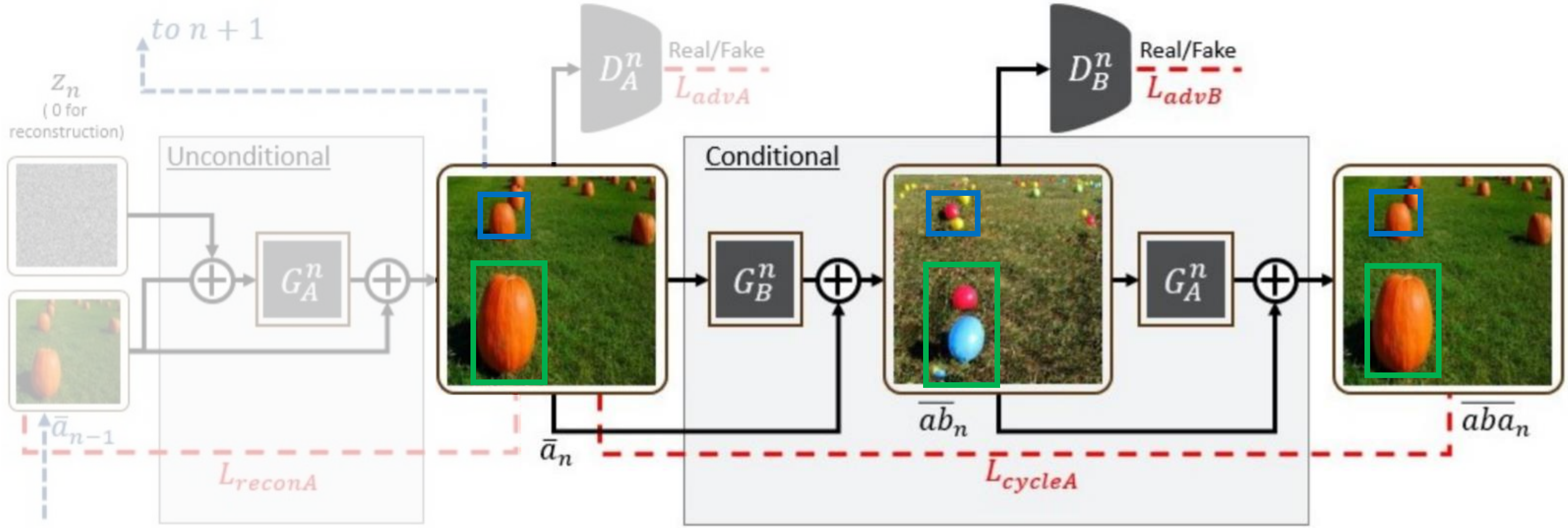
Unconditional Generation (Level n)



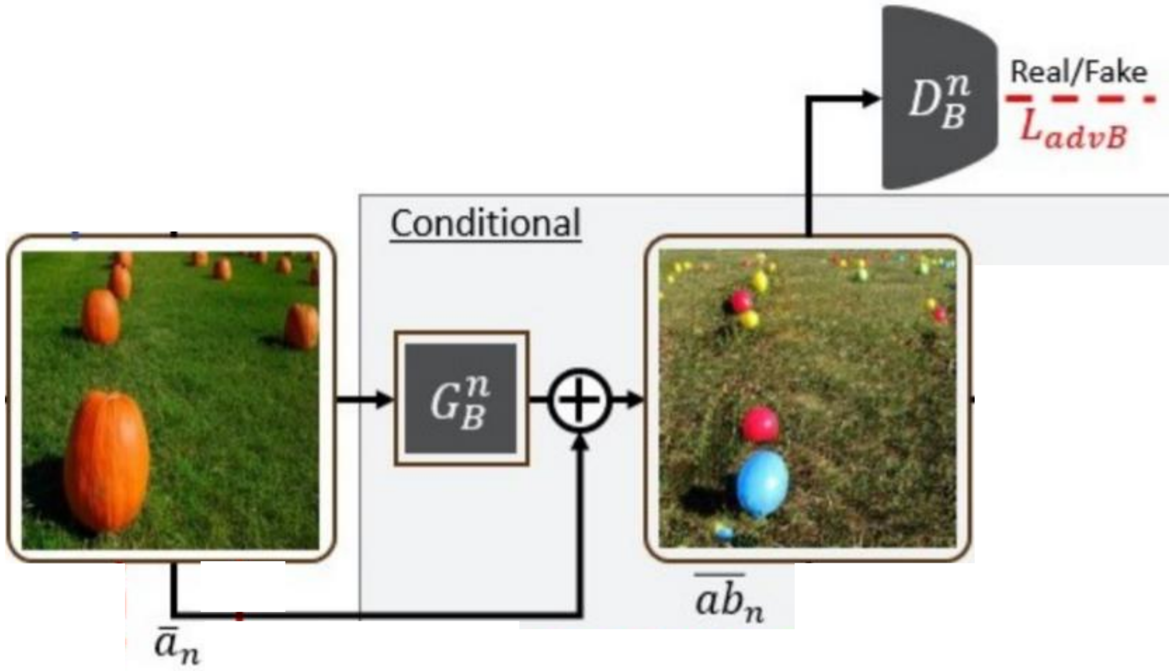
Conditional Generation (Level n)

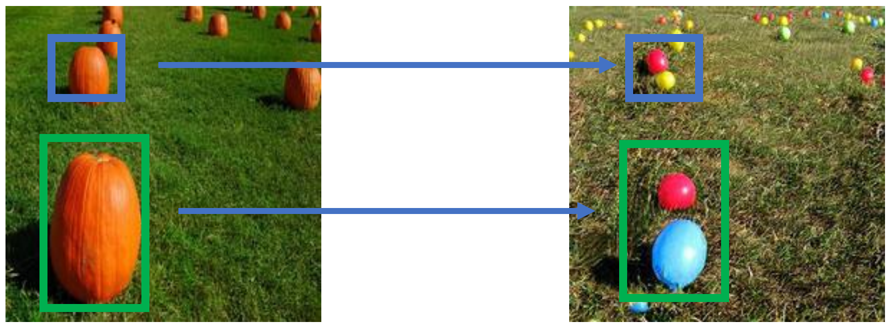


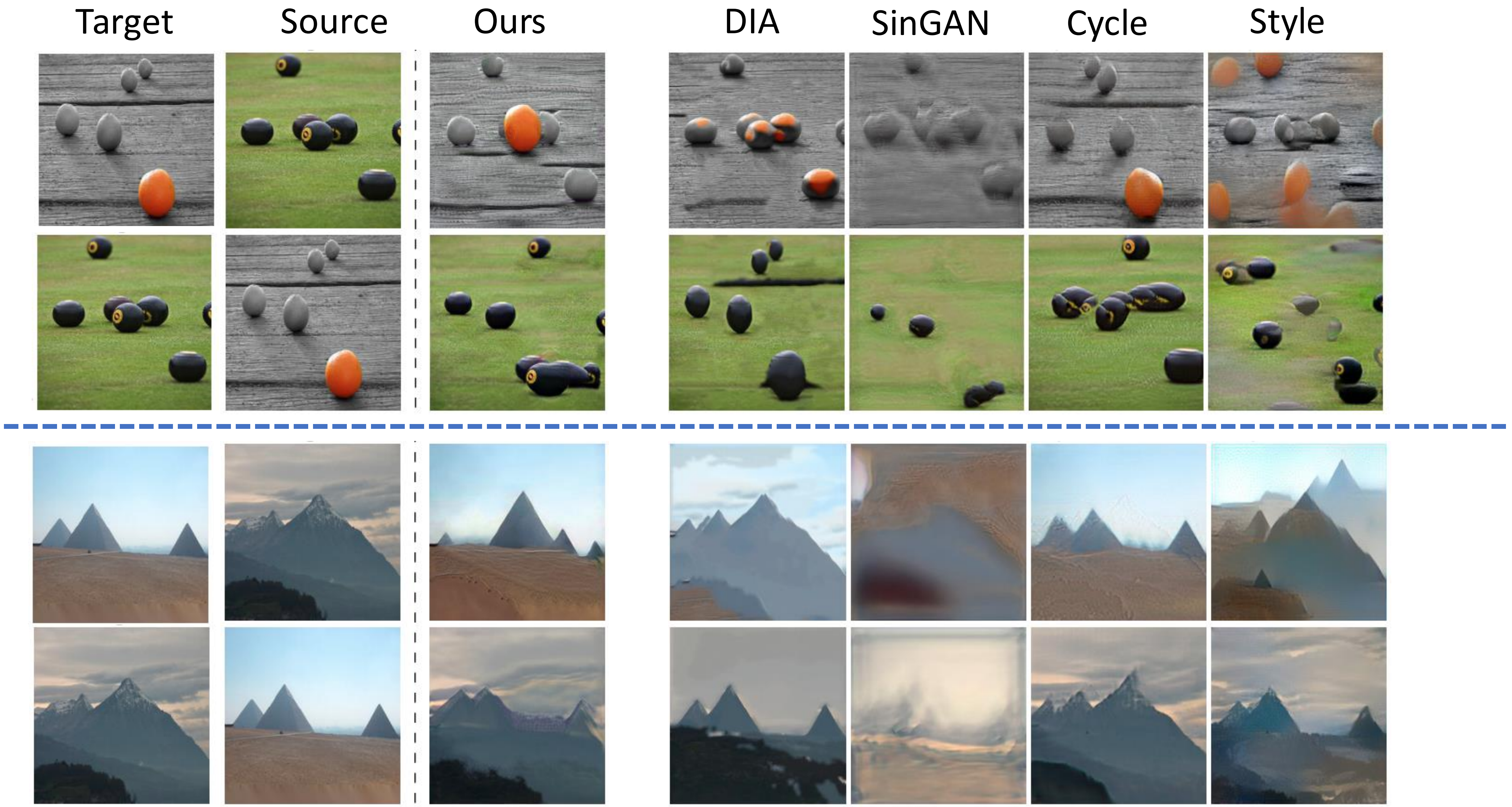
Conditional Generation (Level n)



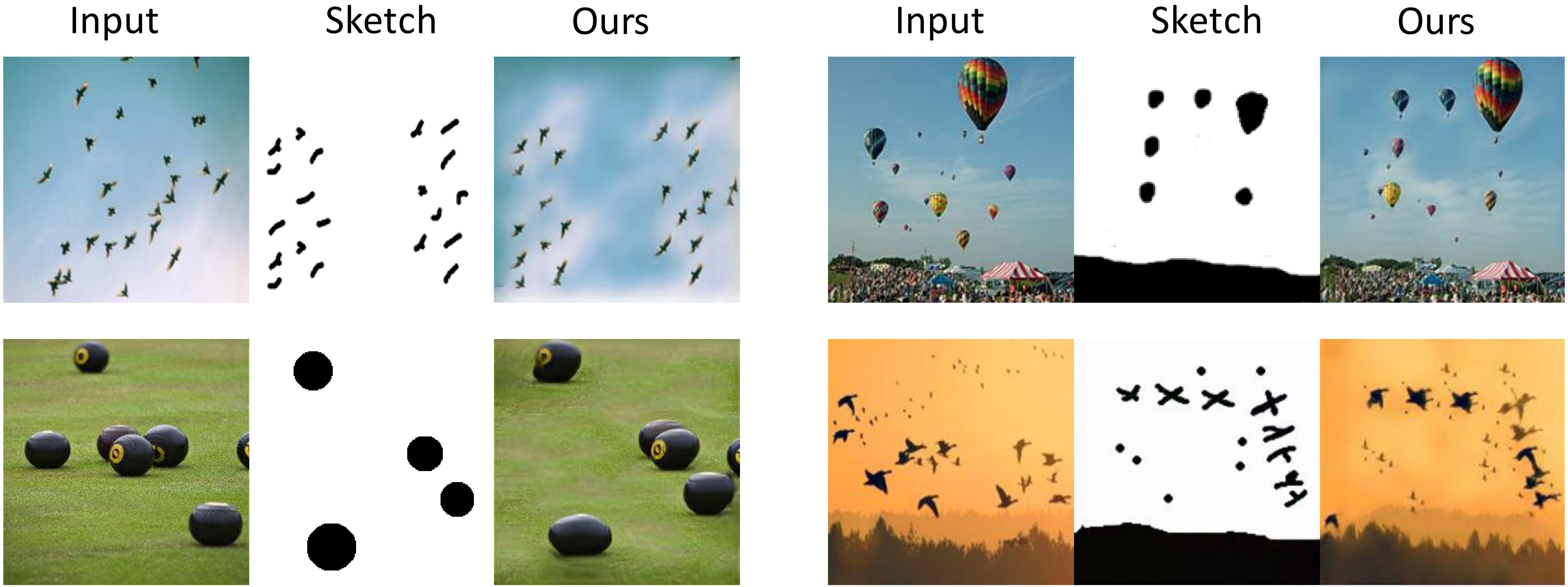
Coarse and Mid Scales: Residual Training






Target
Source
Ours
DIA
SinGAN
Cycle
Style

Paint to Image



Video Generation

Part II: Semantic Manipulation of **Videos**

SpeedNet: **Learning the Speediness in Videos**

S. Benaim, A. Ephrat, O. Lang, I. Mosseri, W. T. Freeman, M. Rubinstein, M. Irani, T. Dekel. CVPR 2020.

Slower

Normal speed

Faster

https://speednet-cvpr20.github.io/

Automatically predict "speediness"

Uniform Speed Up (2x)

Adaptive speed up (2x)

Other Applications:

- Self-supervised action recognition
- Video retrieval

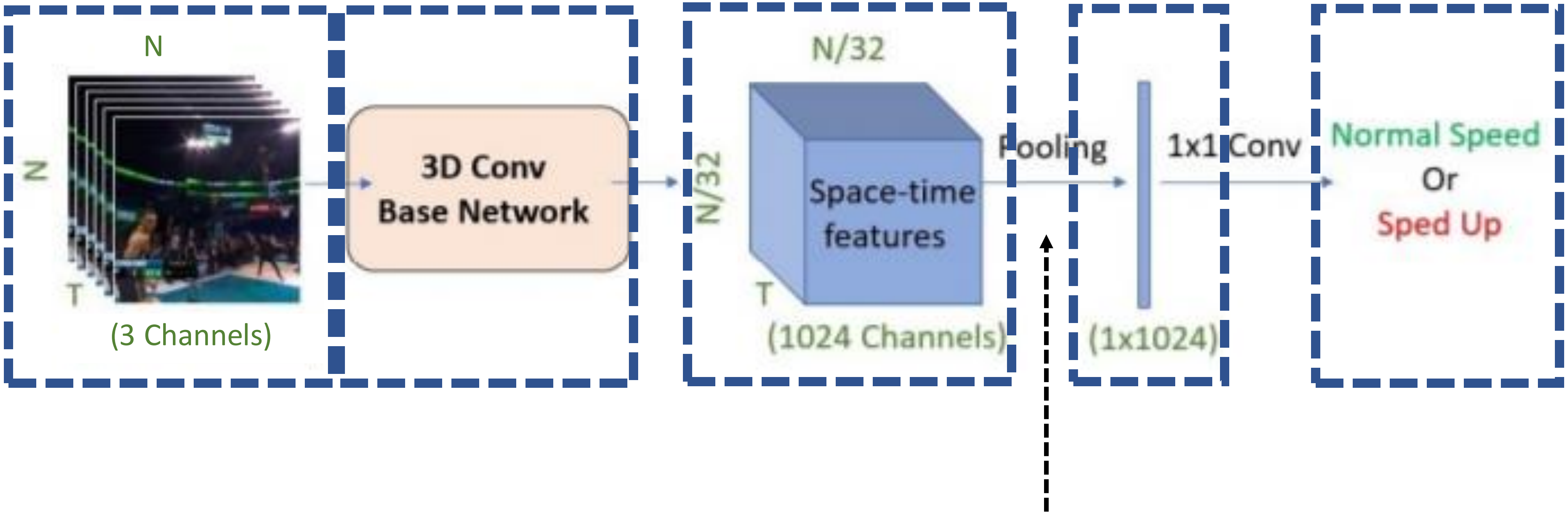
Training SpeedNet

Self-supervised training

Input video

Sped Up

Training SpeedNet



Spatial Max Pooling
Temporal Average Pooling

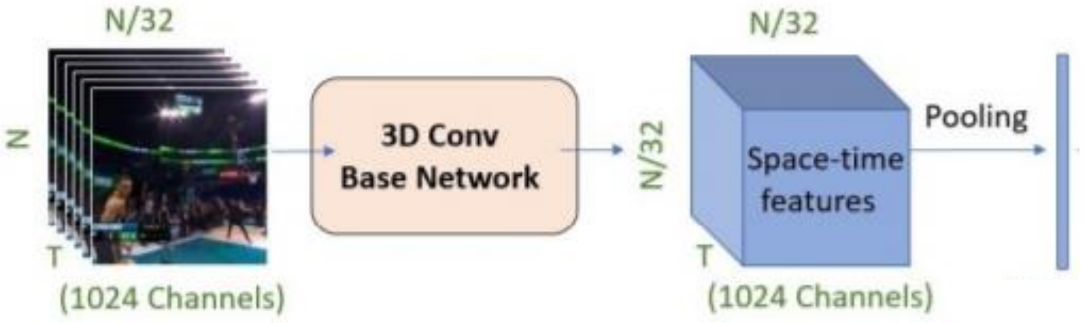
`

Training SpeedNet: Artificial Cues

- Spatial augmentations
- Temporal augmentations
- Same-batch training

Spatial Augmentations

- Random resize of input (both downsample and upsample)
- Network cannot rely on size dependent factors



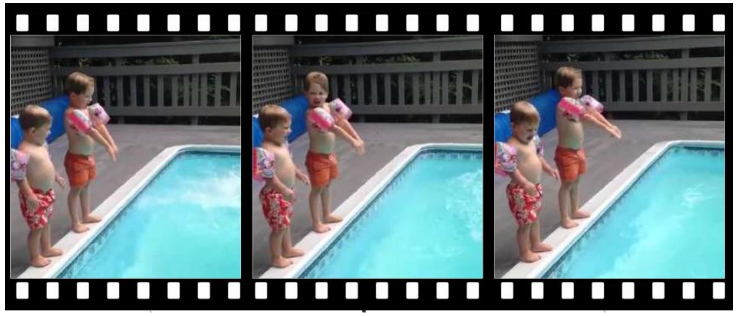
Temporal Augmentations

- Normal speed sample rate: 1-1.2x
- Sped up sample rate: 1.7-2.2x
- Randomly skip frames with probability 1 – 1/f where f is randomly chosen randomly in the desired range.

Same Batch Training

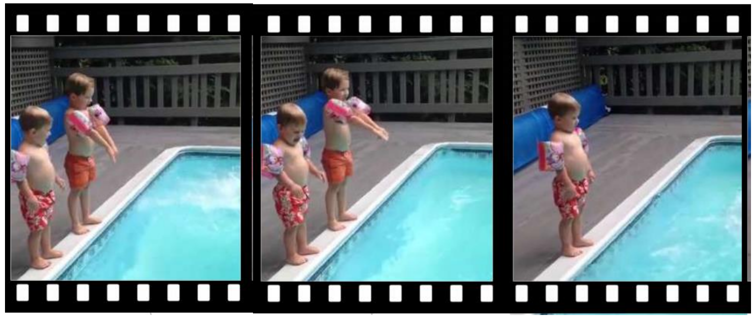
Same Batch

Normal speed



…

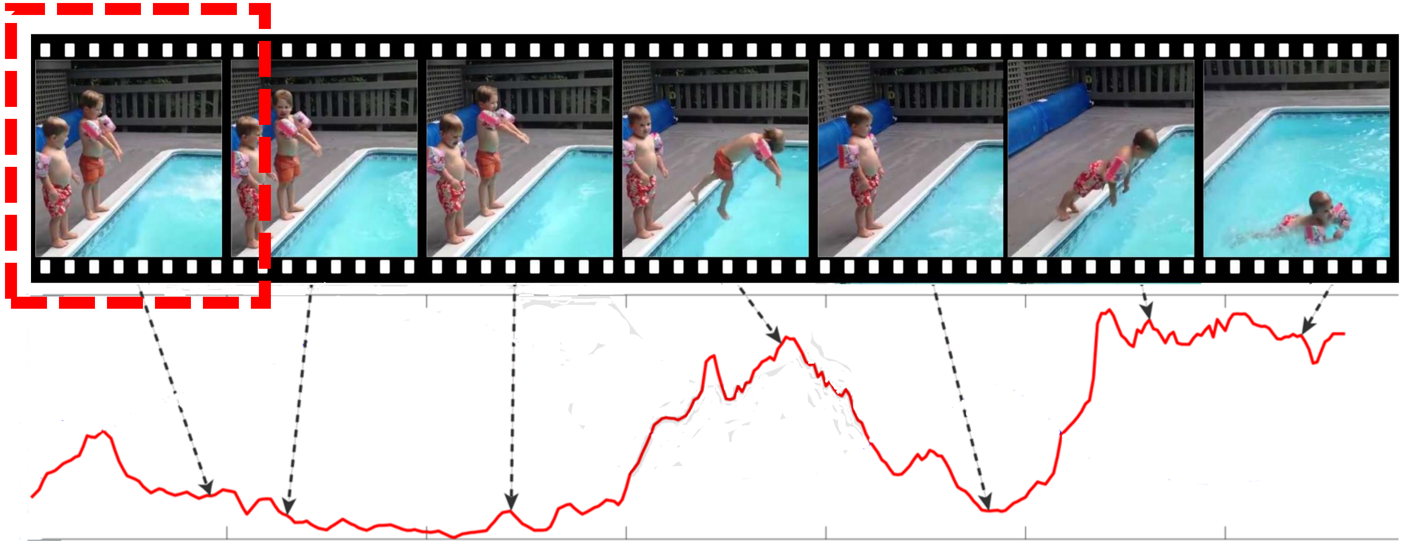
Speed up



…

Adaptive video speedup

Inference on full **sped-up** video



From Speediness to Adaptive Speedup

Original 1x video

N videos of increasing speed

1x video (T frames)

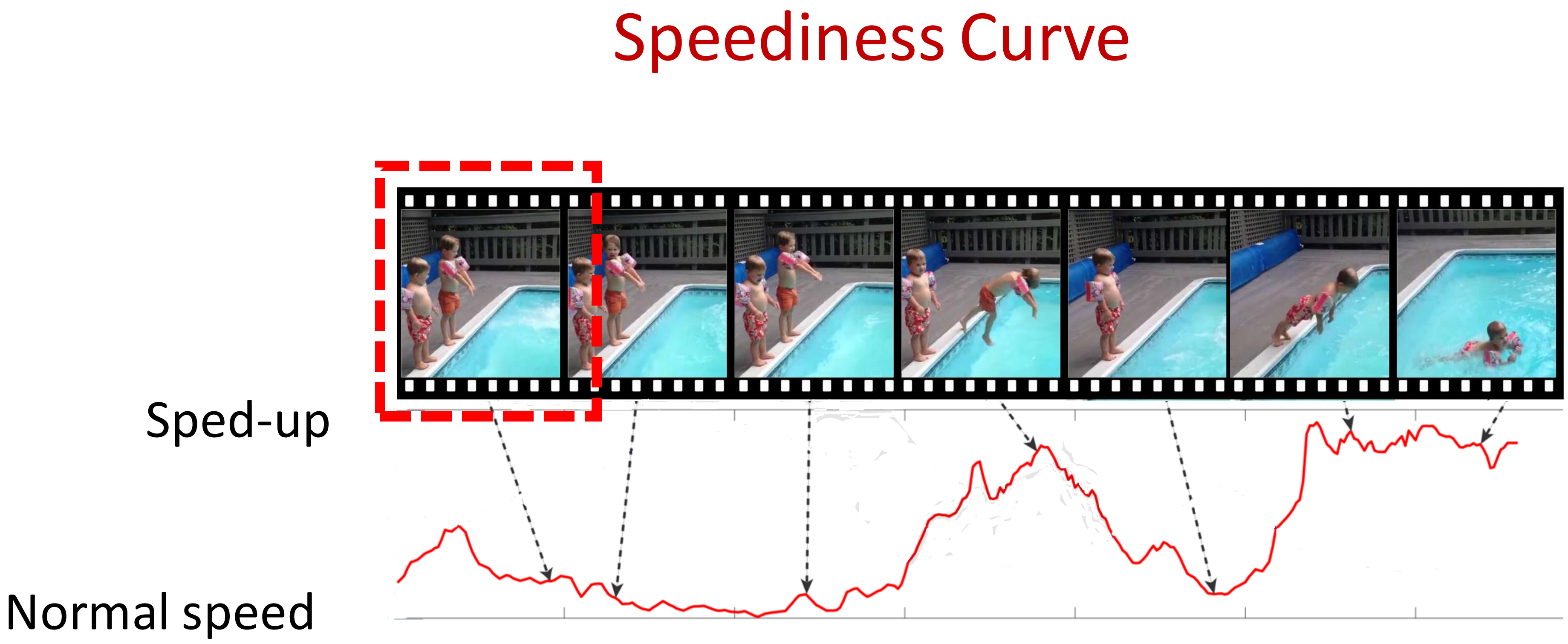
2x video (Interpolate to T Frames)

3x video (Interpolate to T Frames)

...

Nx video (Interpolate to T Frames)

From Speediness to Adaptive Speedup



From Speediness to Adaptive Speedup

Original 1x video

N videos of increasing speed

1x video Speediness Curve

2x video Speediness Curve

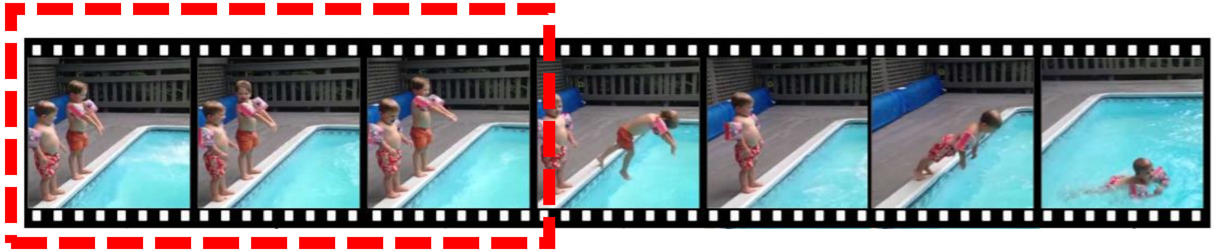
3x video Speediness Curve

...

Nx video Speediness Curve

From Speediness to Adaptive Speedup

Original 1x video



Low Speediness (for most speedup curves)

1x video Speediness Curve

2x video Speediness Curve

3x video Speediness Curve

...

Nx video Speediness Curve

N videos of increasing speed

From Speediness to Adaptive Speedup

Original 1x video

High Speediness (for most speedup curves)

N videos of increasing speed

1x video Speediness Curve

2x video Speediness Curve

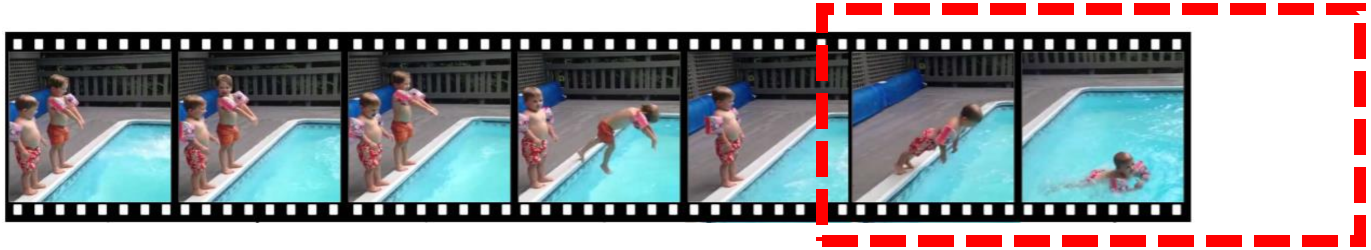
3x video Speediness Curve

...

Nx video Speediness Curve

From Speediness to Adaptive Speedup

Original 1x video



Medium Speediness (only some curves indicate speedup)

N videos of increasing speed

1x video Speediness Curve

2x video Speediness Curve

3x video Speediness Curve

...

Nx video Speediness Curve

From Speediness to Adaptive Speedup

Original 1x video

Speedup Vector V(t) =
Max of

1x binarized video Speediness Curve x1

2x binarized video Speediness Curve x2

3x binarized video Speediness Curve x3

...

Nx binarized video Speediness Curve xN

From Speediness to Adaptive Speedup

Original 1x video

Final step: Estimate a smoothly varying speedup curve (say for 2x)

$$\arg\min_S E_{\text{speed}}(S, V)$$

- S should be close to V(t) – our estimated Speedup Vector

From Speediness to Adaptive Speedup

Original 1x video

Final step: Estimate a smoothly varying speedup curve (say for 2x)

$$\arg\min_S E_{\text{speed}}(S, V) + \beta E_{\text{rate}}(S, R_o)$$

- S should be close to V(t) – our estimated Speedup Vector
- The total frame rate should be the desired frame rate (e.g 2x or 3x)

From Speediness to Adaptive Speedup

Original 1x video

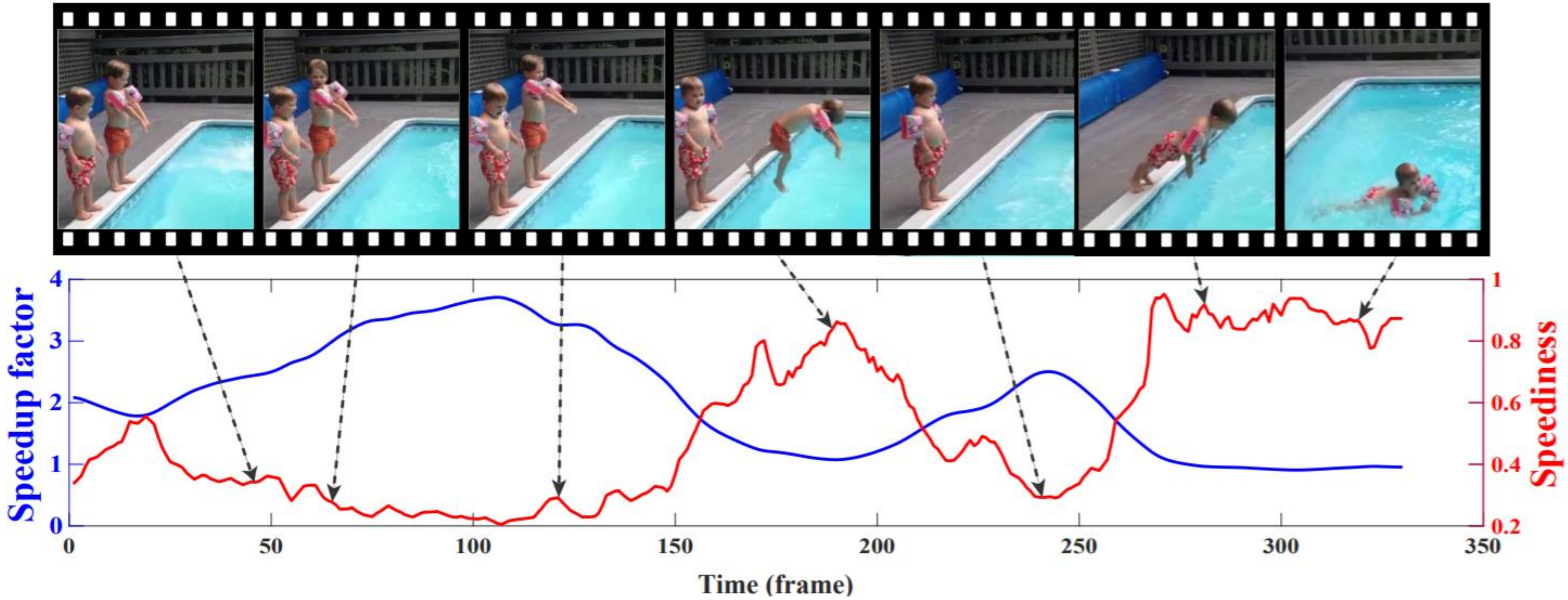
Final step: Estimate a smoothly varying speedup curve

$$\arg\min_S E_{\text{speed}}(S, V) + \beta E_{\text{rate}}(S, R_o) + \alpha E_{\text{smooth}}(S')$$

- S should be close to V(t) – our estimated Speedup Vector
- The total frame rate should be the desired frame rate (e.g 2x or 3x)
- Smoothness regularizer using the first derivatives S’

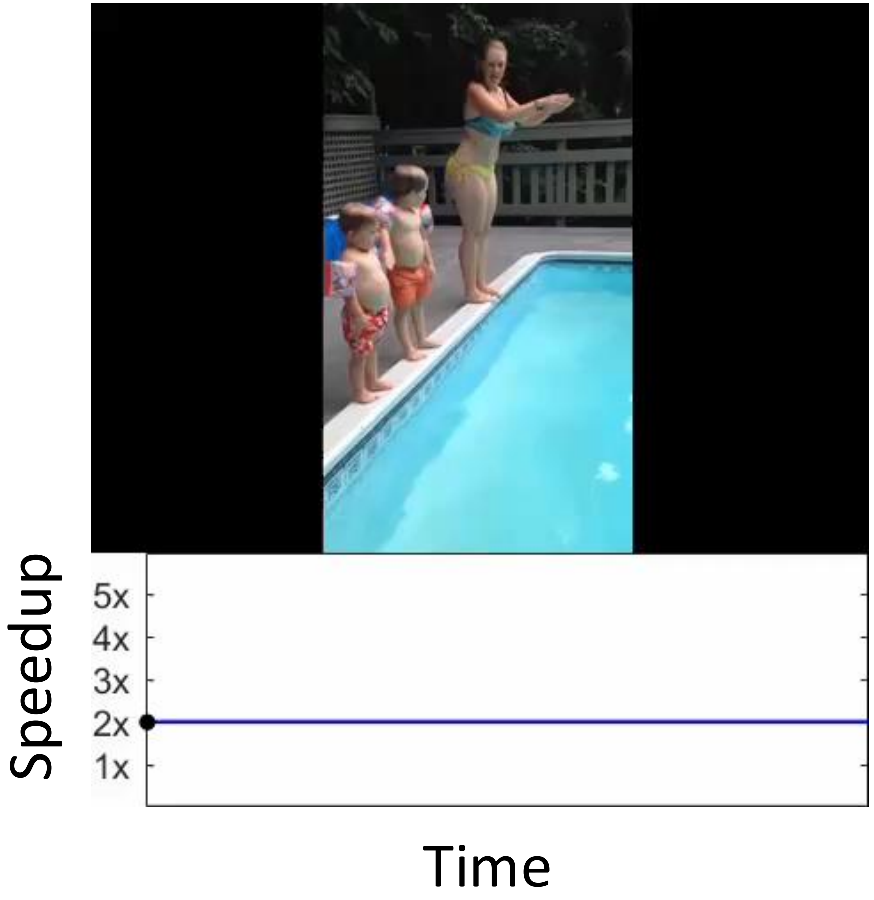
From Speediness to Adaptive Speedup

2x final “speediness curve” (blue):



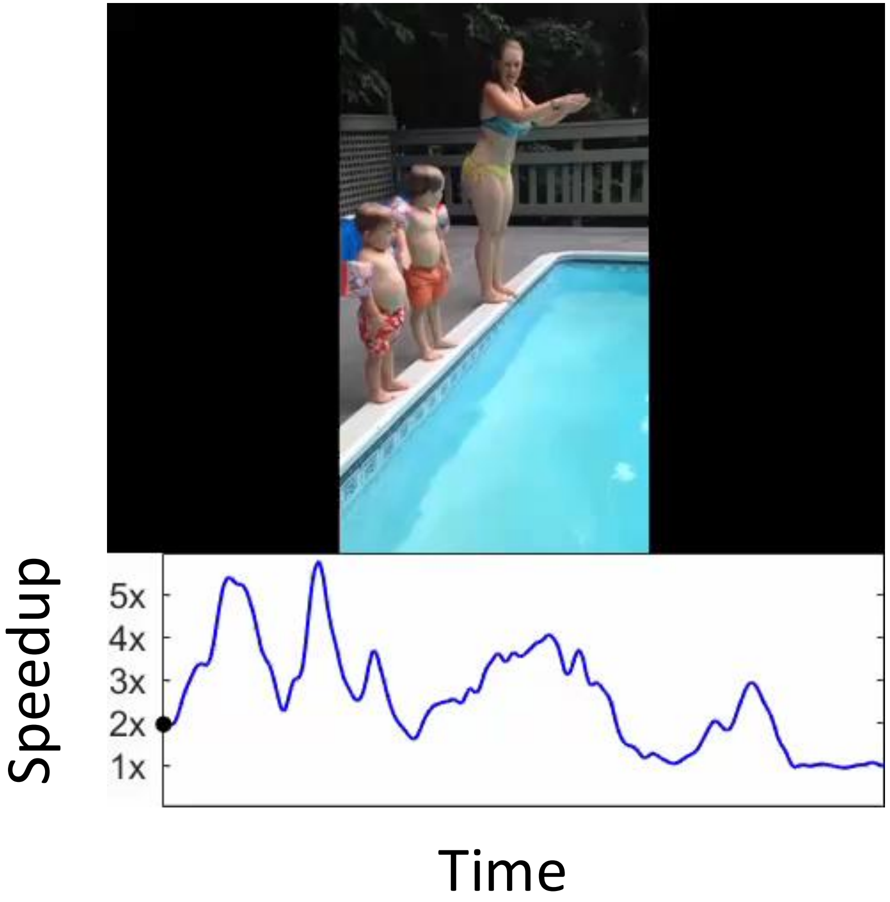
Adaptive video speedup

Total time = $\frac{1}{2}$ input time



Uniform Speedup

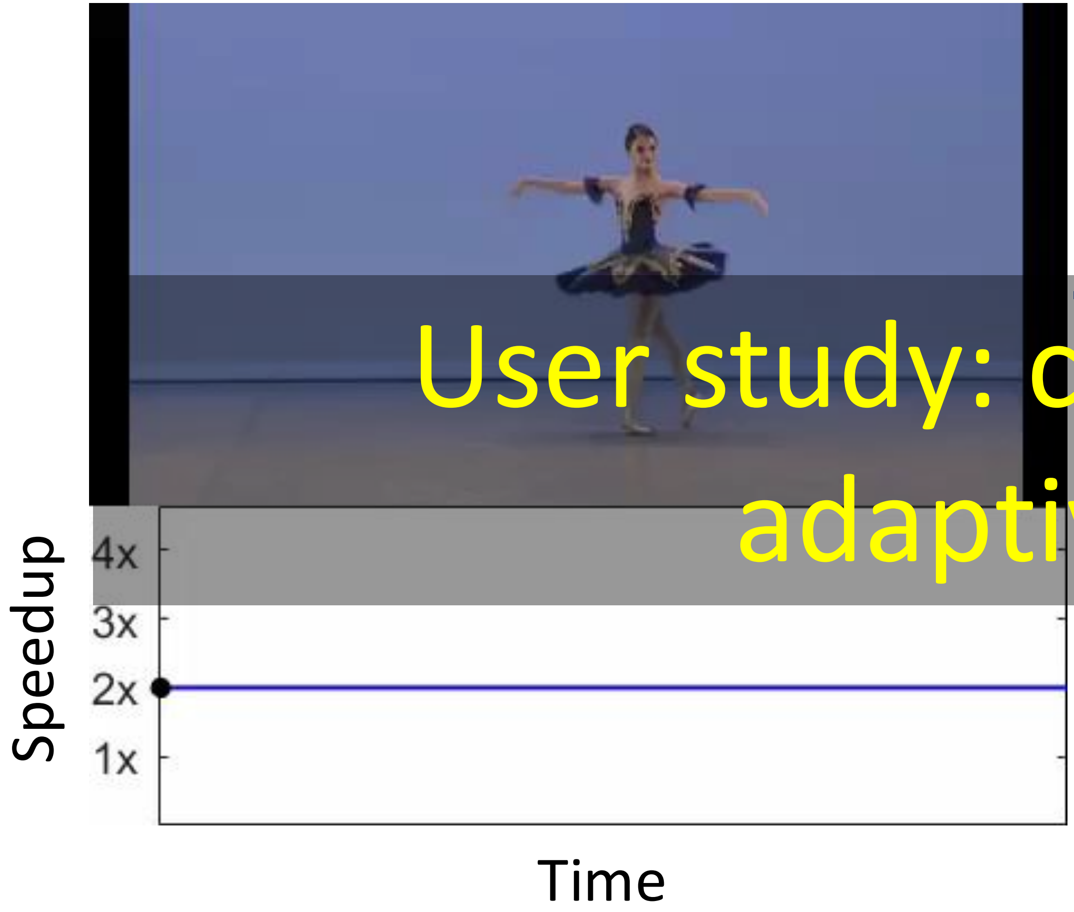
Total time = $\frac{1}{2}$ input time



Adaptive Speedup (ours)

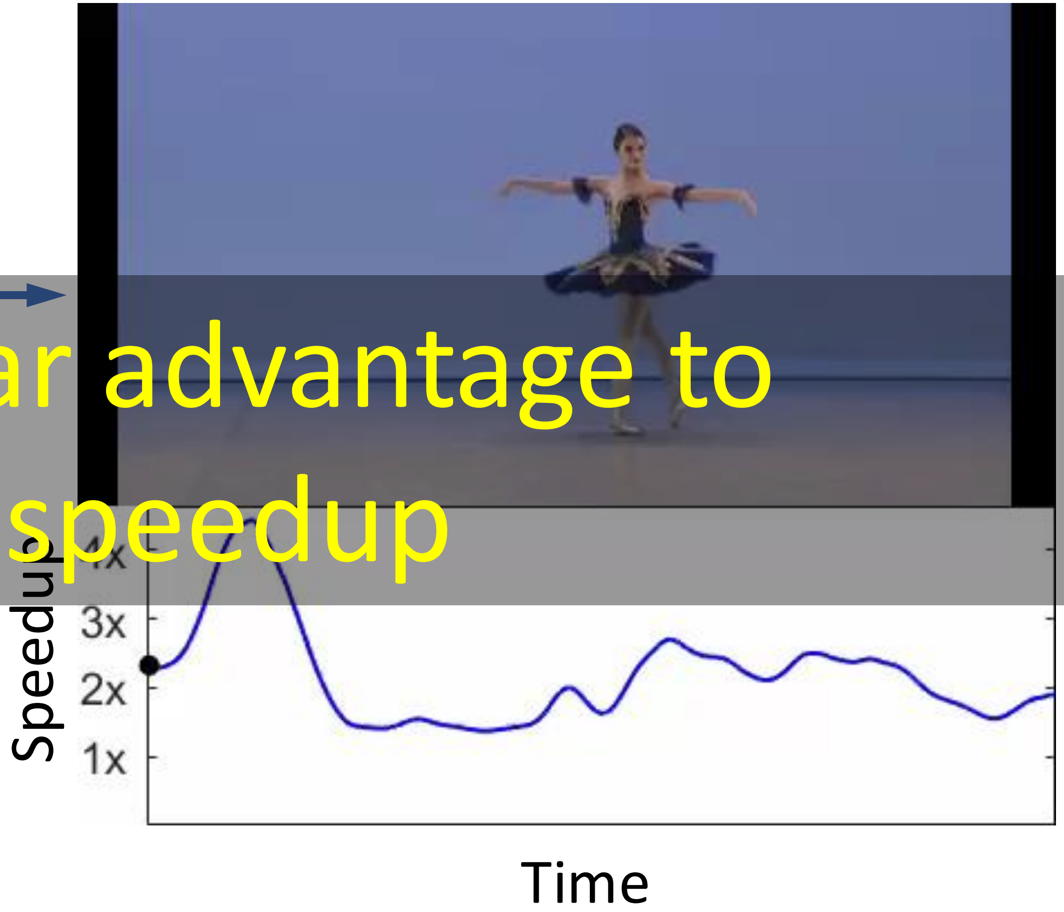
Adaptive video speedup

Total time = $\frac{1}{2}$ input time



Uniform Speedup

Total time = $\frac{1}{2}$ input time

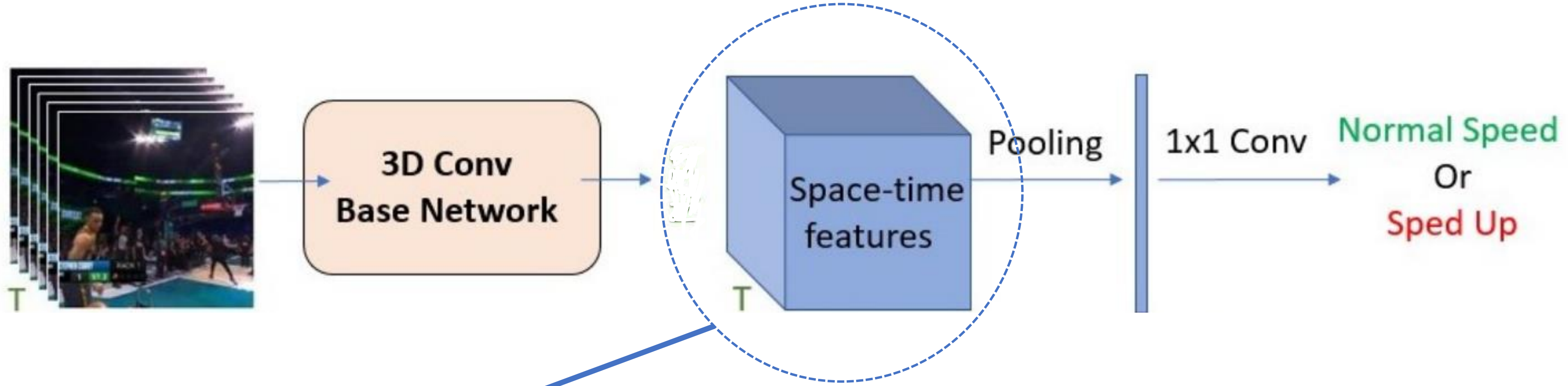


Adaptive Speedup (ours)

User study: clear advantage to adaptive speedup

Other self supervised tasks

Train SpeedNet

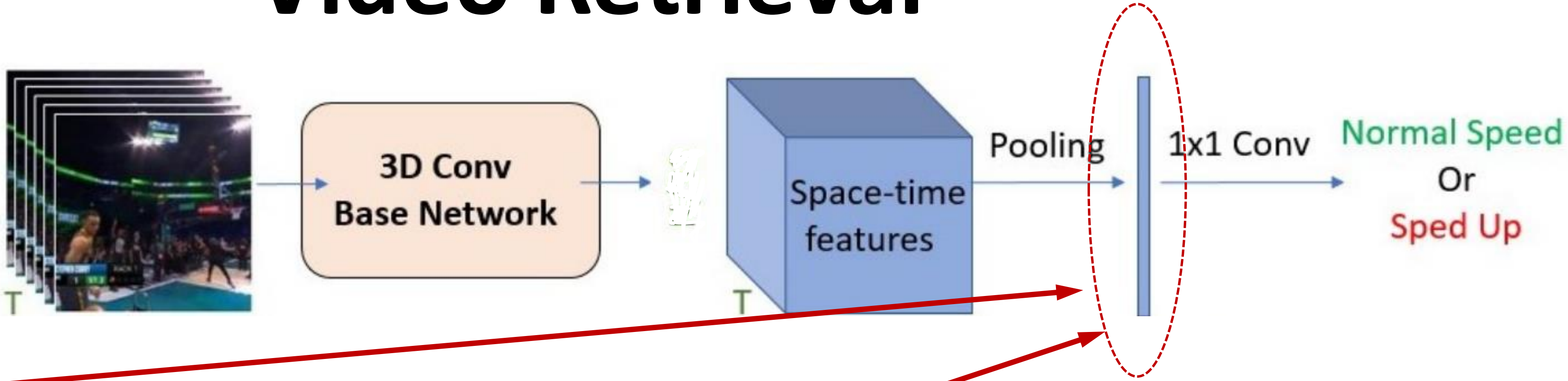


Self Supervised Action Recognition

Initialization		Supervised accuracy	
Method	**Architecture**	**UCF101**	**HMDB51**
Random init	S3D-G	73.8	46.4
ImageNet inflated	S3D-G	86.6	57.7
Kinetics supervised	S3D-G	96.8	74.5
CubicPuzzle [19]	3D-ResNet18	65.8	33.7
Order [40]	R(2+1)D	72.4	30.9
DPC [13]	3D-ResNet34	75.7	35.7
AoT [38]	T-CAM	79.4	-
SpeedNet (Ours)	S3D-G	**81.1**	**48.8**
Random init	I3D	47.9	29.6
SpeedNet (Ours)	I3D	66.7	43.7

Other self supervised tasks: Video Retrieval

Train SpeedNet

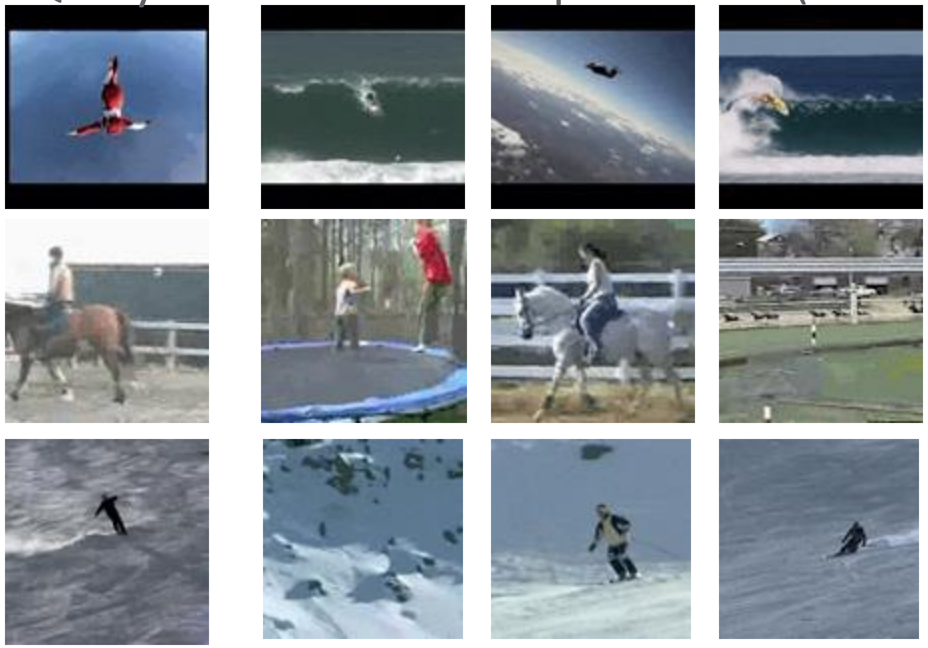


Query

Retrieved top-3 results (Within)

Query

Retrieved top-3 results (Across)



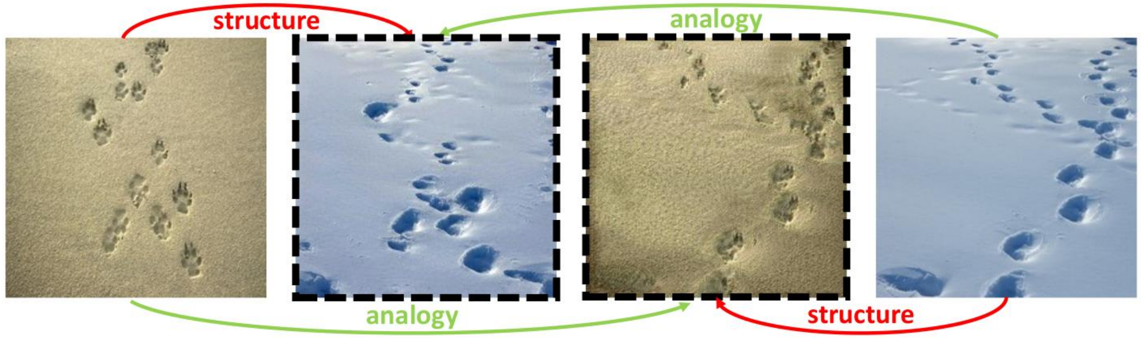
“Memory Eleven”: An artistic video by Bill Newsinger:
https://www.youtube.com/watch?v=djylS0Wi_lo

Spatio-Temporal Visualizations

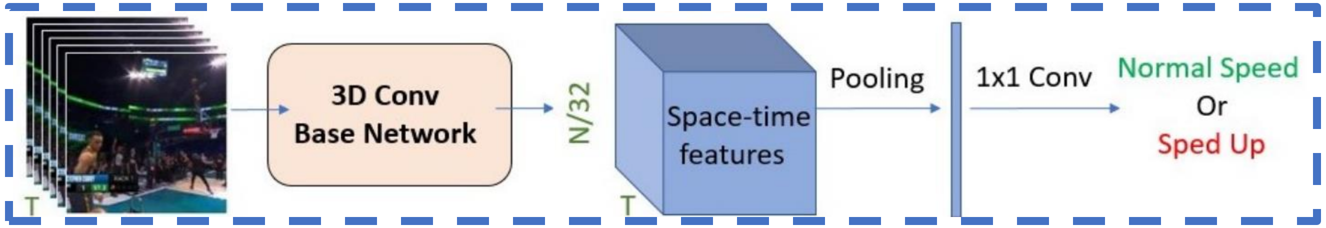
blue/green =
normal speed

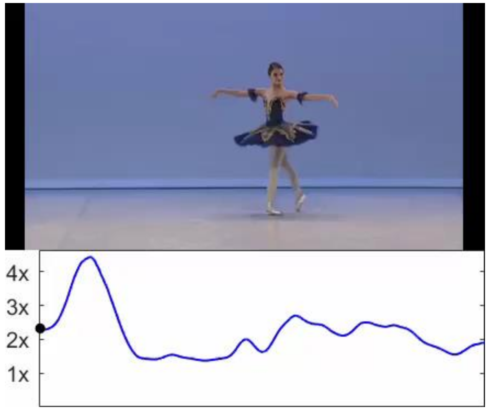
yellow/orange =
slowed down

Part I: Semantic Manipulation of **Images**



Part II: Semantic Manipulation of Videos

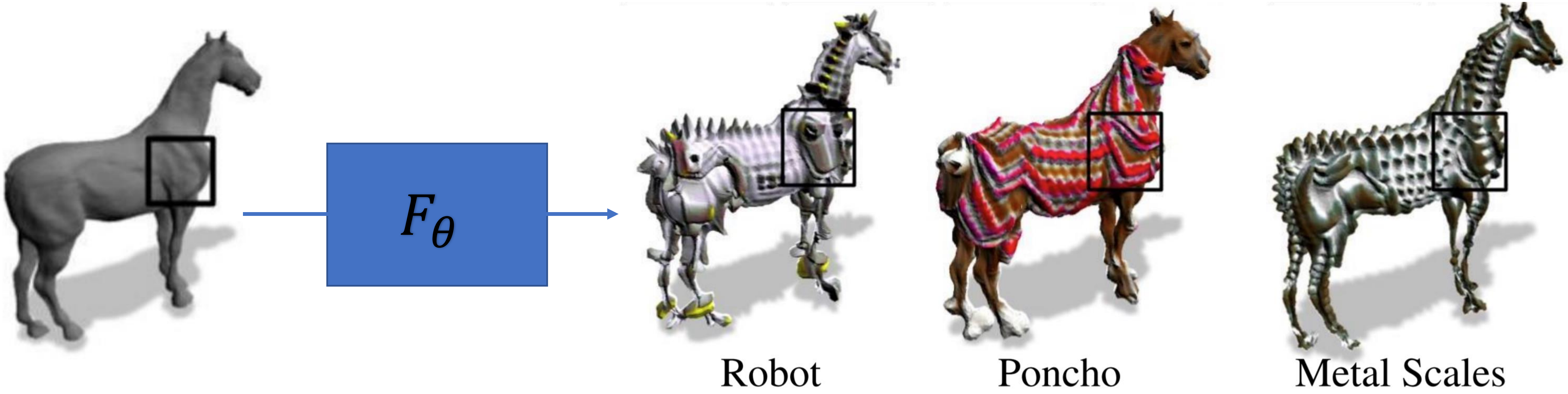




Part III: Semantic Manipulation of **3D Objects**

Text2Mesh: Text-Driven Stylization for Meshes

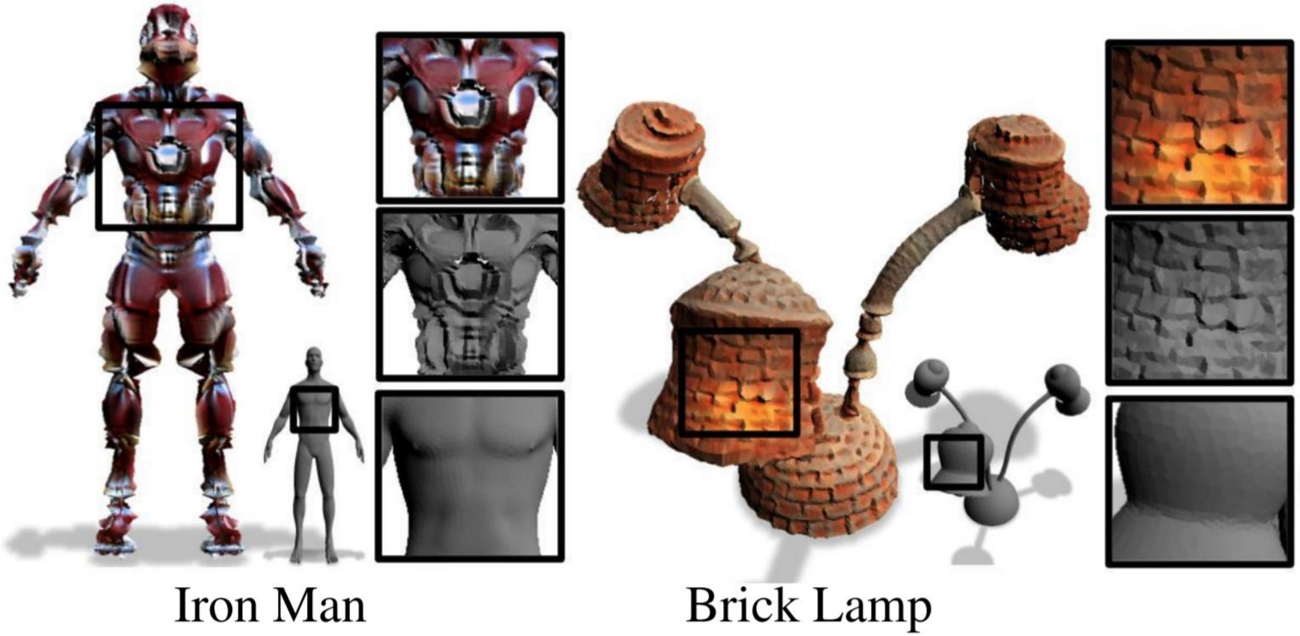
O. Michel, R Bar-On, R Liu, **S. Benaim**, R. Hanocka. In Submission.



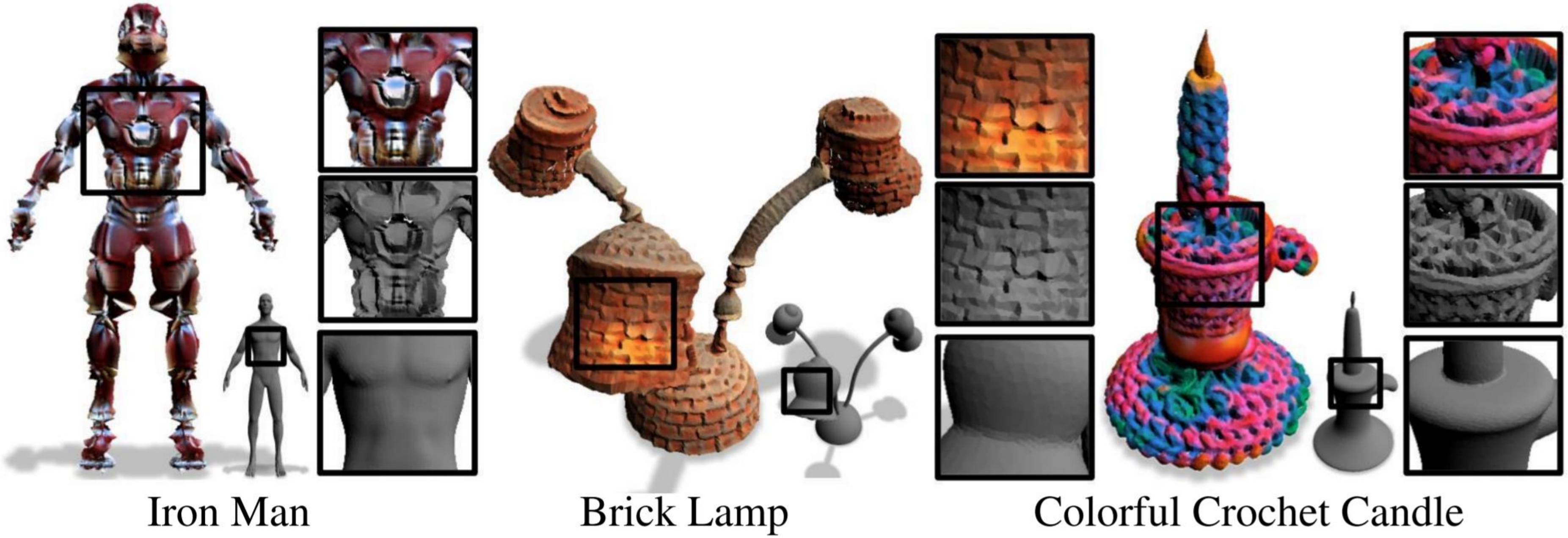
Part Aware Global Semantics

Iron Man

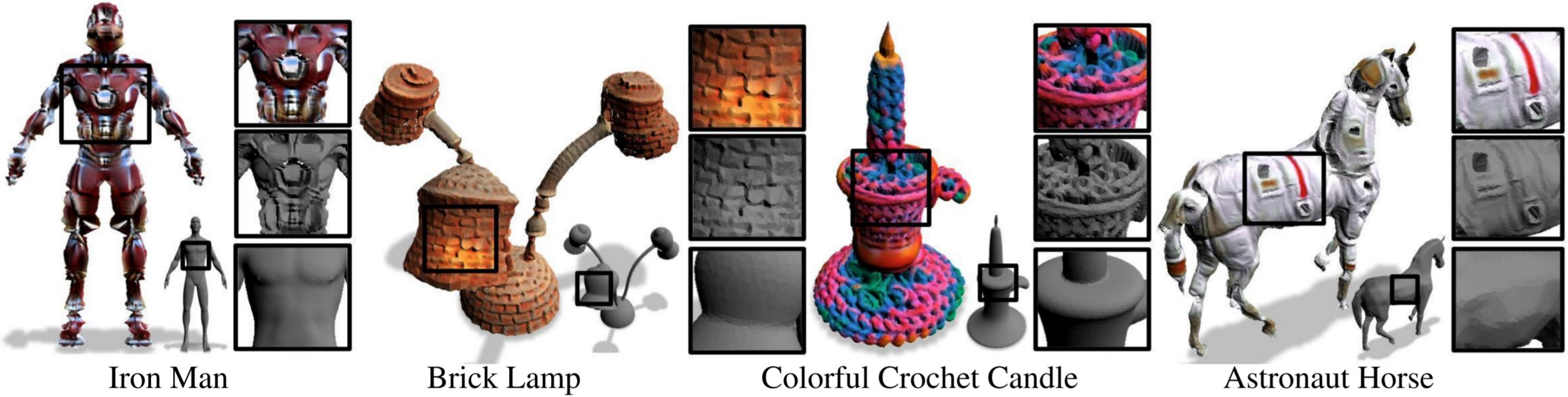
Structured Textures with Lighting



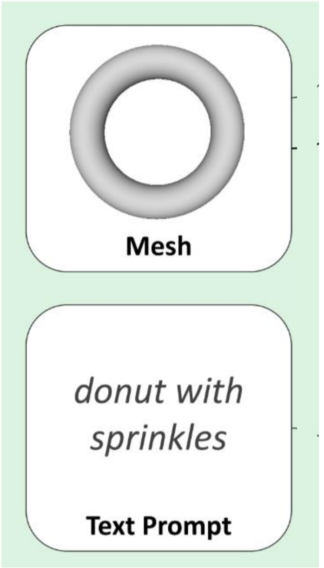
Variety of Textures and Materials



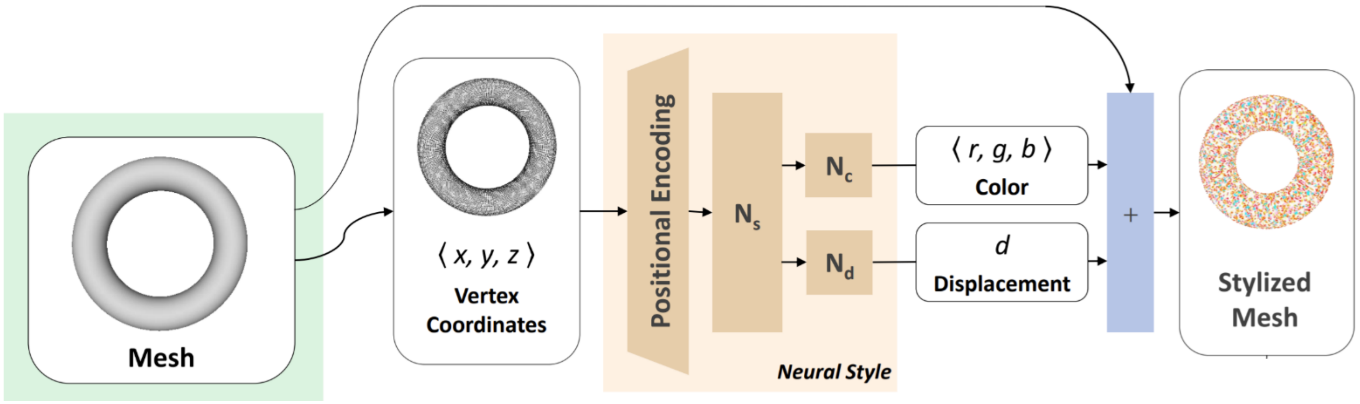
Out of Domain Generations



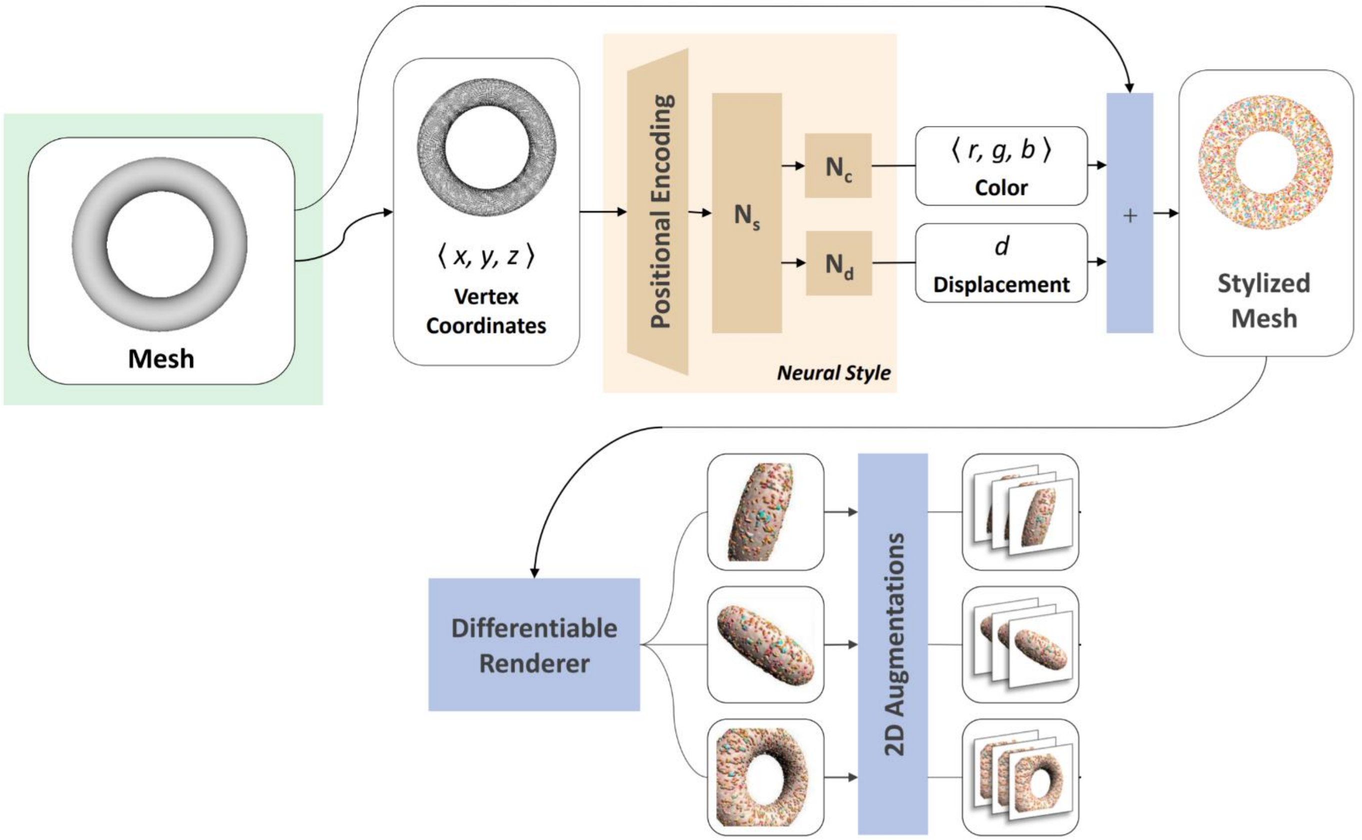
Overview: Input



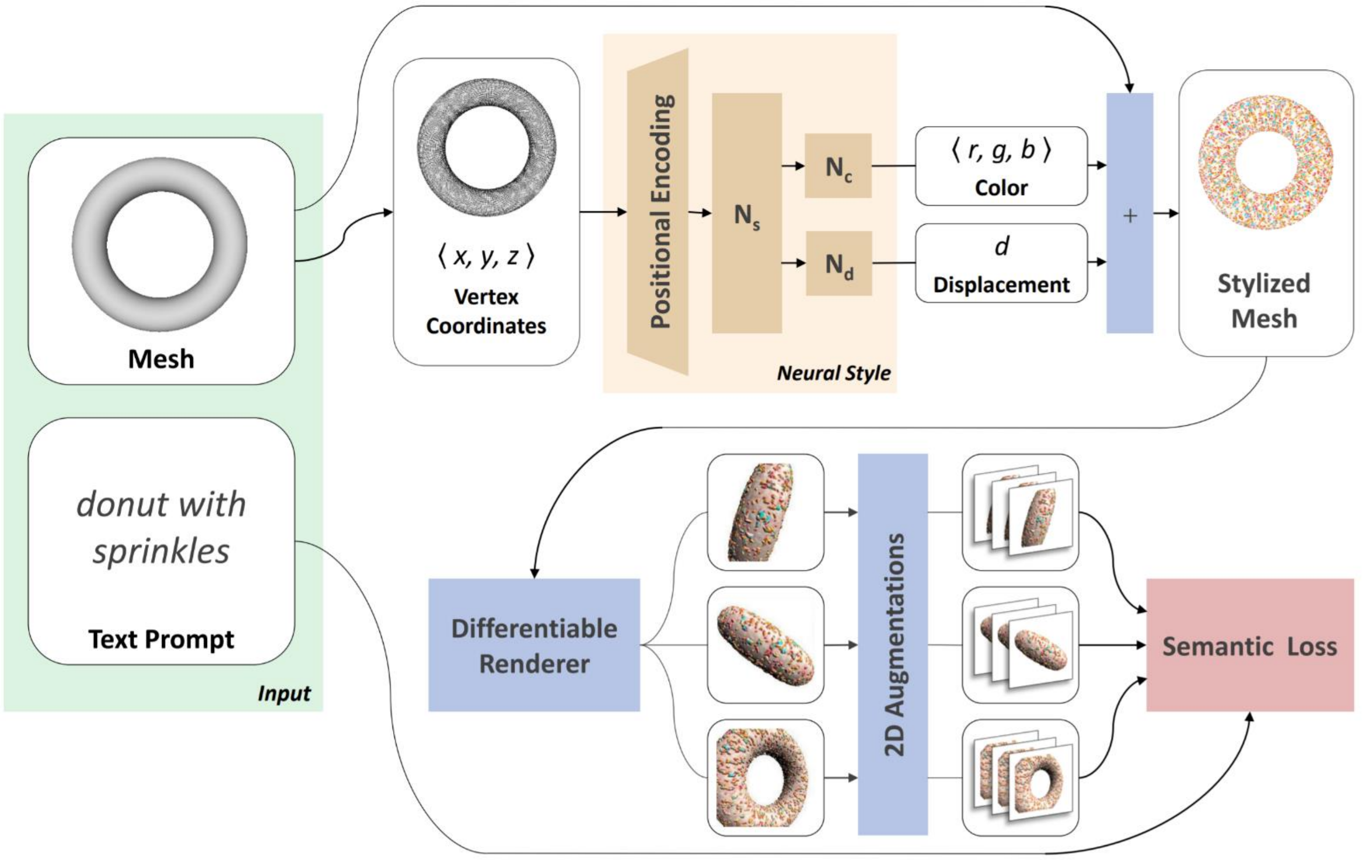
Overview: Neural Style Field



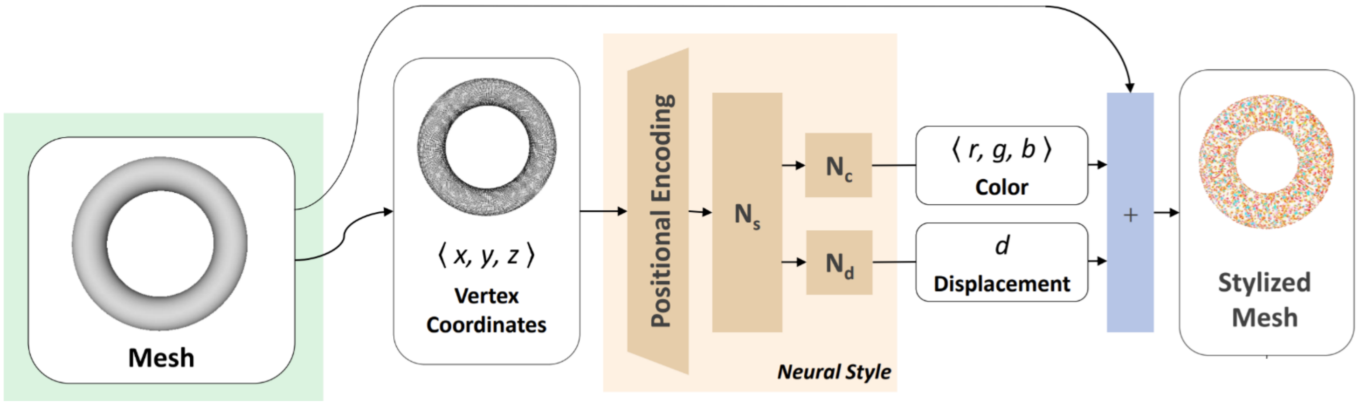
Overview: Neural Rendering and Augmentations



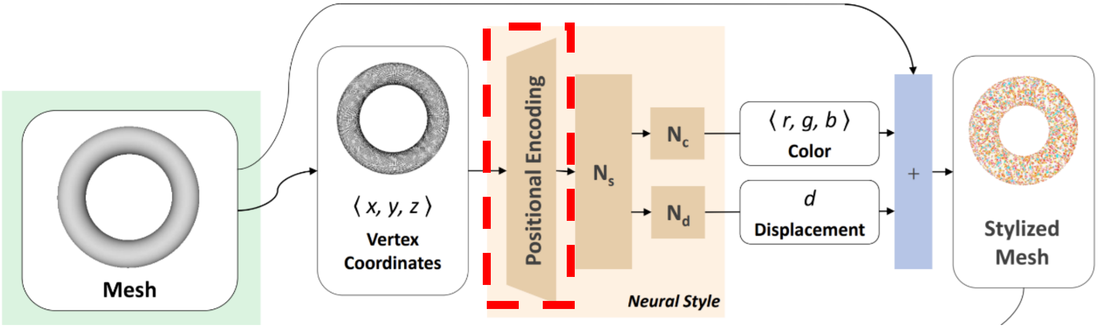
Overview: CLIP Based Semantic Loss



Neural Style Field



Positional Encoding



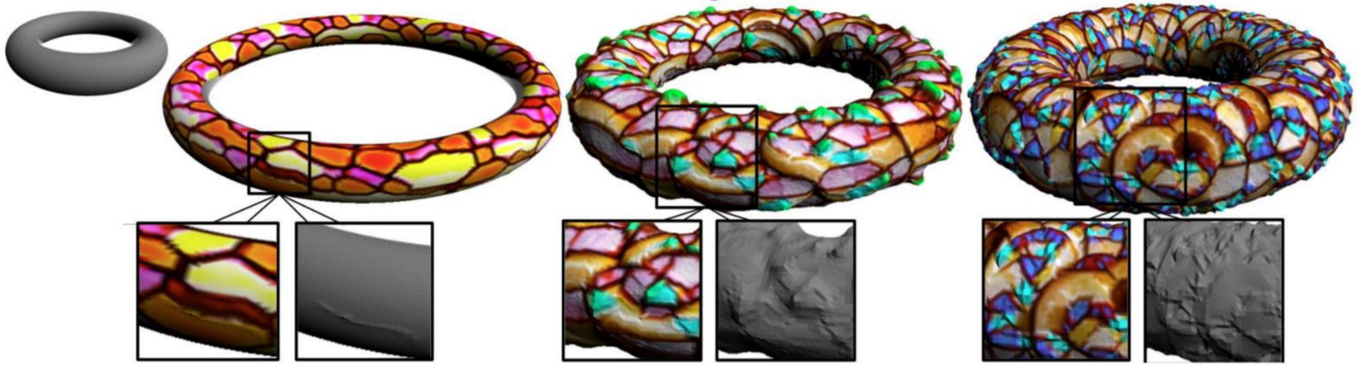
- Frequency based encoding:

$$\gamma(p) = \left[\cos\left(2\pi \mathbf{B} p\right), \sin\left(2\pi \mathbf{B} p\right)\right]^{\mathrm{T}}$$

$\boldsymbol{B} \in R^{n\times 3}$ randomly drawn from $N(0, \sigma)$

- σ is a hyperparameter which controls the output frequency:

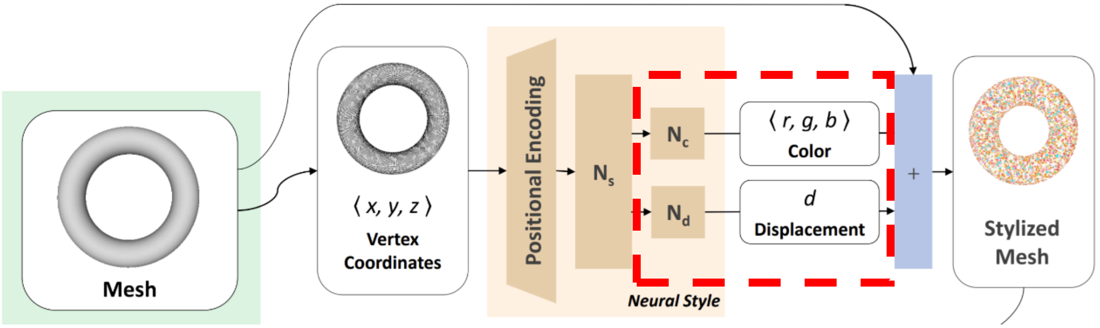
‘Stained glass donut’

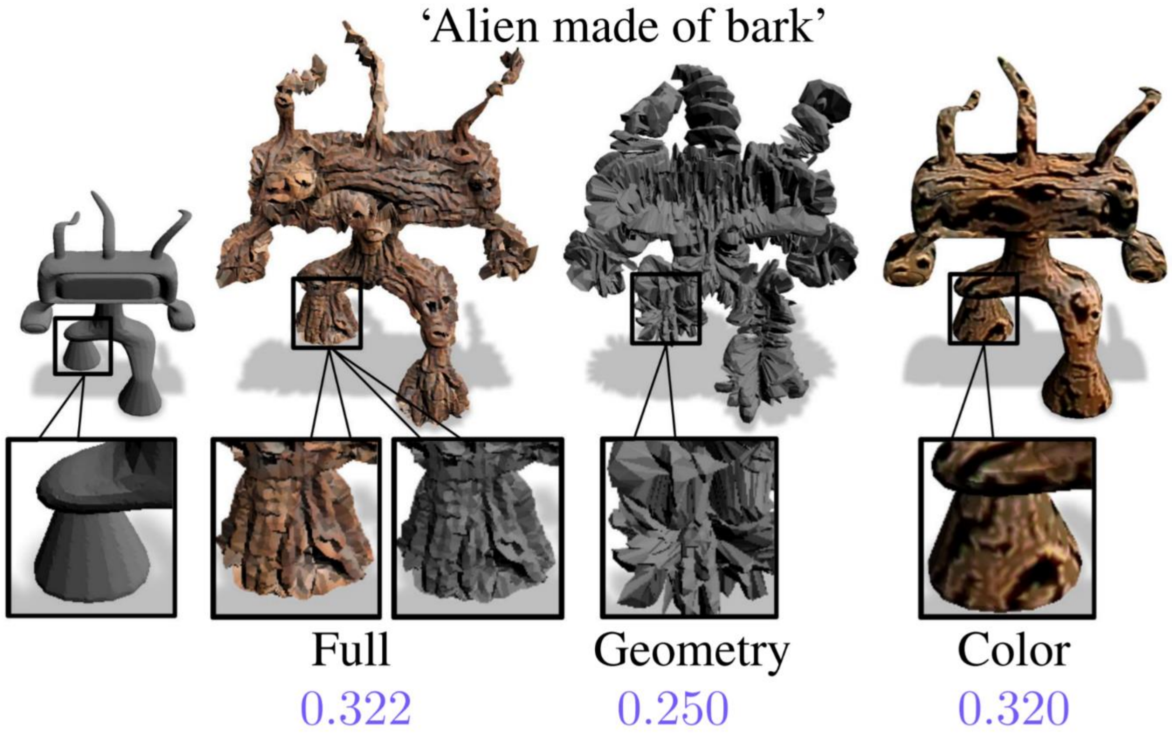


$\sigma = 3$ $\sigma = 5$ $\sigma = 8$

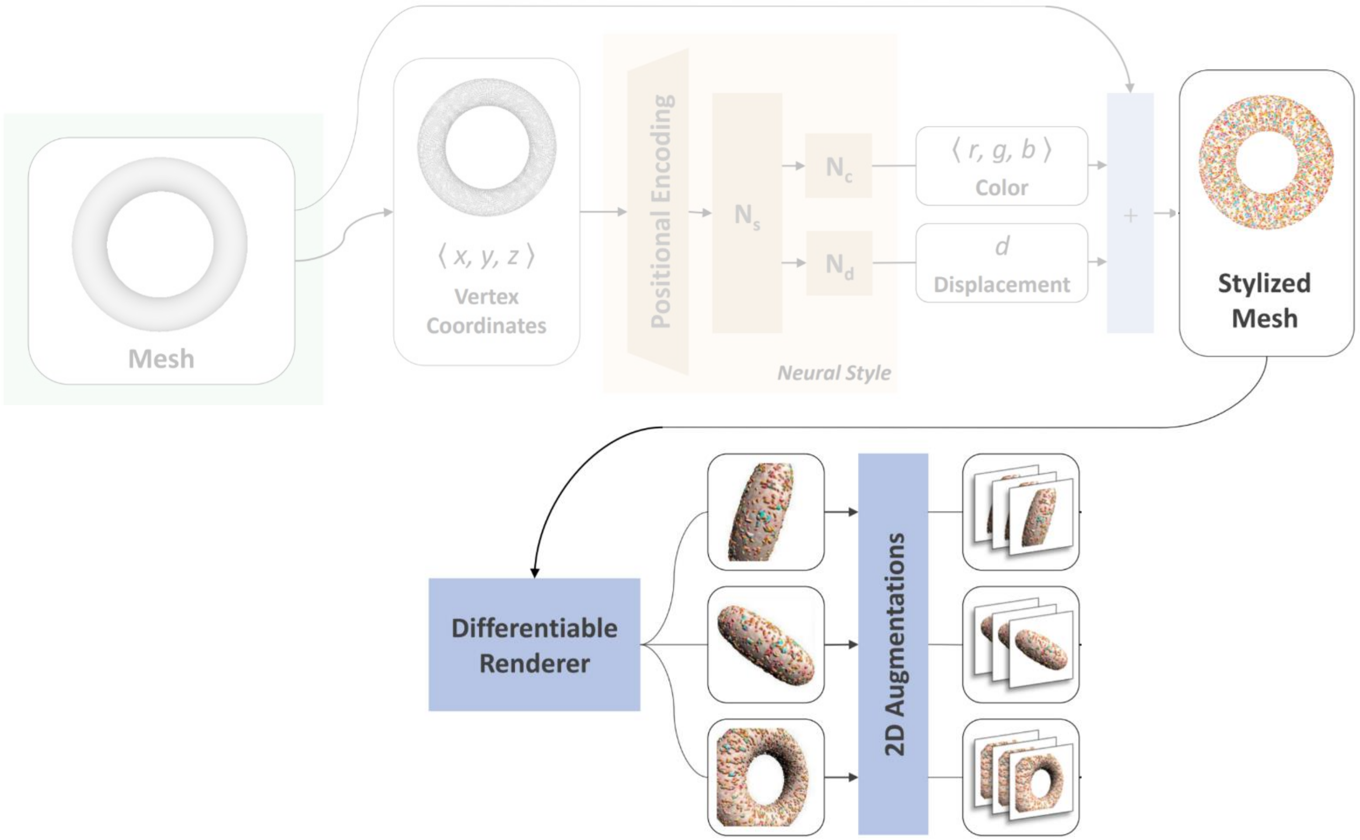
M. Tancik et al., “Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains”. NeurIPS 2020.

Geometry and Color



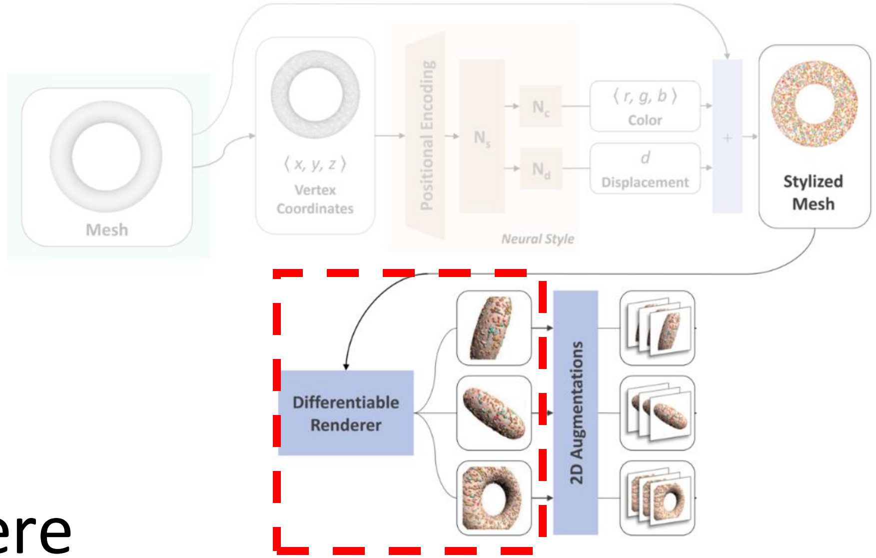


Neural Rendering and Augmentations



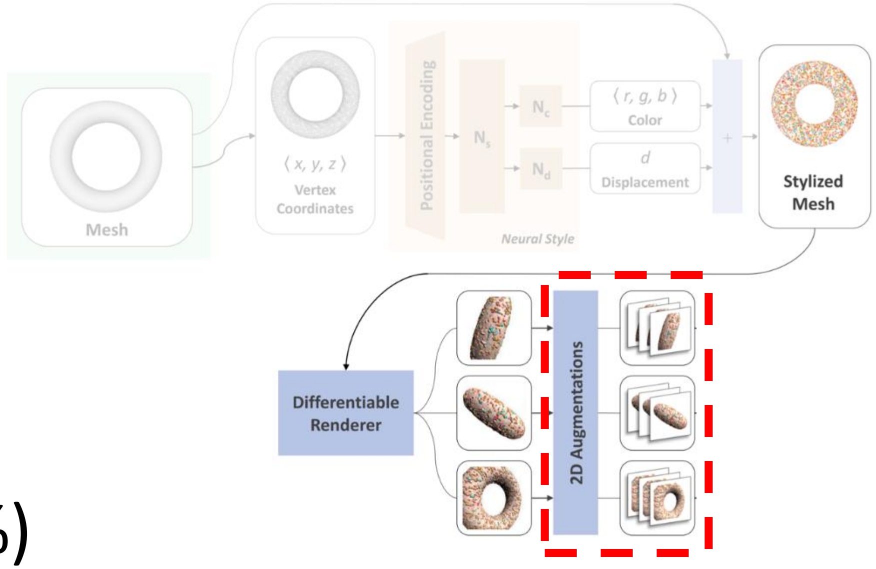
How are views selected?

- *Anchor view v:* view with high similarity to target in CLIP space
- Many such views exist!
- Sample random views from a $N(v, \sigma)$ where $\sigma = \pi/4$.
- 5 views are sufficient.

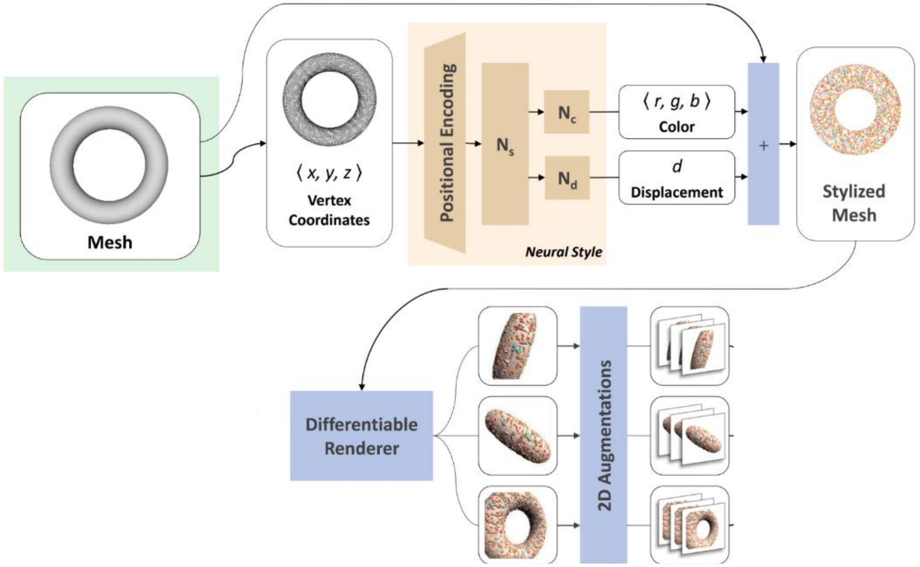


Augmentations are crucial!

- *Global Augmentations:*
  Random Perspective
- *Local Augmentations:*
  Random Perspective + Random Crop (10%)

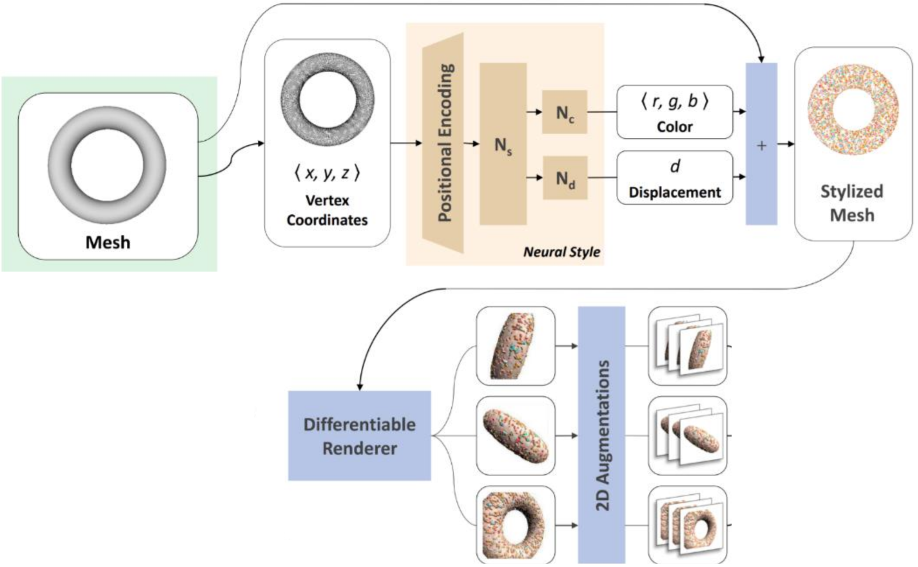


Importance of Components



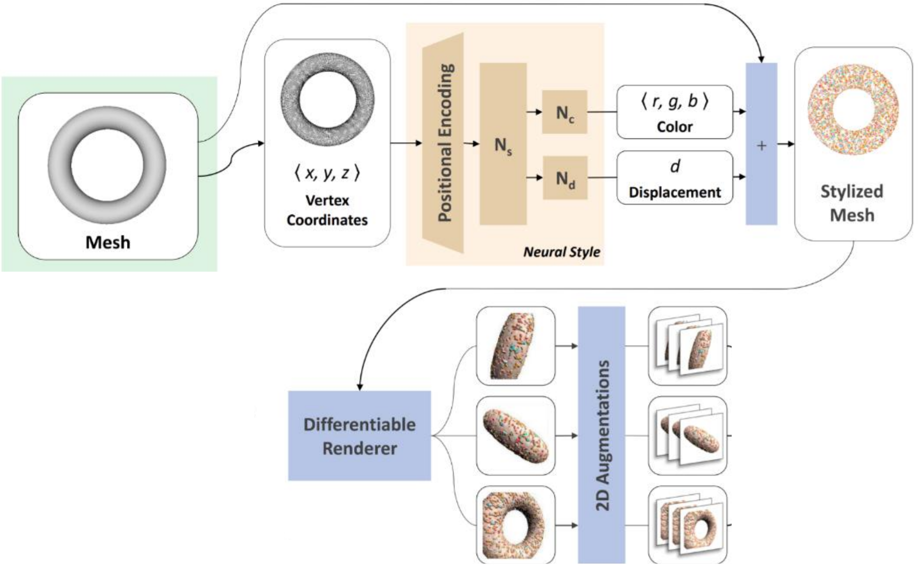
'Candle made of bark'

Importance of Components



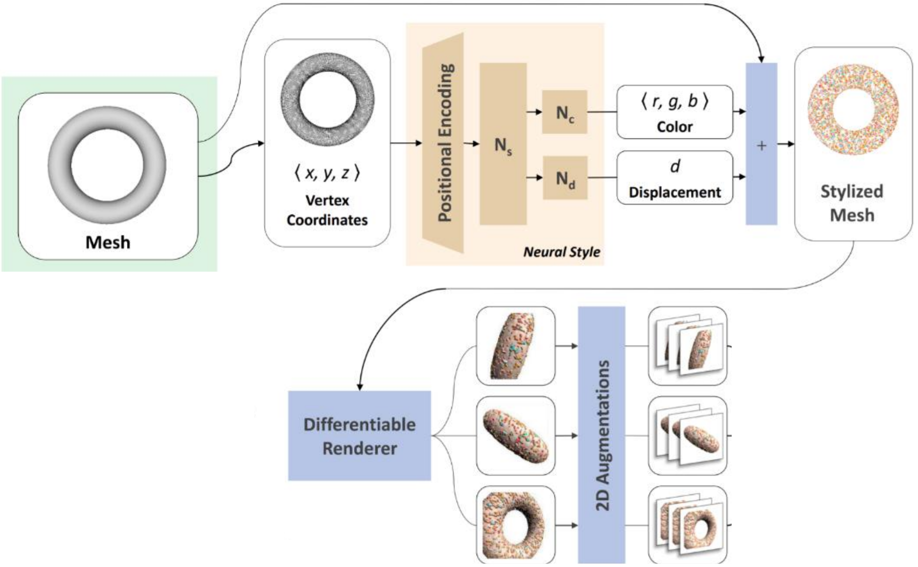
'Candle made of bark'

Importance of Components

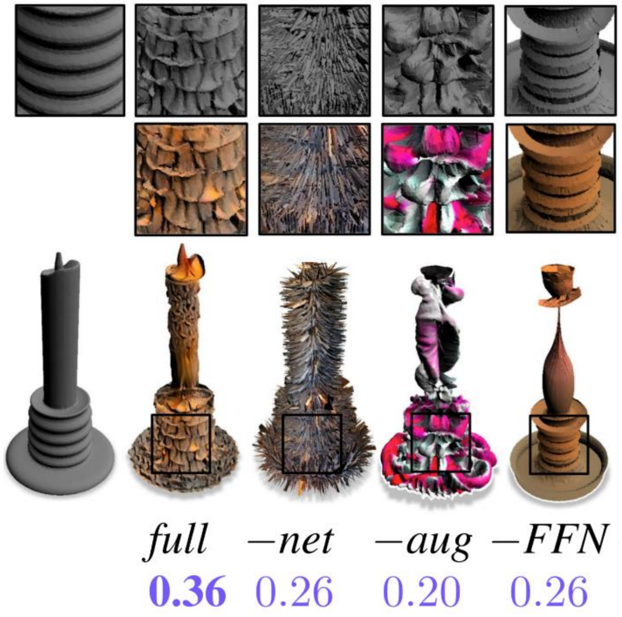


'Candle made of bark'

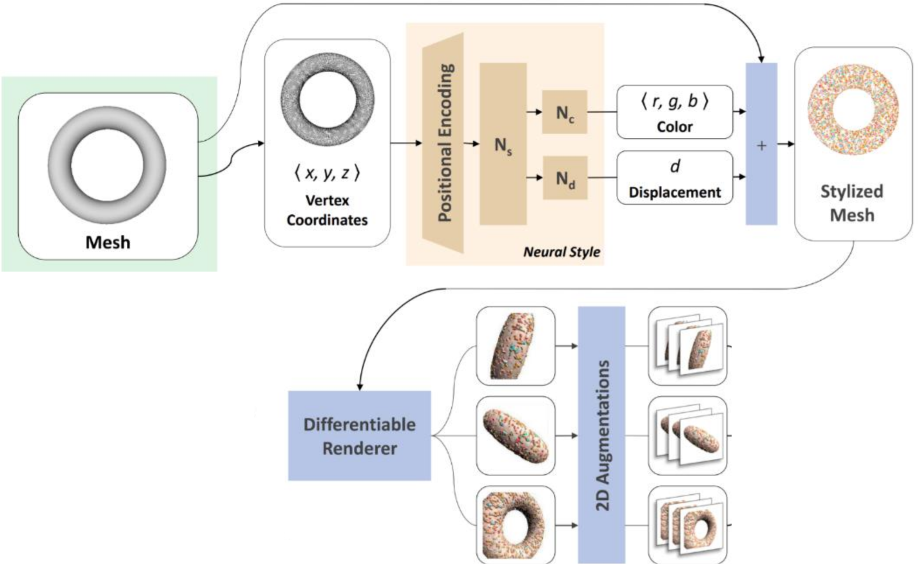
Importance of Components



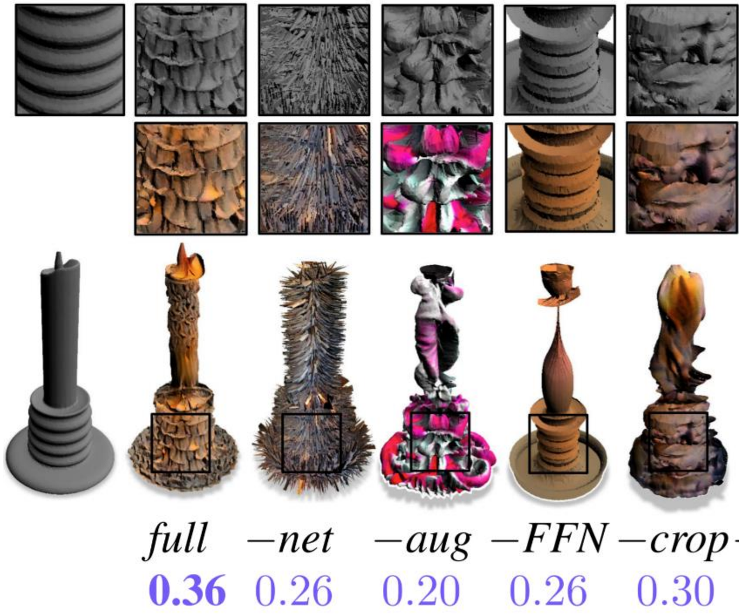
‘Candle made of bark’



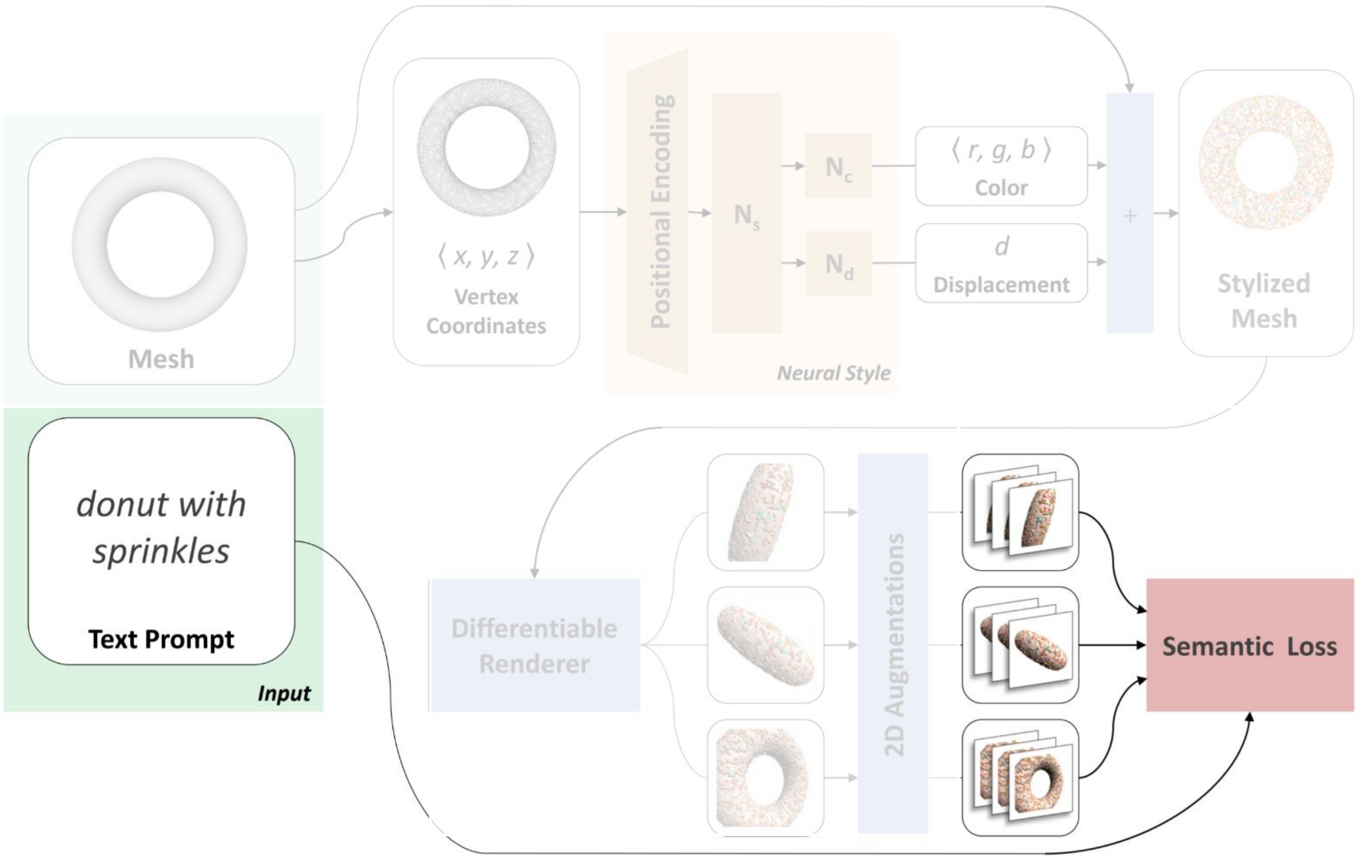
Importance of Components



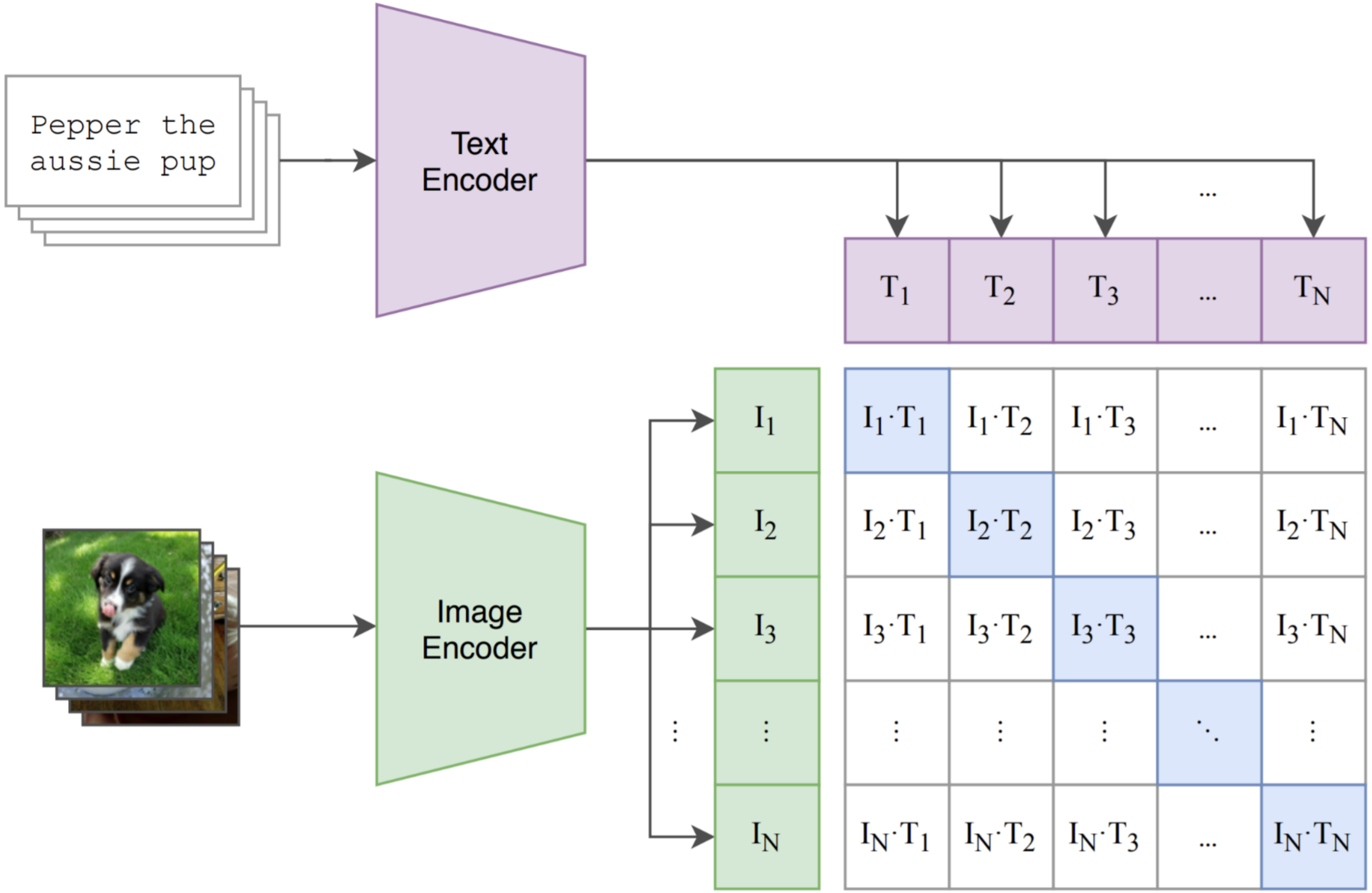
'Candle made of bark'



CLIP Based Semantic Loss

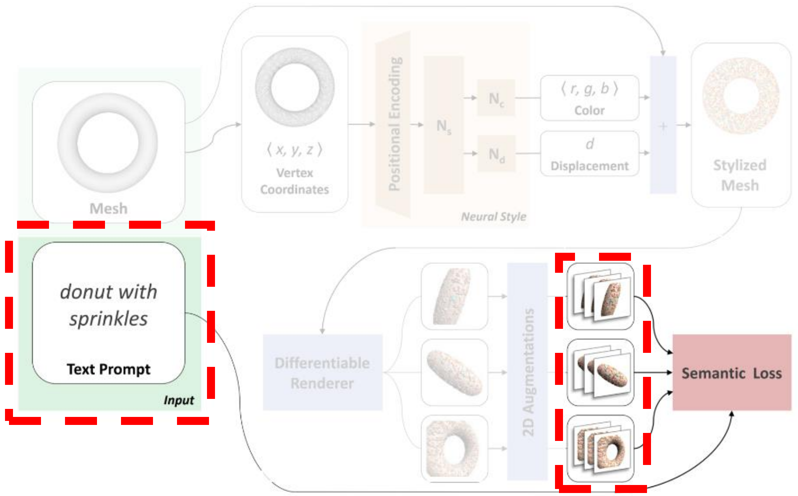


What is CLIP?



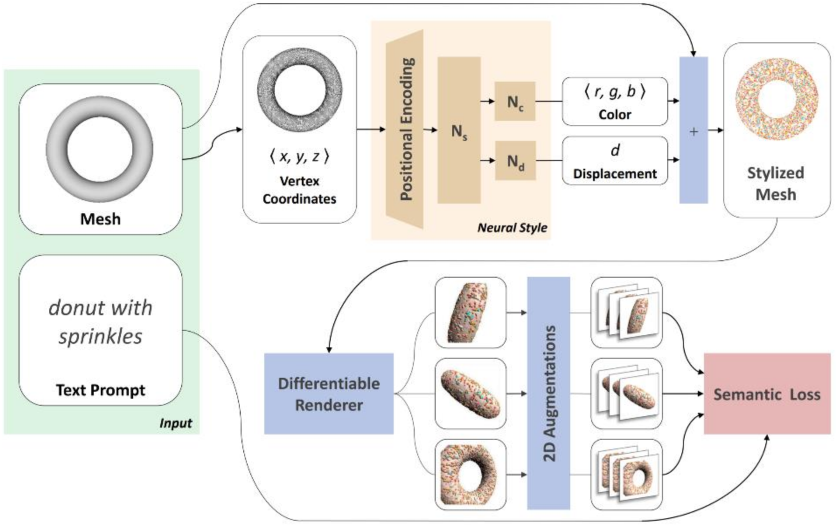
A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision". ICML 2021.

CLIP Based Semantic Loss



- Embed all augmented views and average to get S
- Embed text prompt to get T
- Maximize cosine similarity between S and T (Loss)

Important Advantages

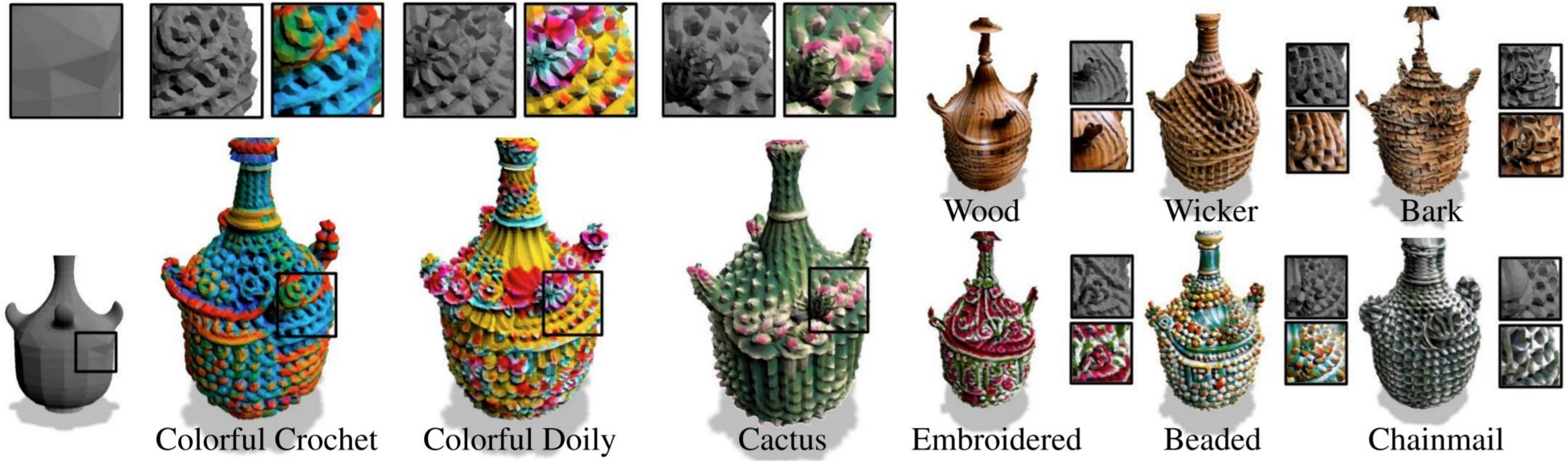


- No GAN or 3D Dataset needed – Only CLIP.
  And so, our method is zero-shot!
- Arbitrarily high resolutions can be rendered.
  Triangulation of the mesh can be arbitrarily dense.
- Disentanglement into an explicit mesh *content* and an *implicit* neural style field.
- In-the-wild meshes, arbitrary styles. Out-of-domain stylizations.

Results

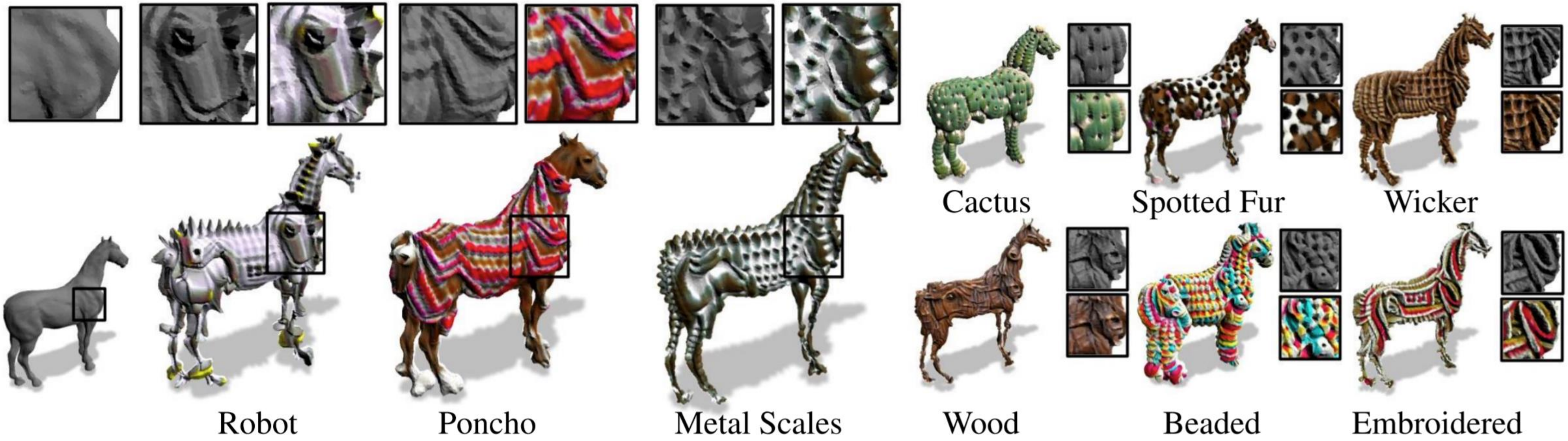
Stained Glass
Fur
Burgundy Crochet
Cactus
Bark
Brick
Wood
Chainmail
Wicker

Results

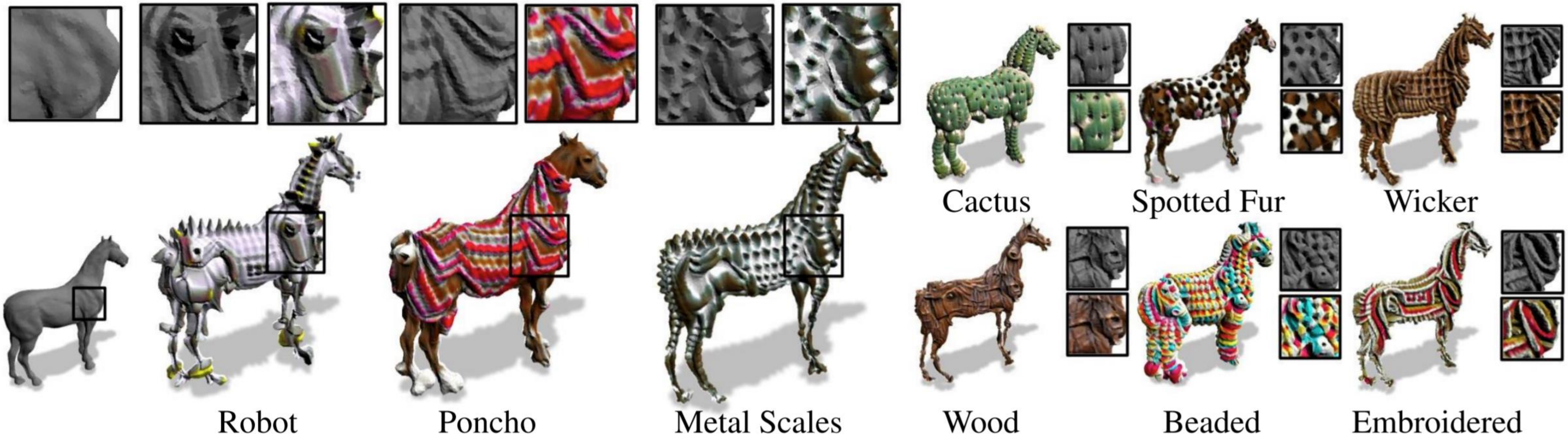


Wood
Wicker
Bark
Colorful Crochet
Colorful Doily
Cactus
Embroidered
Beaded
Chainmail

Results


Cactus
Spotted Fur
Wicker
Robot
Poncho
Metal Scales
Wood
Beaded
Embroidered

Results


Robot
Poncho
Metal Scales
Cactus
Spotted Fur
Wicker
Wood
Beaded
Embroidered

Humans

Input
Steve Jobs
Yeti
Astronaut
Buzz Lightyear
Ninja
Lawyer
Messi
Batman
Hulk

Increasing Granularity of Text

"Lamp"

Increasing Granularity of Text

"Luxo lamp"

Increasing Granularity of Text

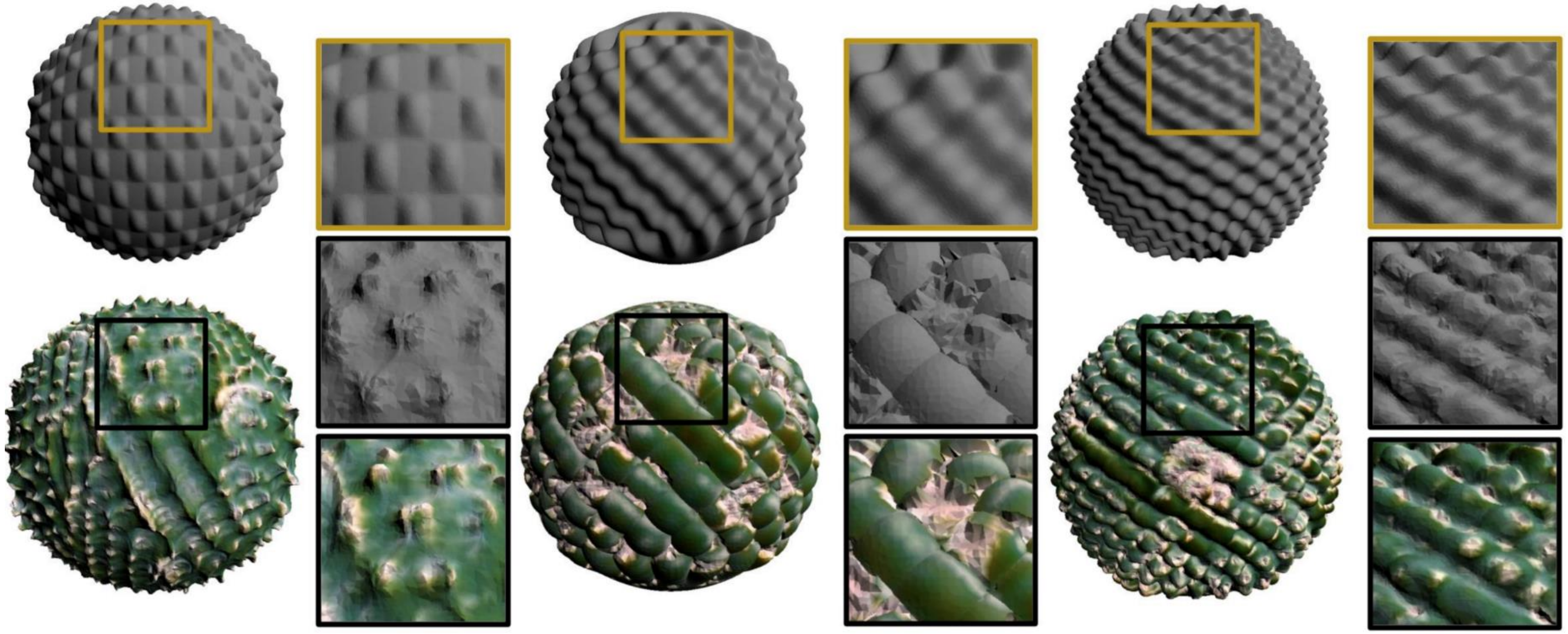
"Blue steel luxo lamp"

Increasing Granularity of Text

"Blue steel luxo lamp with corrugated metal"

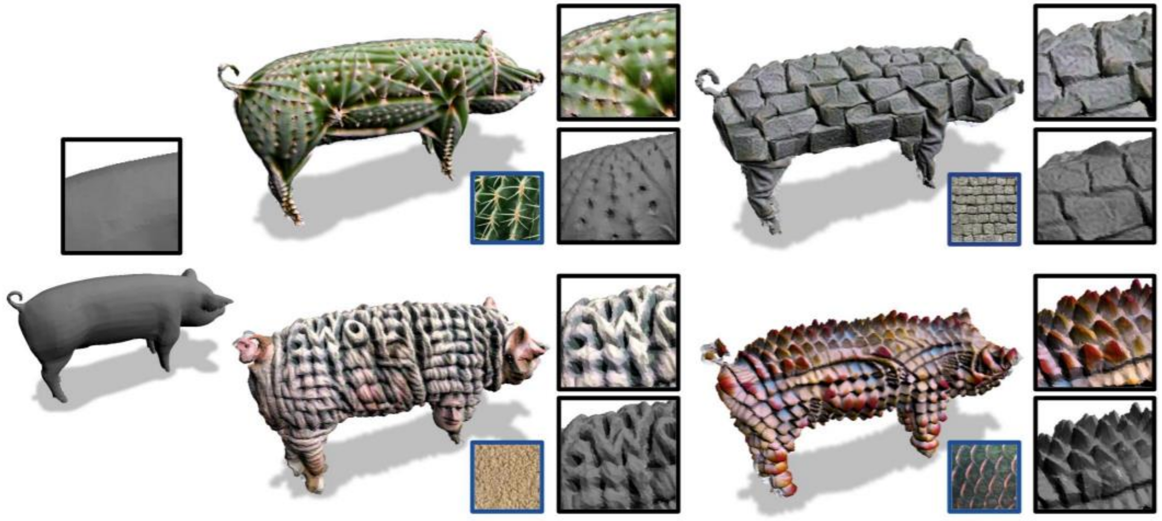
Increasing Mesh Granularity

"Cactus"



Different Target Modality

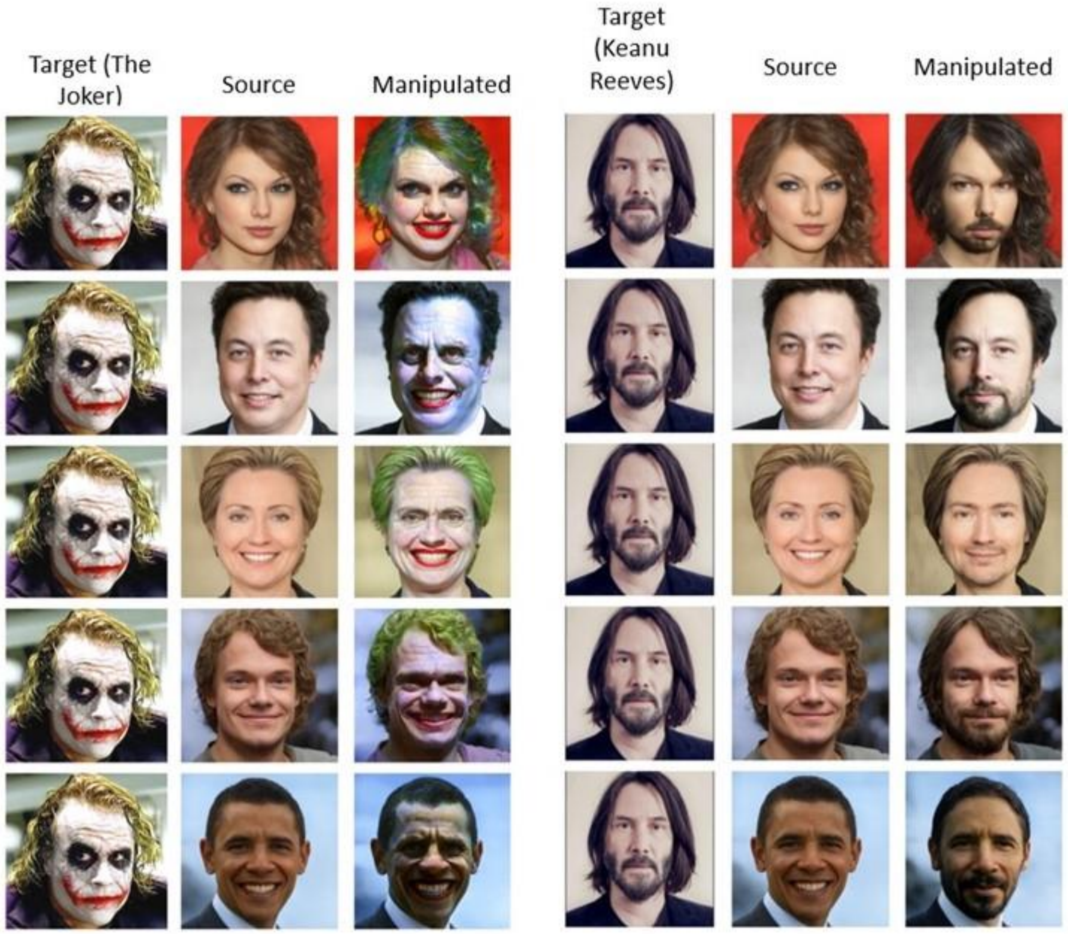
Image Target



Mesh Target

(a) (b) Target

Aside: CLIP for Semantic Image Stylization



See "Image-Based CLIP-Guided Essence Transfer". H. Chefer, **S. Benaim**, R. Paiss, L. Wolf. In Submission.

Images

- Multi-sample approaches
- Structural analogies via patches of image pair

Videos

- Speed up videos "gracefully" using "speed" as supervision

3D Objects

- Semantic stylization using text (CLIP-based)

Visual Understanding via Semantic Manipulation

Next?

- Manipulation of multiple 3D objects in complex scenes.
- Manipulation under "constraints" derived for AR devices.
- Functional relationships: A person riding a bike vs a person beside a bike

Images

- Multi-sample approaches
- Structural analogies via patches of image pair

Videos

- Speed up videos “gracefully” using “speed” as supervision

3D Objects

- Semantic stylization using text (CLIP-based)

Thank You! Questions?

Visual Understanding via Semantic Manipulation

Next?

- Manipulation of multiple 3D objects in complex scenes.
- Manipulation under “constraints” derived for AR devices.
- Functional relationships: A person riding a bike vs a person beside a bike