Structure-Aware Manipulation of Images and Videos

Sagie Benaim
School of Computer Science, Tel Aviv University

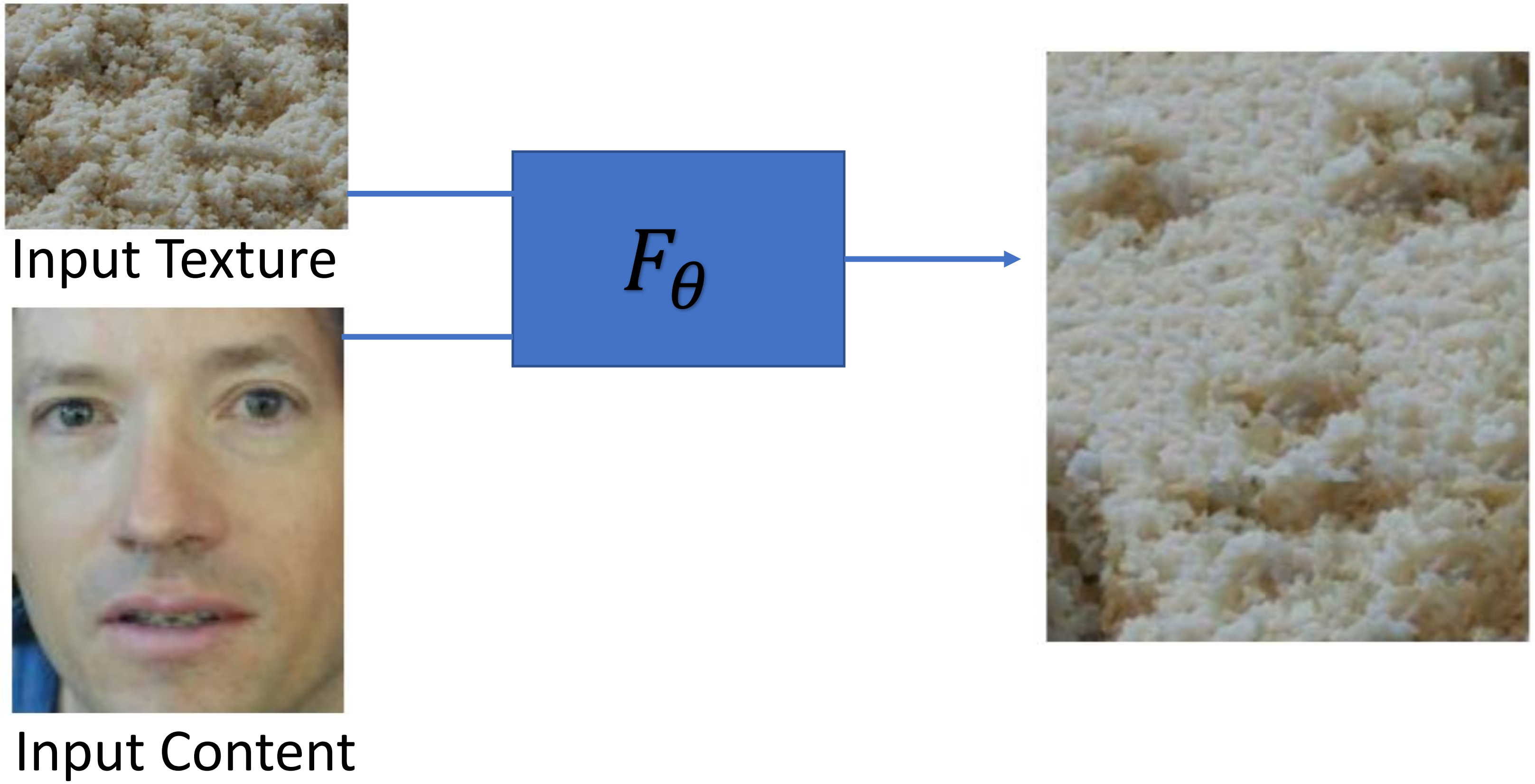
TEL AVIV UNIVERSITY

What is a natural image?

Intelligent machines must **understand** perceived content

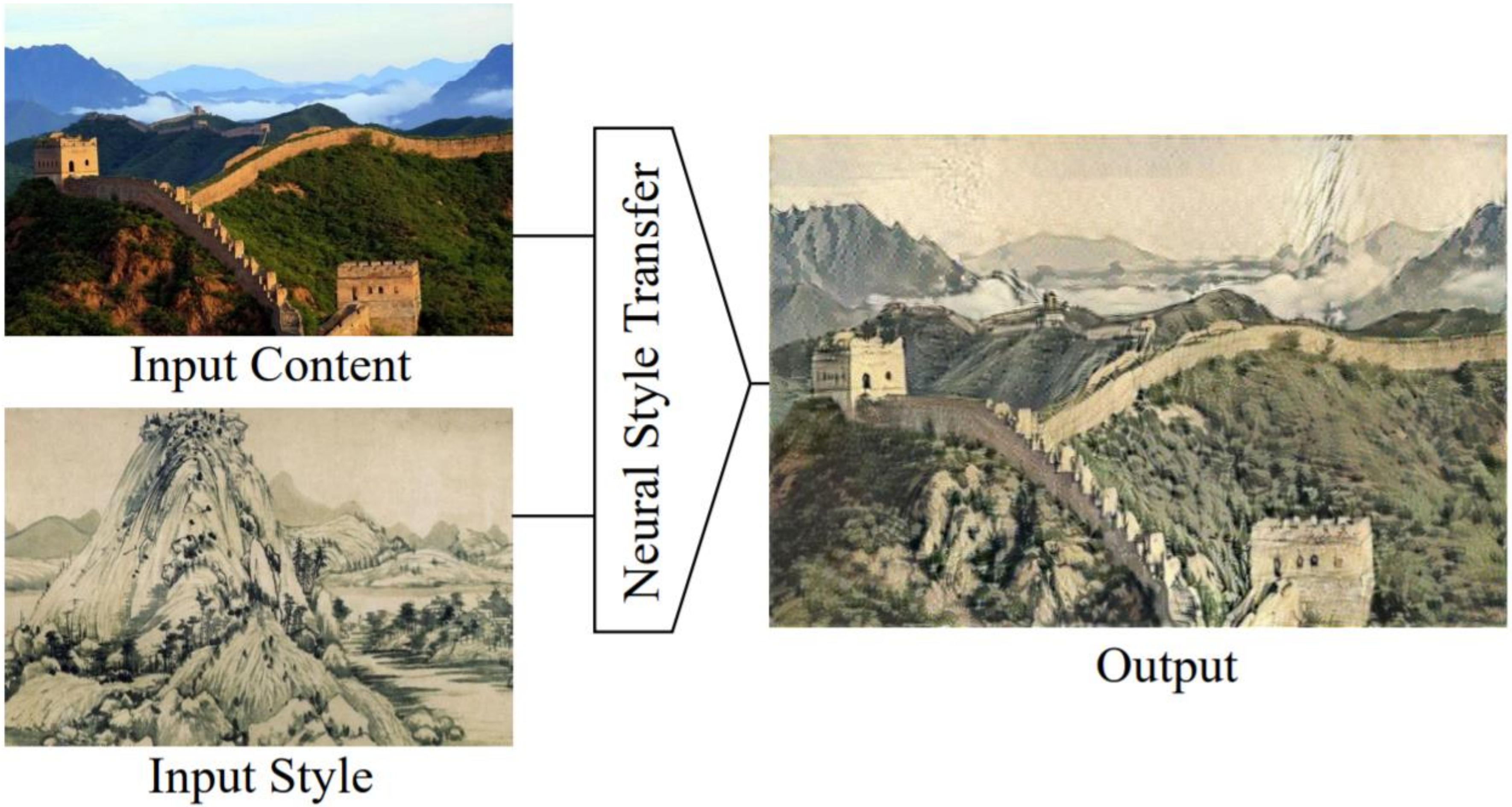
Understanding by creating/manipulating: "What I cannot create, I do not understand" (Richard Feynman)

Manipulating Texture



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

Manipulating Style

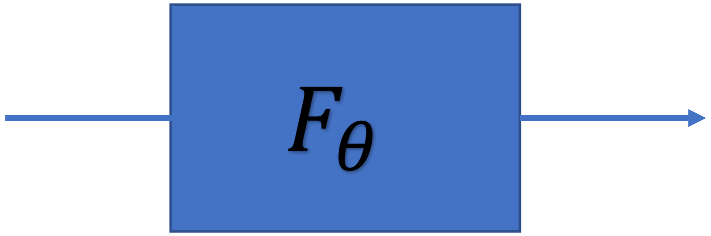


L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style”. 2015.

Manipulating Structure

Target

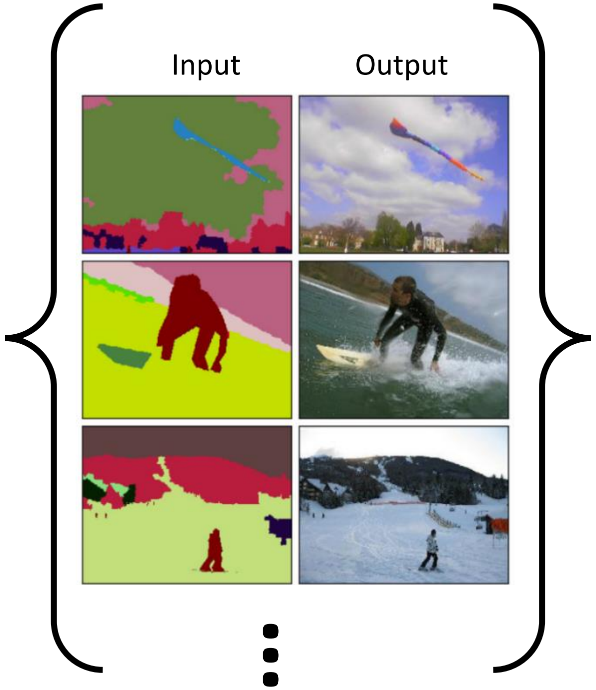
Source Structure



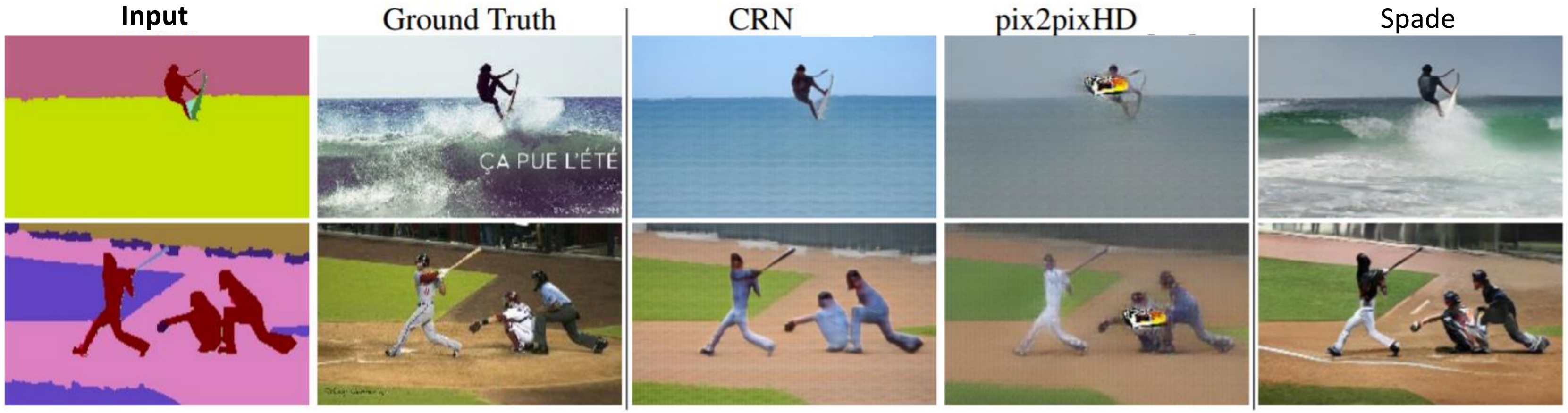
Multi-Sample Approaches

Supervised (Paired) Setting

Train



Test



Unsupervised (Unpaired) Setting

A

Faces with glasses

B

Faces <u>without</u> 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.**

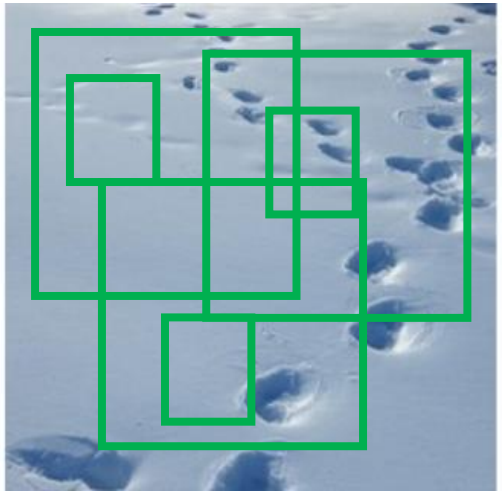
R. Mokady, **S. Benaim**, L. Wolf, A. Bermano. Mask Based Unsupervised Content Transfer. In **ICLR, 2020.**

Require a large collection of images from both domains

Patch-Based Approaches

Multi-Image Distribution

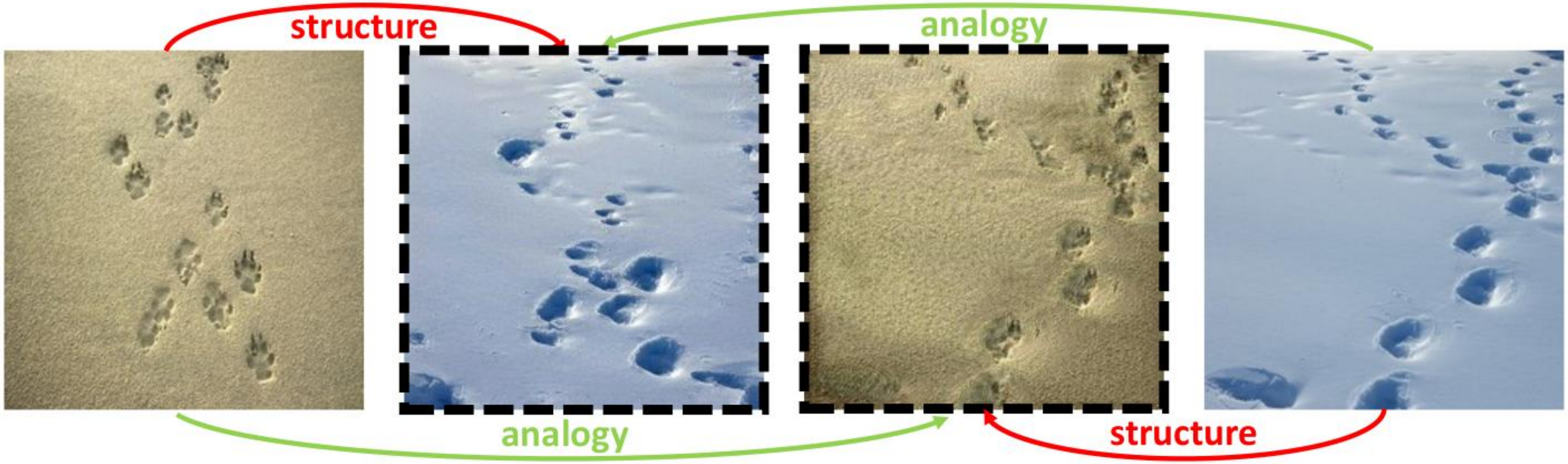
Multi-Scale Patch Distribution



Brock et al., Large Scale GAN Training for High Fidelity Natural Image Synthesis. ICLR 2019

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

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



Hierarchical Patch VAE-GAN:
Generating Diverse Videos from a **Single Sample**

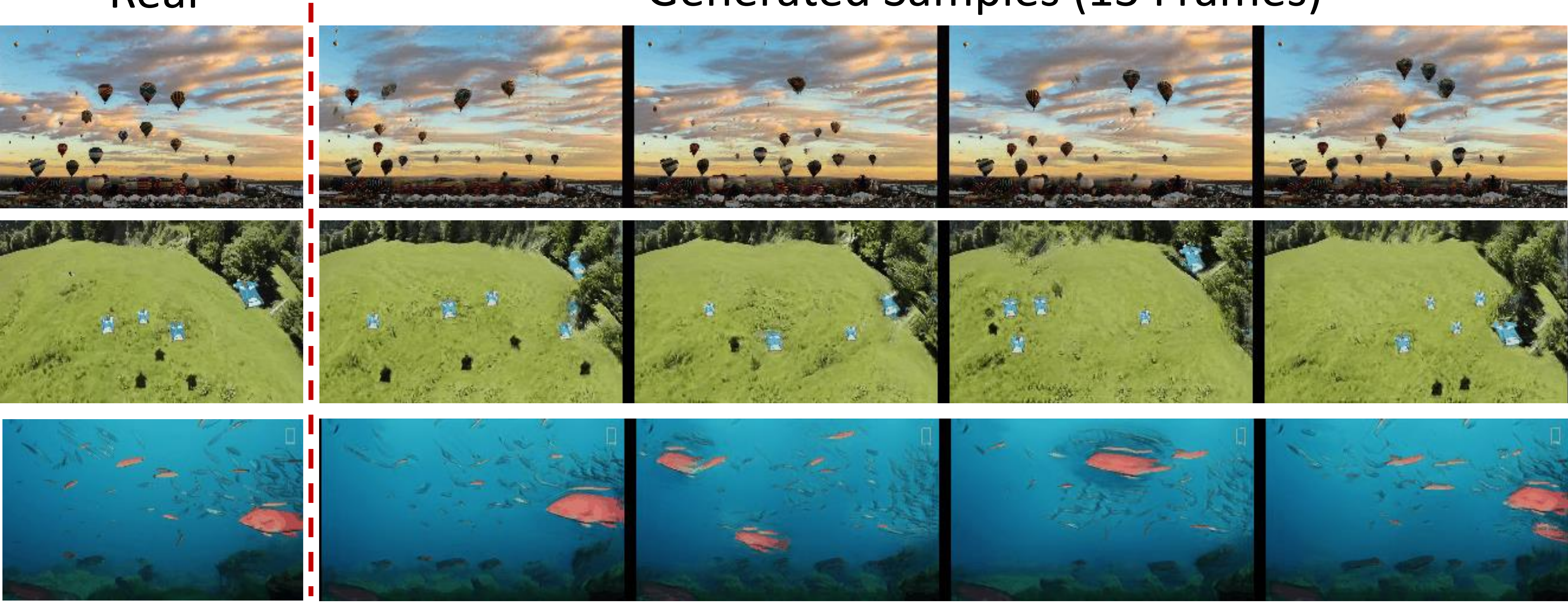
S. Gur*, **S. Benaim***, L. Wolf. NeurIPS 2020 (*Equal contribution)

Real

Hierarchical Patch VAE-GAN:
Generating Diverse Videos from a Single Sample

S. Gur*, **S. Benaim***, L. Wolf. NeurIPS 2020 (*Equal contribution)

Real Generated Samples (13 Frames)



Extending 2D to 3D

Real

Ours

Real

SinGAN [1] + 3D Convolution

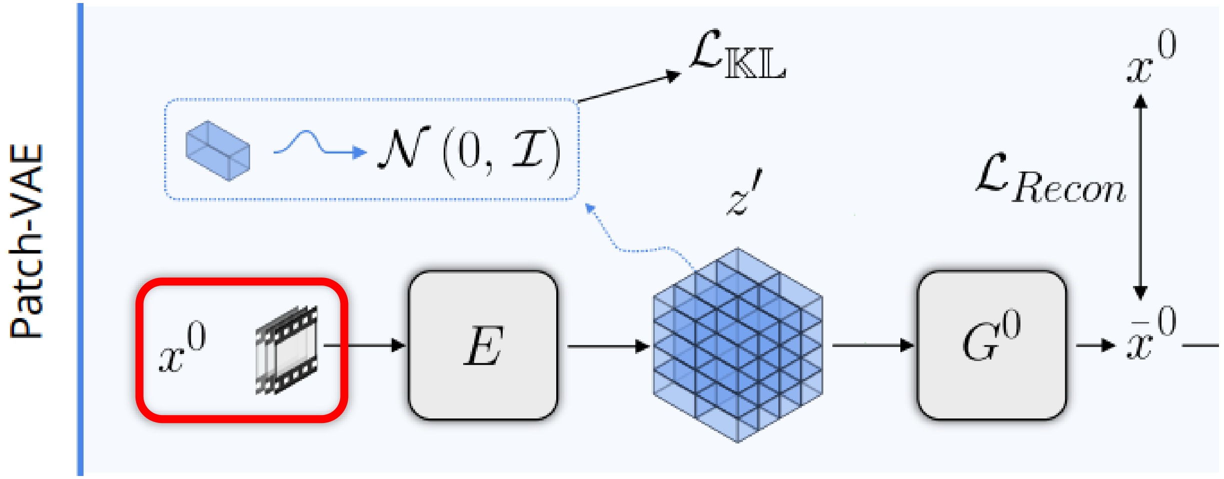
Real

ConSinGAN [2] + 3D Convolution

[1] “SinGAN: Learning a Generative Model from a Single Natural Image”, Shaham et al., ICCV 2019
[2] “Improved Techniques for Training Single-Image GANs”, Hinz et al., arXiv 2020

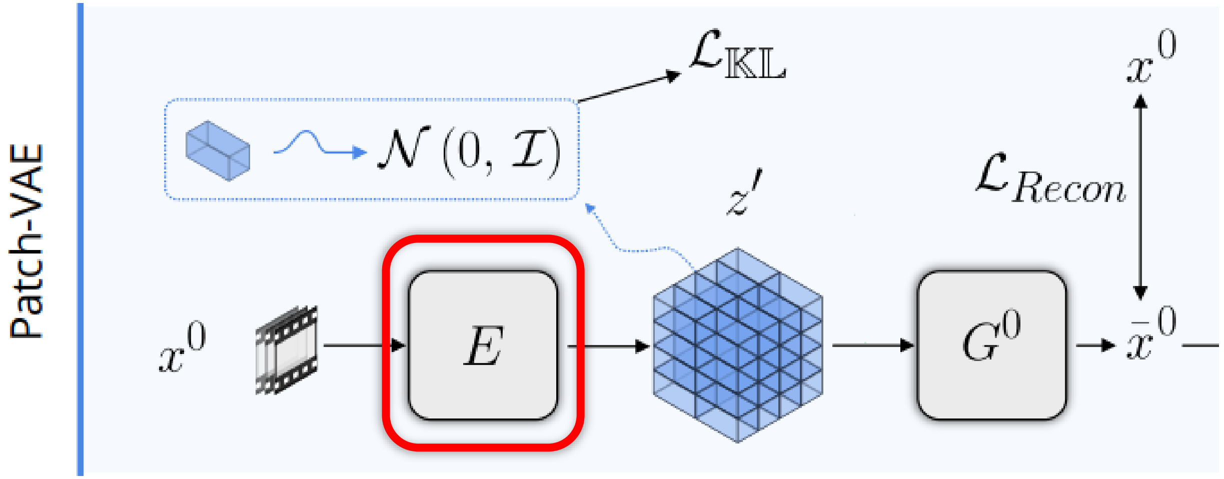
Proposed Approach: Patch VAE

Input video - x^0



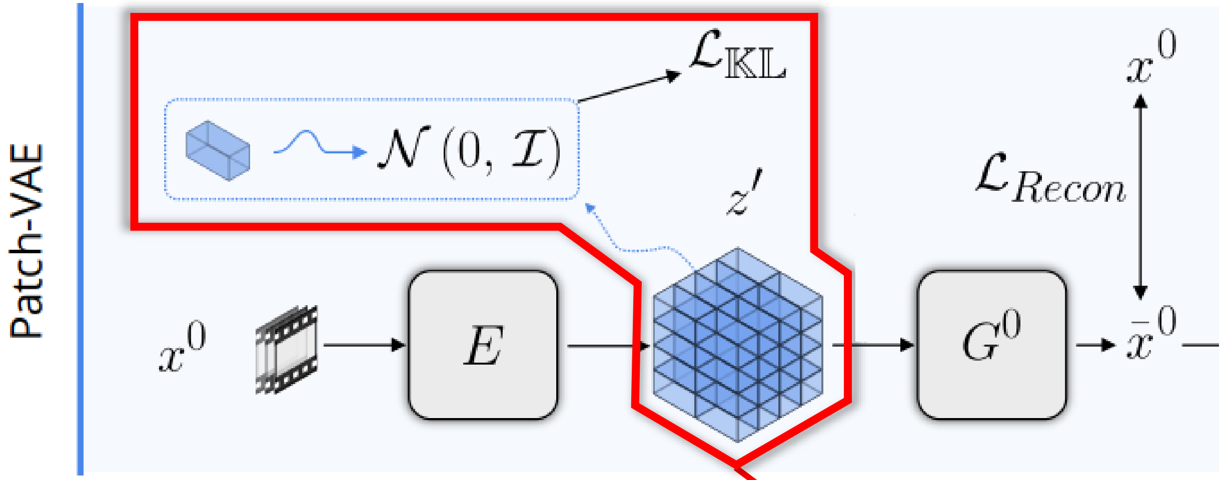
Proposed Approach: Patch VAE

Encoder – $\mathrm{E}(x^0)$

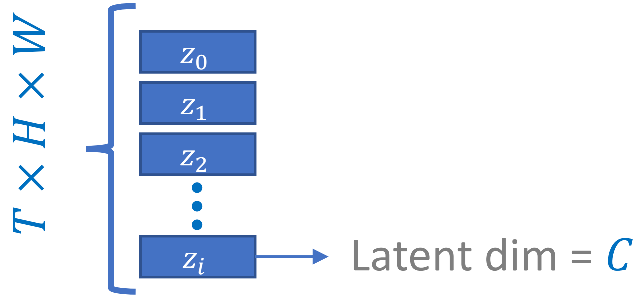


Proposed Approach: Patch VAE

KL Loss – z'

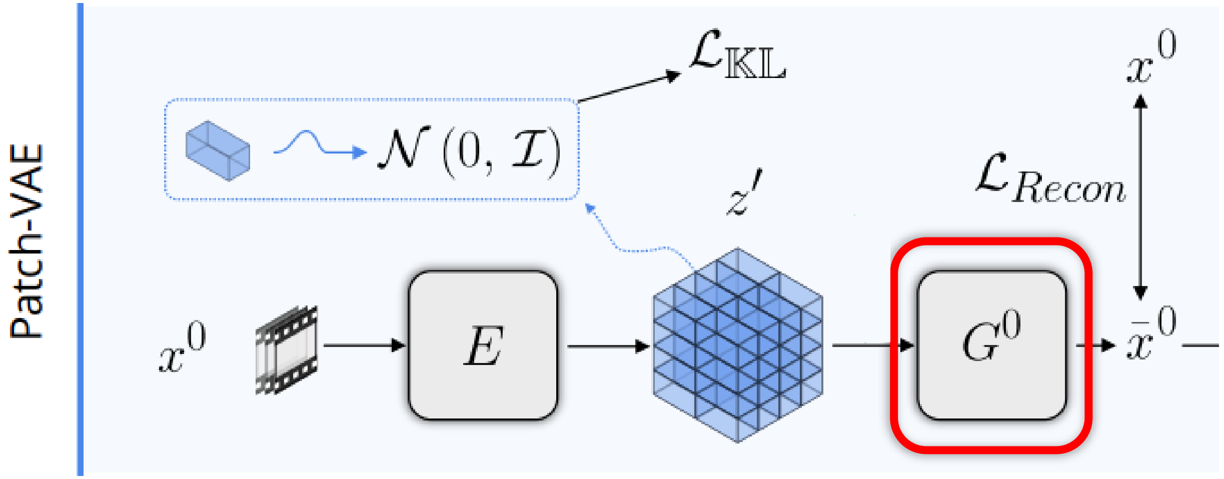


Each feature z_i , $i = [1 \ldots K], K = T \times H \times W$, in the latent space is associated with a patch ω_i



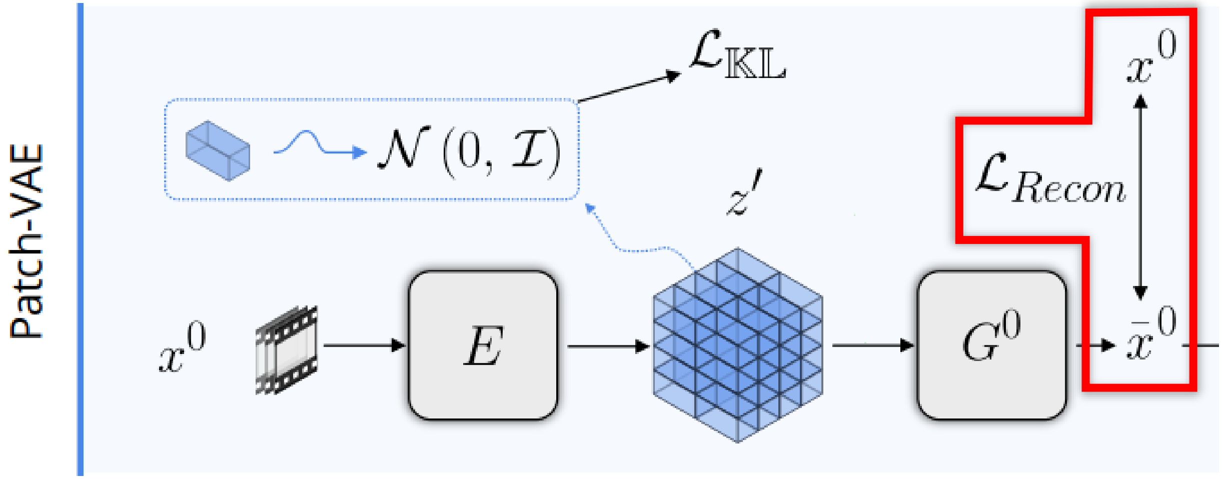
Proposed Approach: Patch VAE

Decoder - $\bar{x}^0$



Proposed Approach: Patch VAE

Reconstruction loss



Proposed Approach: Hierarchical Patch VAE

Coarsest scale:
Low resolution
and frame rate

x^0 (Real)
$\bar{x}^0$(Generated)

LEVEL $= 0$

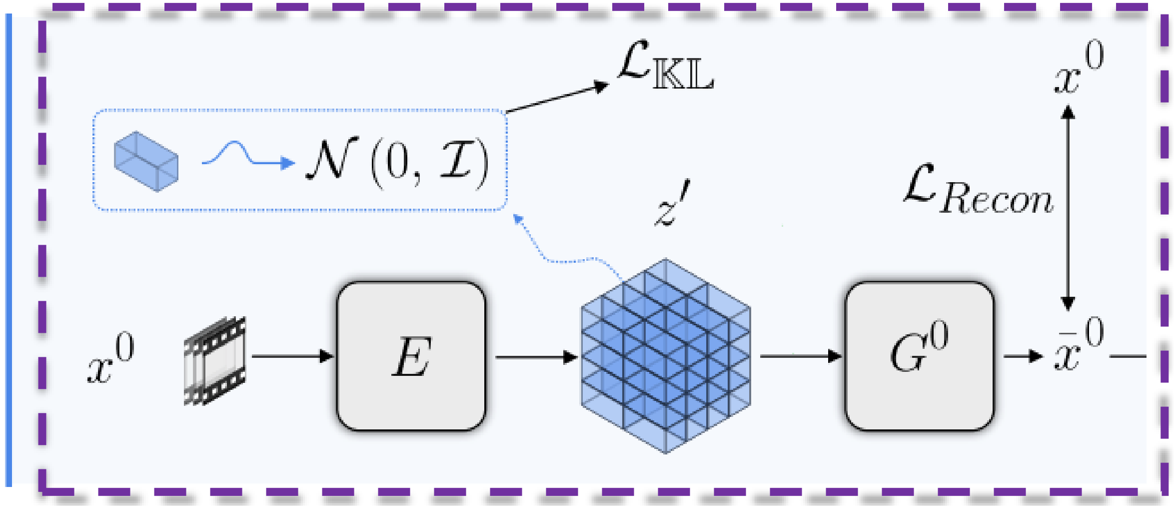
Finest scale:
High resolution
and frame rate

x^N (Real)
$\bar{x}^N$(Generated)

LEVEL $= N$

Proposed Approach: Hierarchical Patch VAE

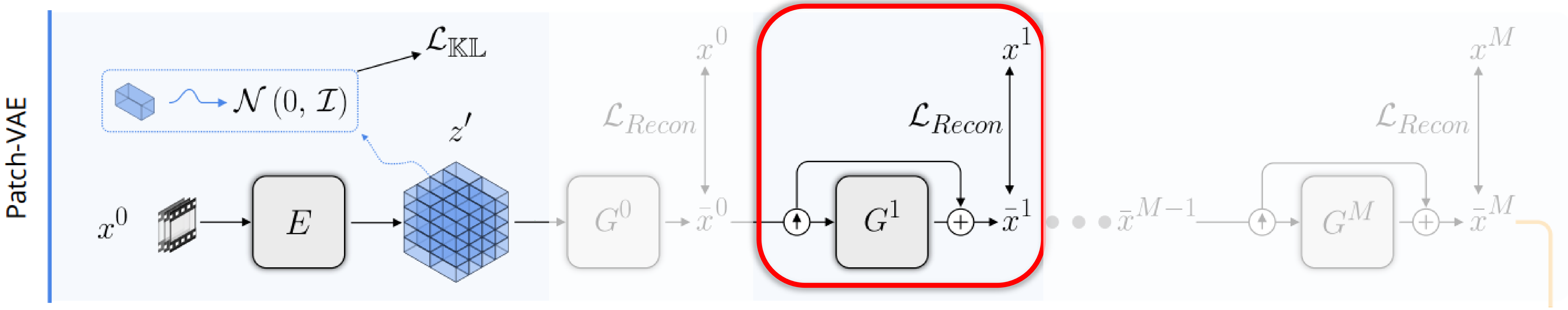
Patch-VAE



LEVEL = 0

Proposed Approach: Hierarchical Patch VAE

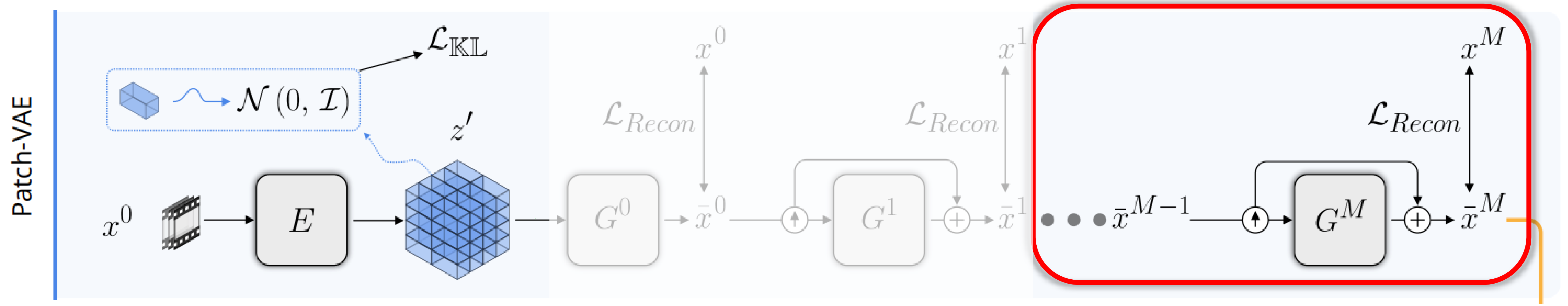
Up-sampling block - $\bar{x}^1$



LEVEL = 1

Proposed Approach: Hierarchical Patch VAE

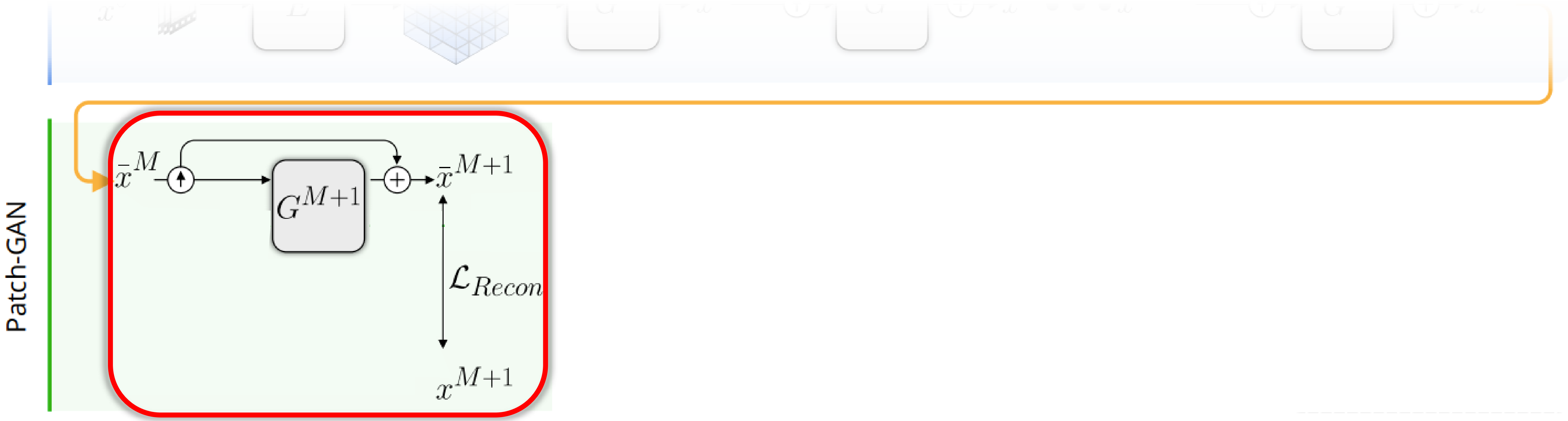
Hierarchical up-sampling up to $\bar{x}^M$



LEVEL $\leq M$

Proposed Approach: Hierarchical Patch VAE GAN

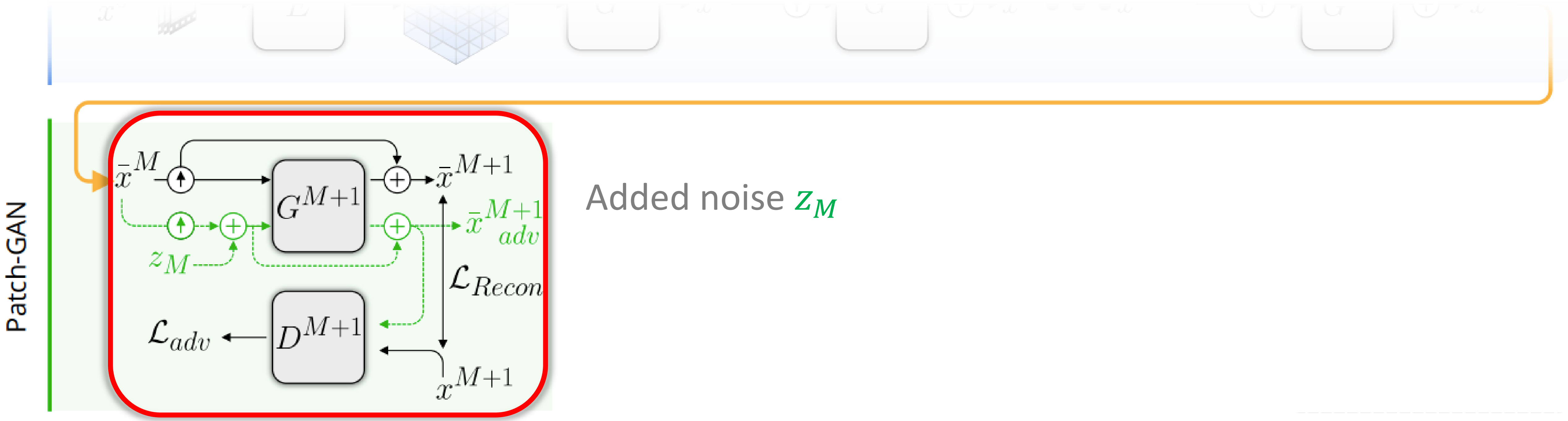
Up-sampling block $\bar{x}^{M+1}$



LEVEL $= M + 1$

Proposed Approach: Hierarchical Patch VAE GAN

Adversarial training

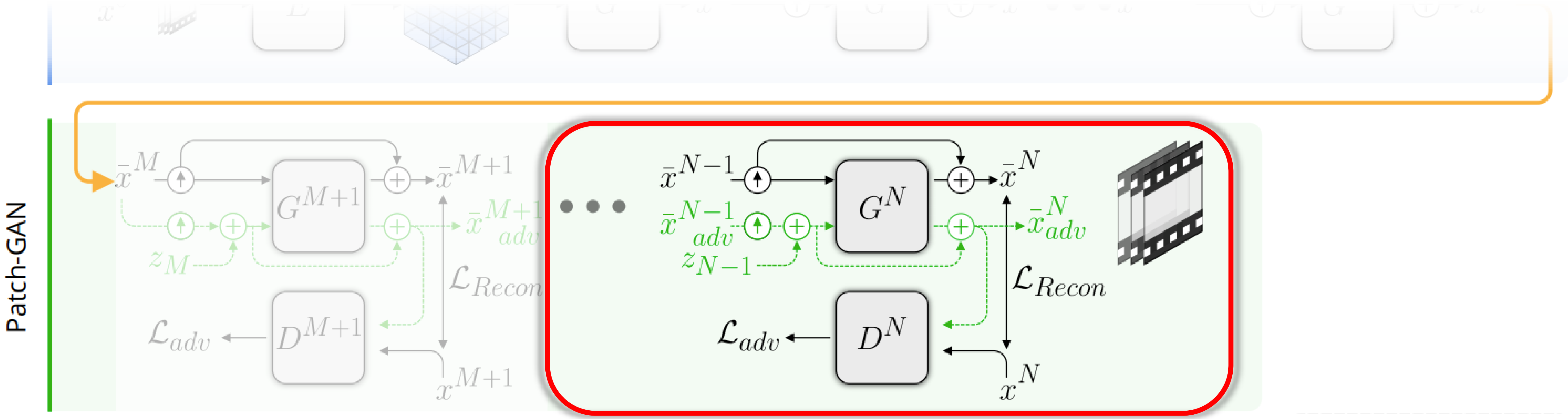


Added noise z_M

LEVEL $= M + 1$

Proposed Approach: Hierarchical Patch VAE GAN

Hierarchical up-sampling up to final resolution $\bar{x}^N$



$M + 1 <$ LEVEL $\leq N$

Effect of Number of patch-VAE levels

Training Video

9 Levels Total

1 p-VAE – 8 p-GAN

8 p-VAE – 1 p-GAN

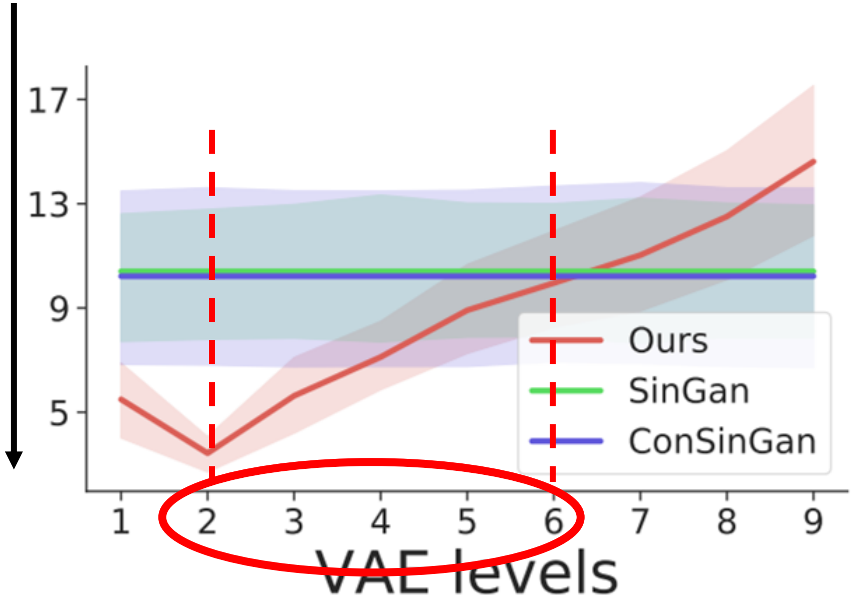
3 p-VAE – 6 p-GAN

Effect of Number of patch-VAE levels

Total of 9 layers

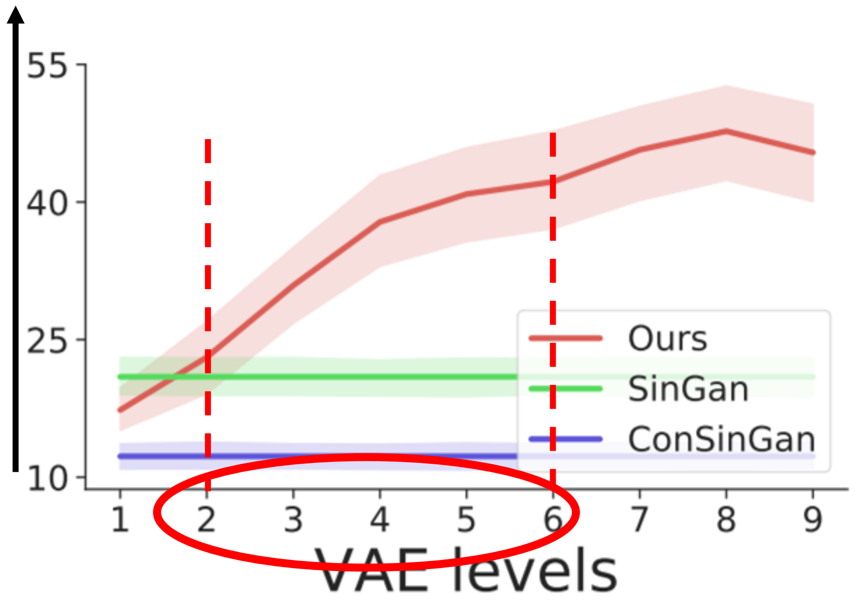
Quality

(**Lower** is Better)



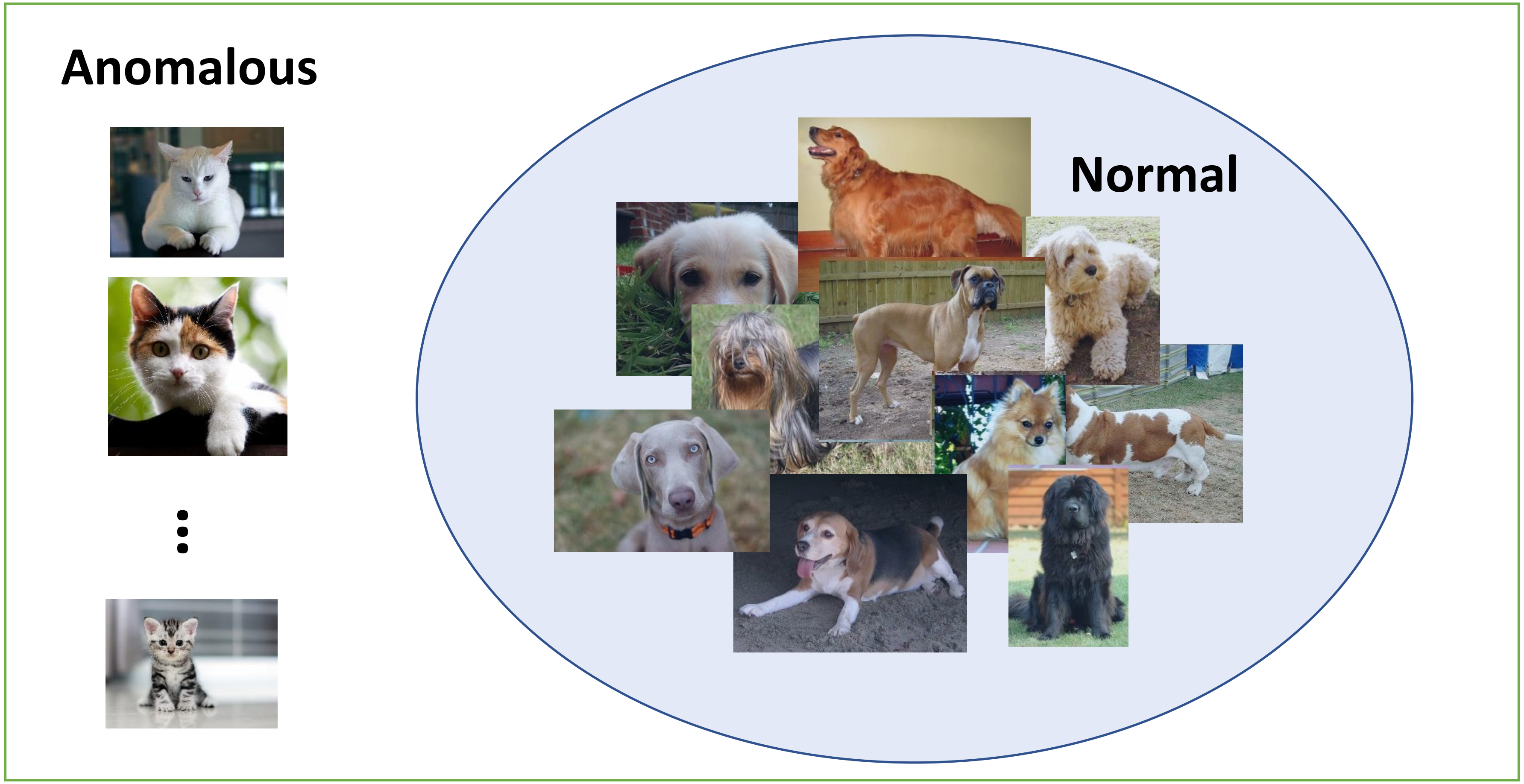
Diversity

(**Higher** is Better)



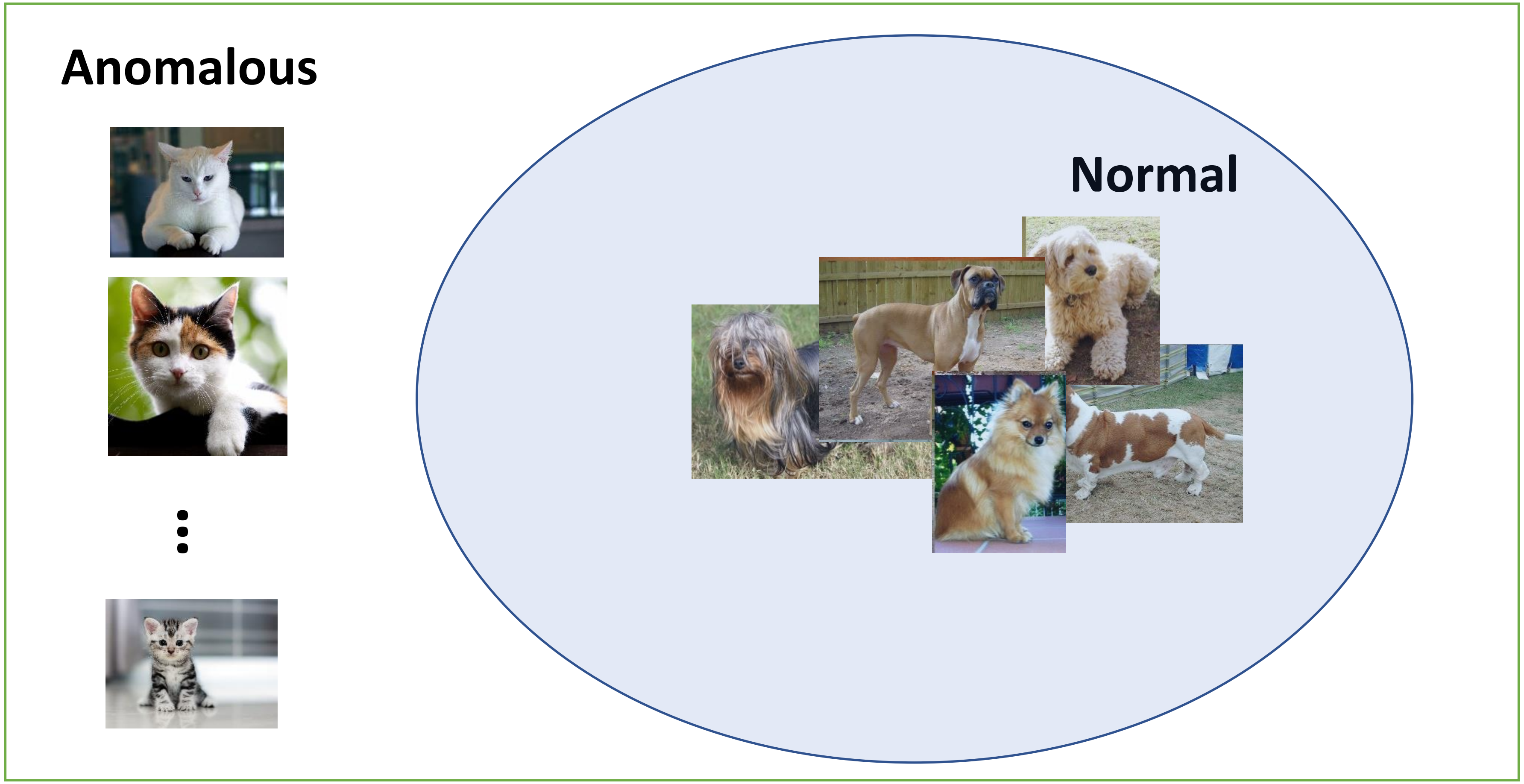
A Hierarchical Transformation-Discriminating Generative Model for **Few Shot Anomaly Detection**

S. Sheynin*, **S. Benaim***, L. Wolf. In Submission to ICCV 2021. (*Equal contribution)



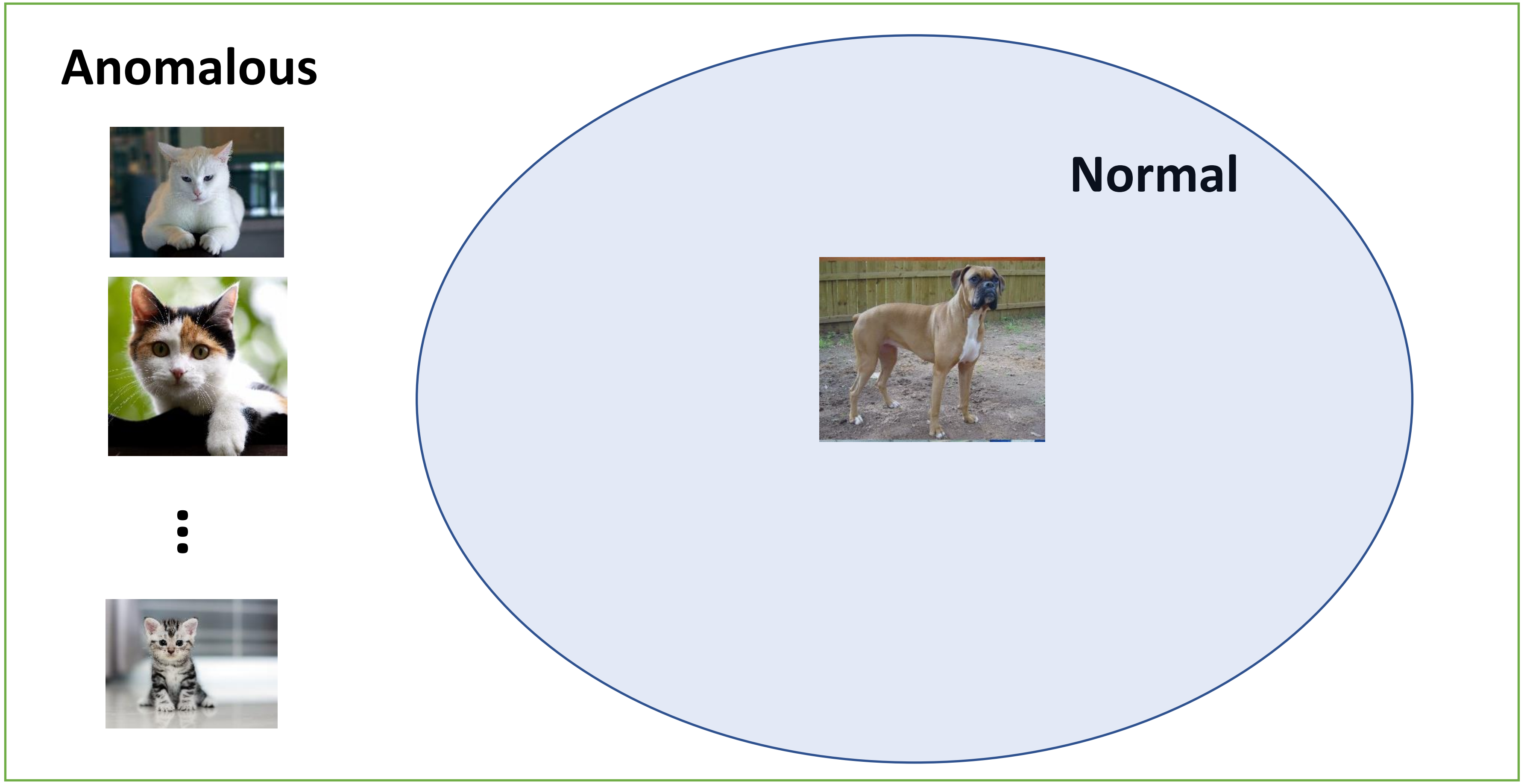
A Hierarchical Transformation-Discriminating Generative Model for **Few Shot Anomaly Detection**

S. Sheynin*, **S. Benaim***, L. Wolf. In Submission to ICCV 2021. (*Equal contribution)

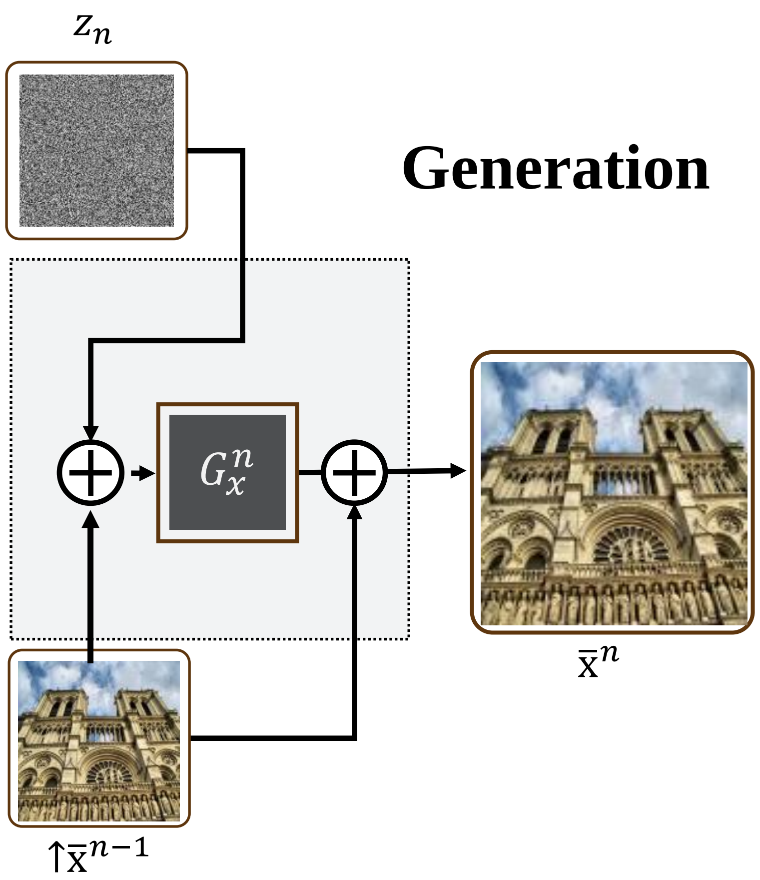


A Hierarchical Transformation-Discriminating Generative Model for **Few Shot Anomaly Detection**

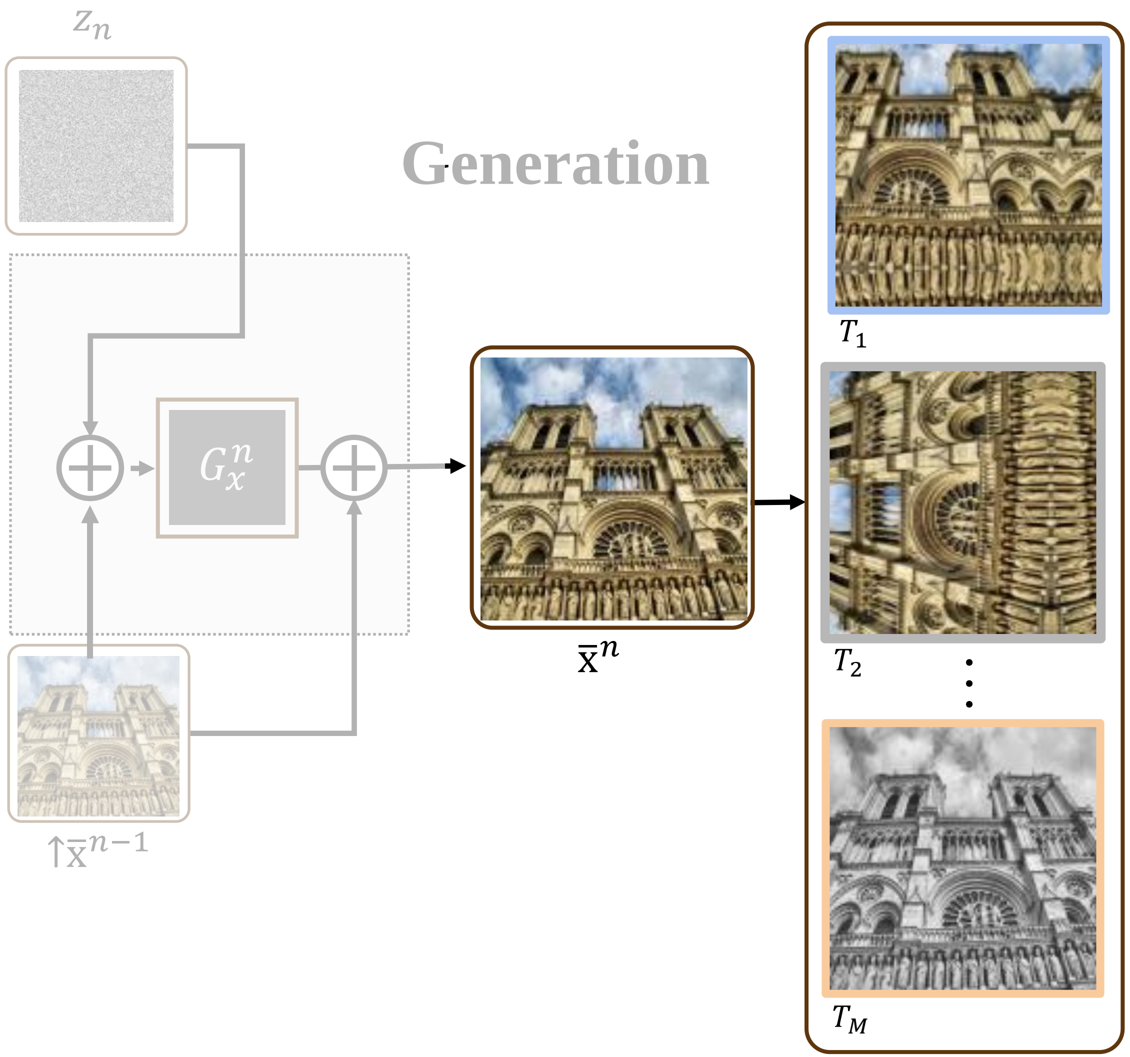
S. Sheynin*, **S. Benaim***, L. Wolf. In Submission to ICCV 2021. (*Equal contribution)



Multi-Scale **Generation** (Level n)



Transform Generated Sample



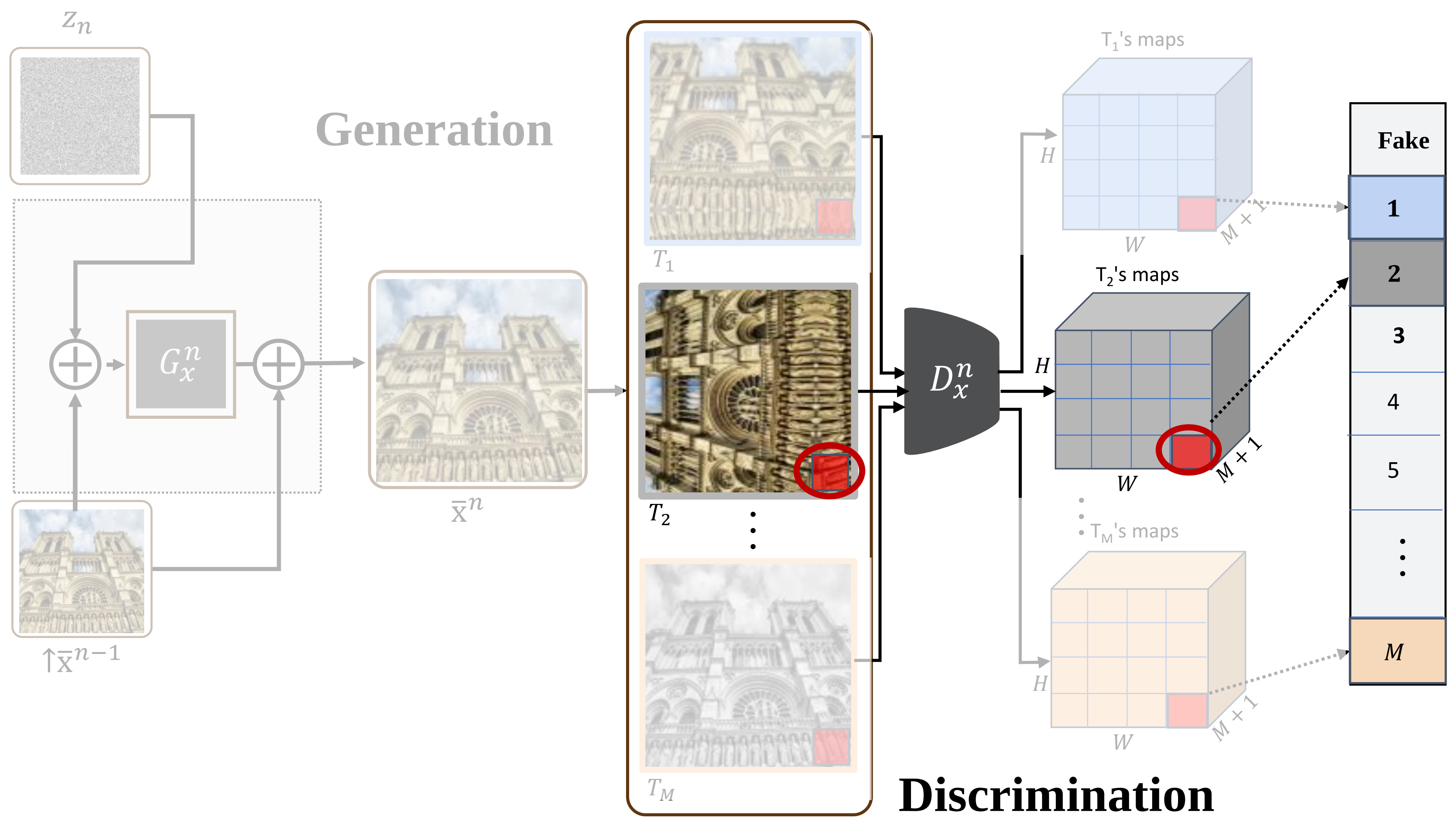
T_1: Horizontal Flip, Translation (y-axis)

T_2: 90° Rotation, Translation (x-axis), Translation (y-axis)

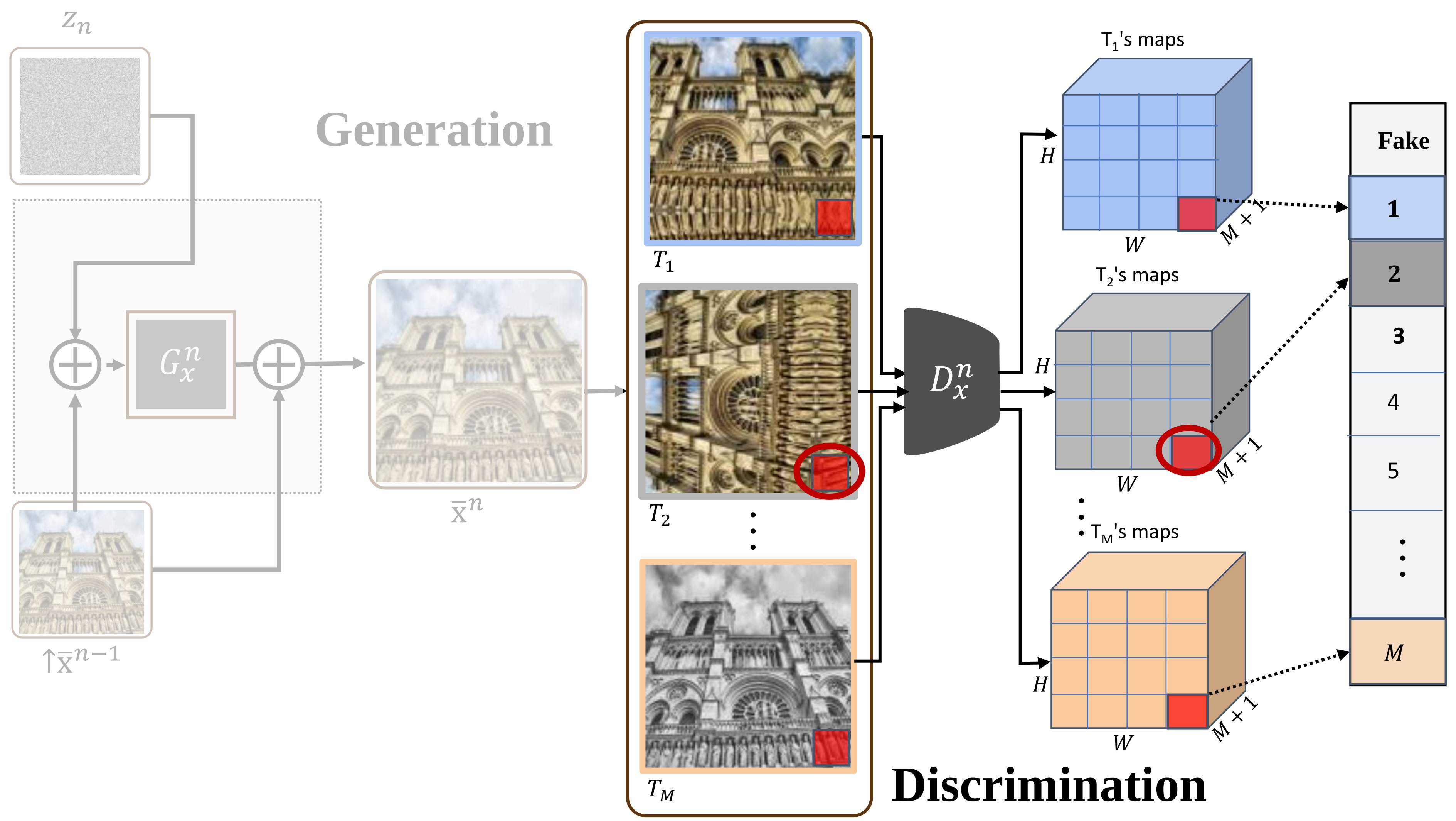
...

T_M: Grayscale (y-axis)

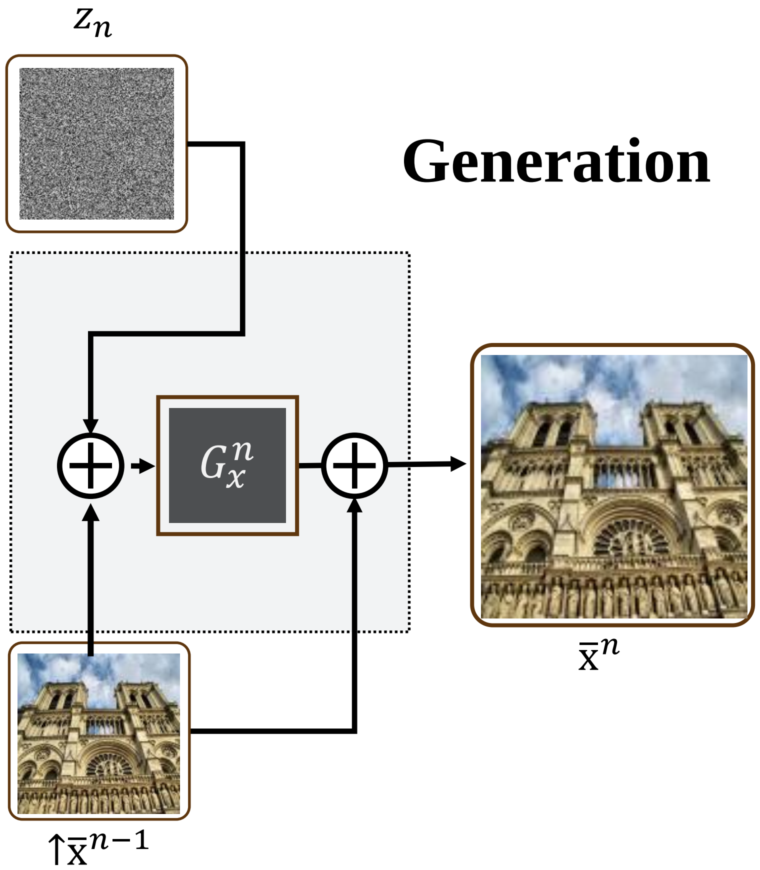
Patch-Based **Self Supervised Task**



Patch-Based **Self Supervised Task**

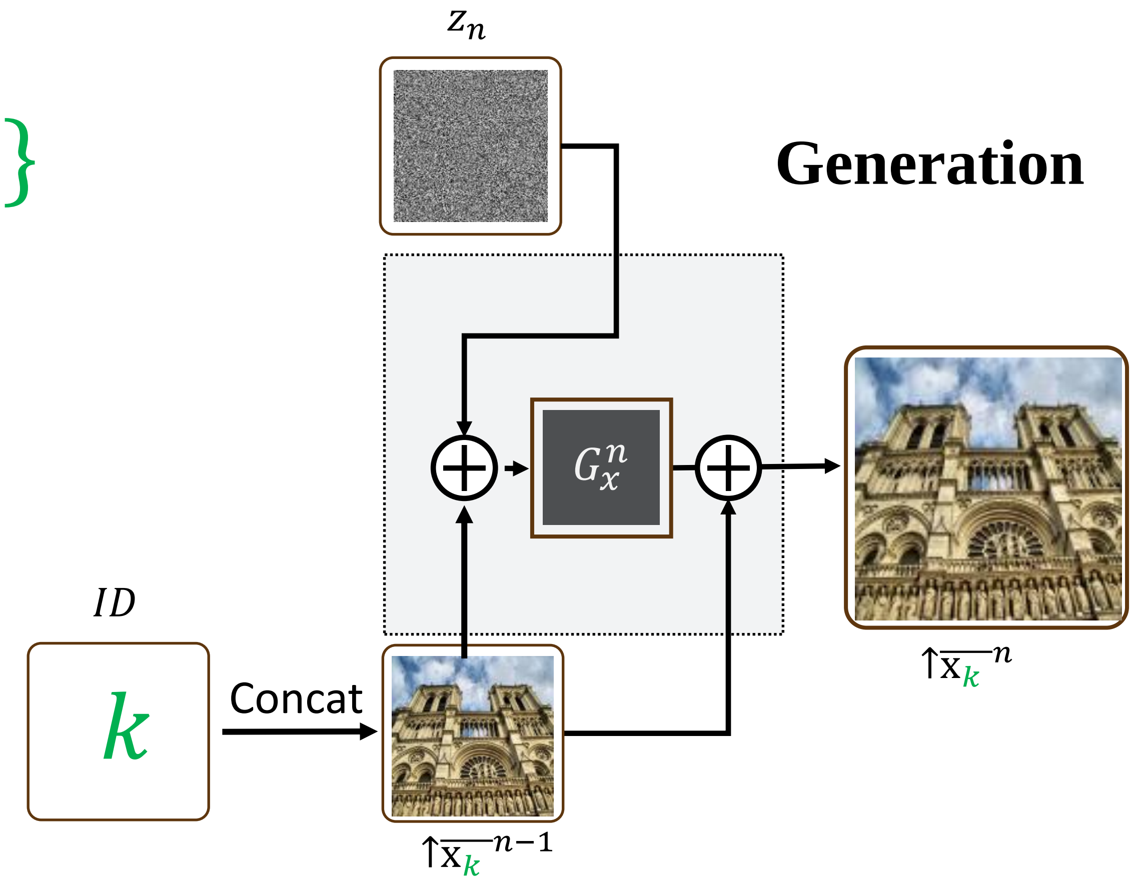


Single Sample

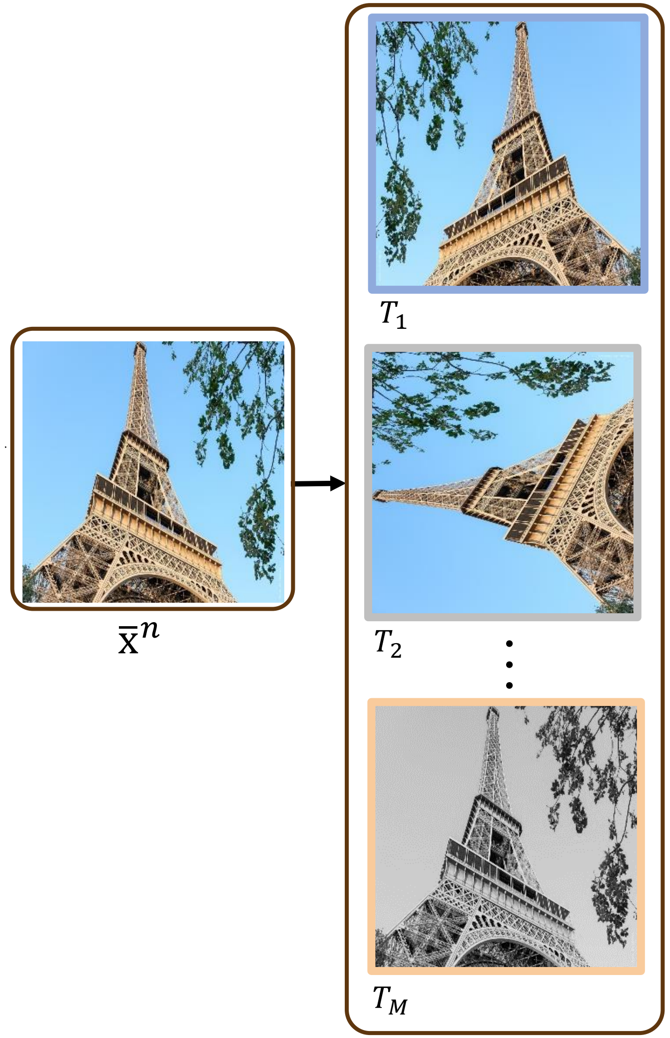


Multiple Samples

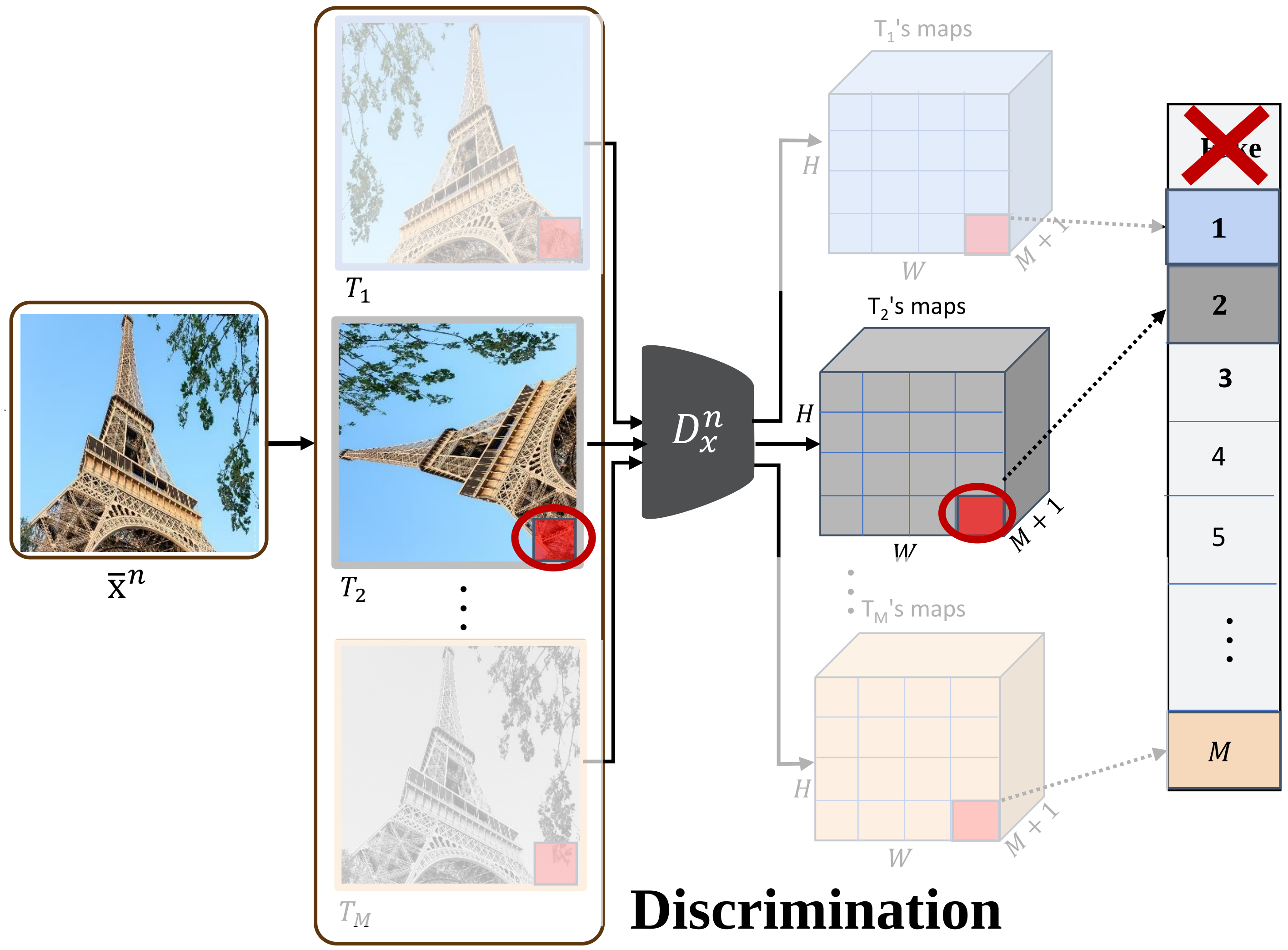
$k \in \{1..K\}$



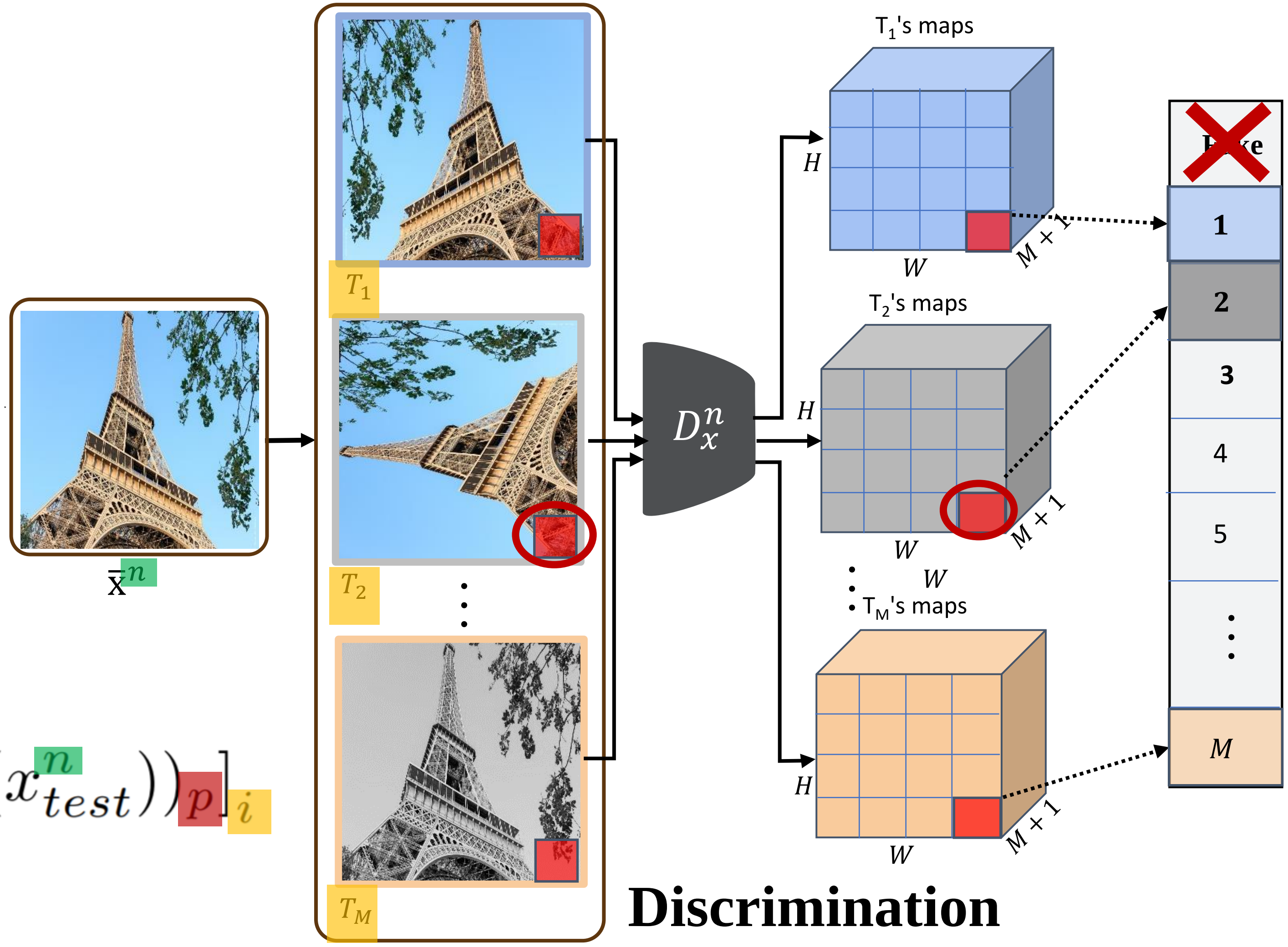
Test Time: Anomaly Score (Scale n)



Test Time: Anomaly Score (Scale n)

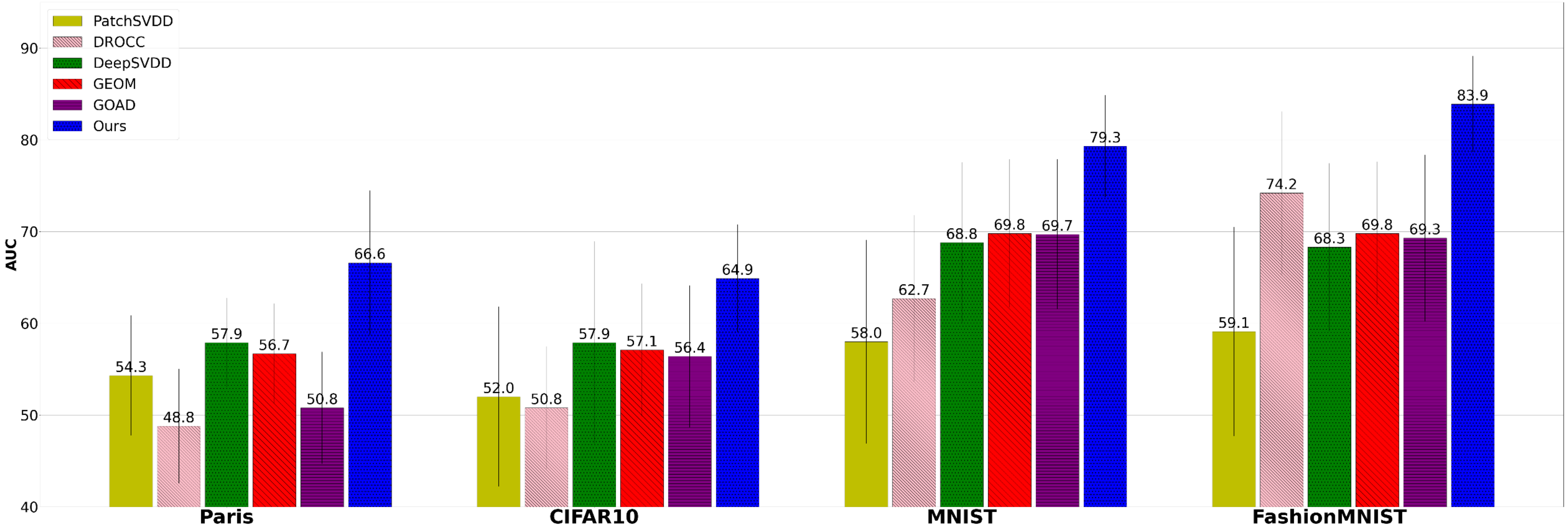


Test Time: Anomaly Score (Scale n)



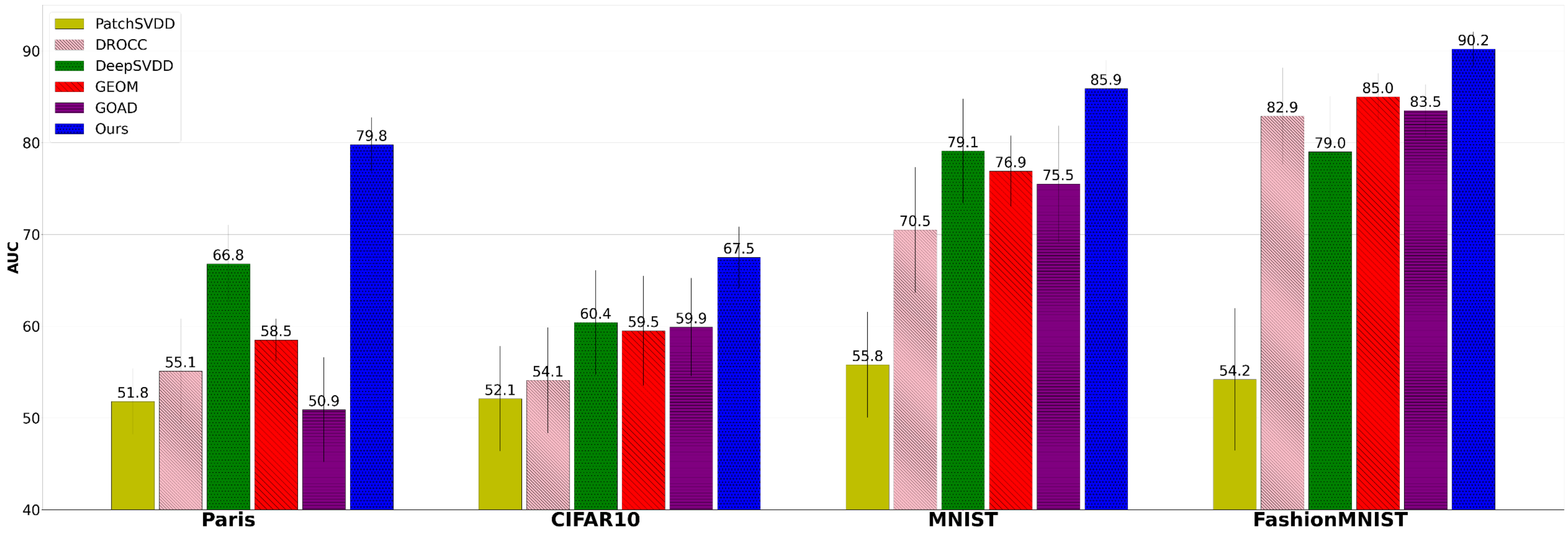
$$\sum_{n=0}^{N}\sum_{i=1}^{M}\sum_{p\in H\,x\,W}[D_x^{n*}(T_i(x^n_{test}))_p]_i$$

One-Shot



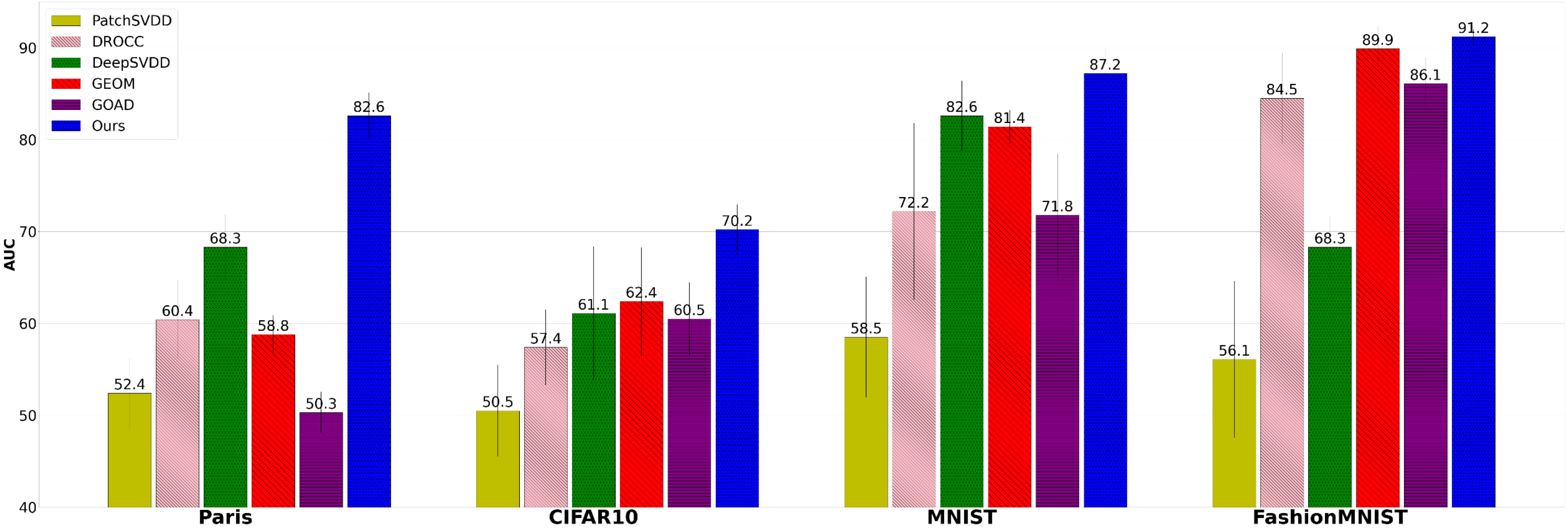
PatchSVDD
DROCC
DeepSVDD
GEOM
GOAD
Ours
AUC
90
80
70
60
50
40
54.3
48.8
57.9
56.7
50.8
66.6
52.0
50.8
57.9
57.1
56.4
64.9
58.0
62.7
68.8
69.8
69.7
79.3
59.1
74.2
68.3
69.8
69.3
83.9
Paris
CIFAR10
MNIST
FashionMNIST

Five-Shot



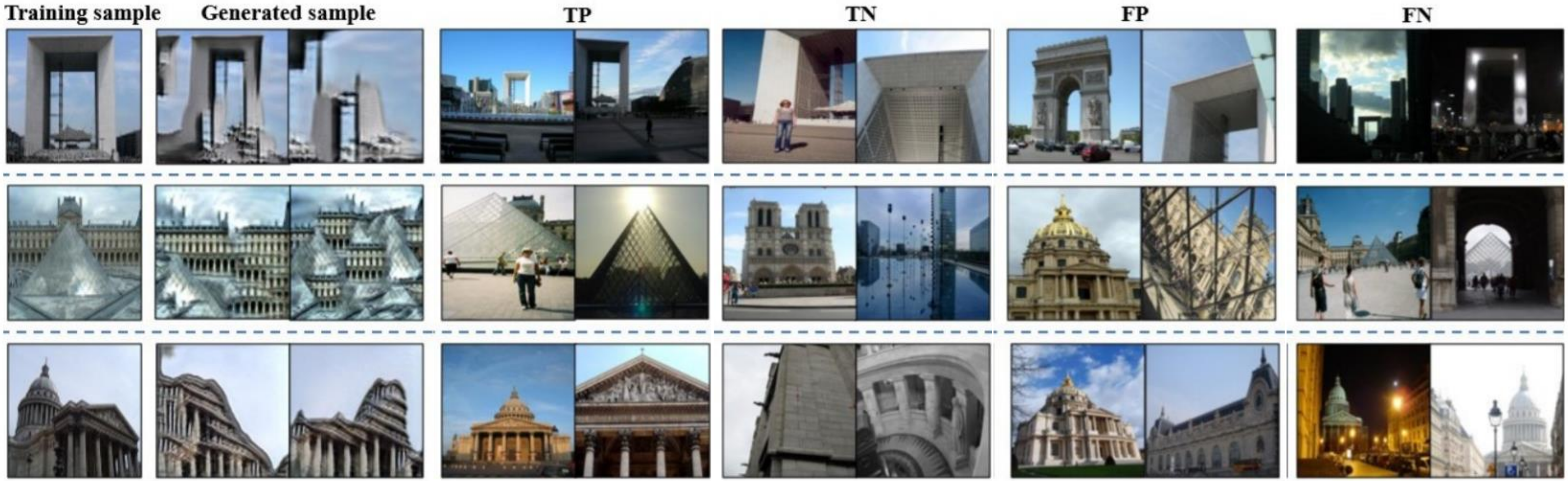
PatchSVDD
DROCC
DeepSVDD
GEOM
GOAD
Ours
AUC
40
50
60
70
80
90
51.8
55.1
66.8
58.5
50.9
79.8
52.1
54.1
60.4
59.5
59.9
67.5
55.8
70.5
79.1
76.9
75.5
85.9
54.2
82.9
79.0
85.0
83.5
90.2
Paris
CIFAR10
MNIST
FashionMNIST

Ten-Shot

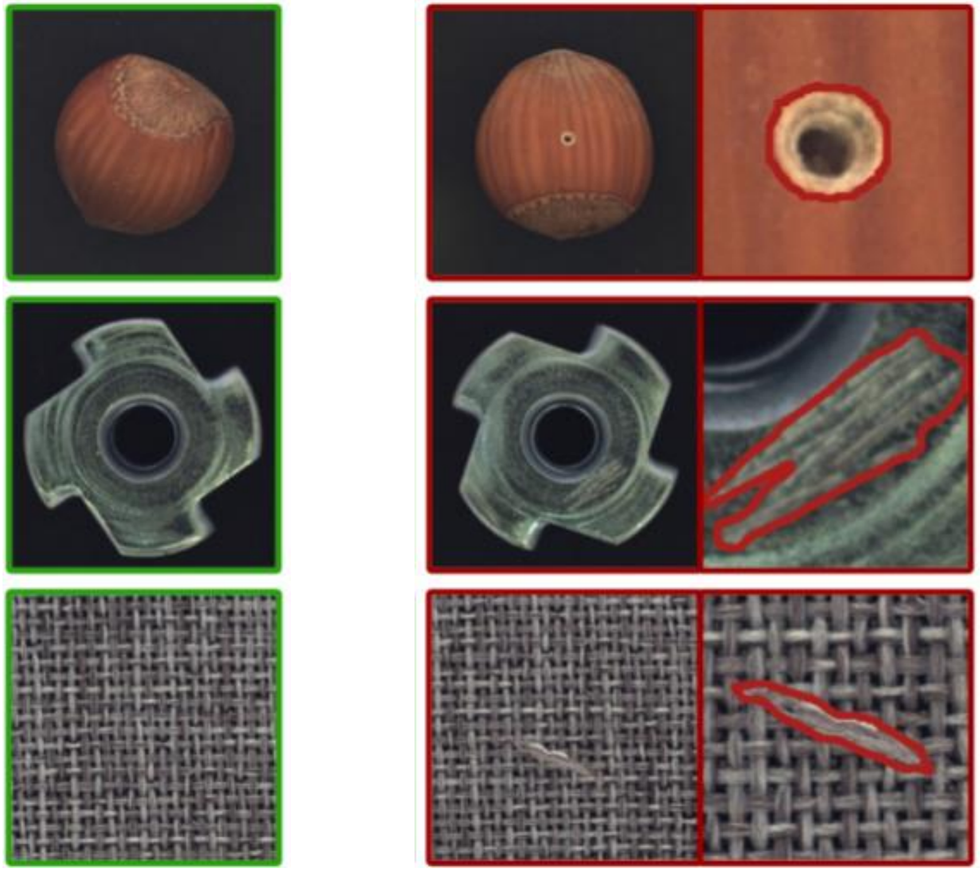


PatchSVDD
DROCC
DeepSVDD
GEOM
GOAD
Ours
AUC
90
80
70
60
50
40
52.4
60.4
68.3
58.8
50.3
82.6
50.5
57.4
61.1
62.4
60.5
70.2
58.5
72.2
82.6
81.4
71.8
87.2
56.1
84.5
68.3
89.9
86.1
91.2
Paris
CIFAR10
MNIST
FashionMNIST

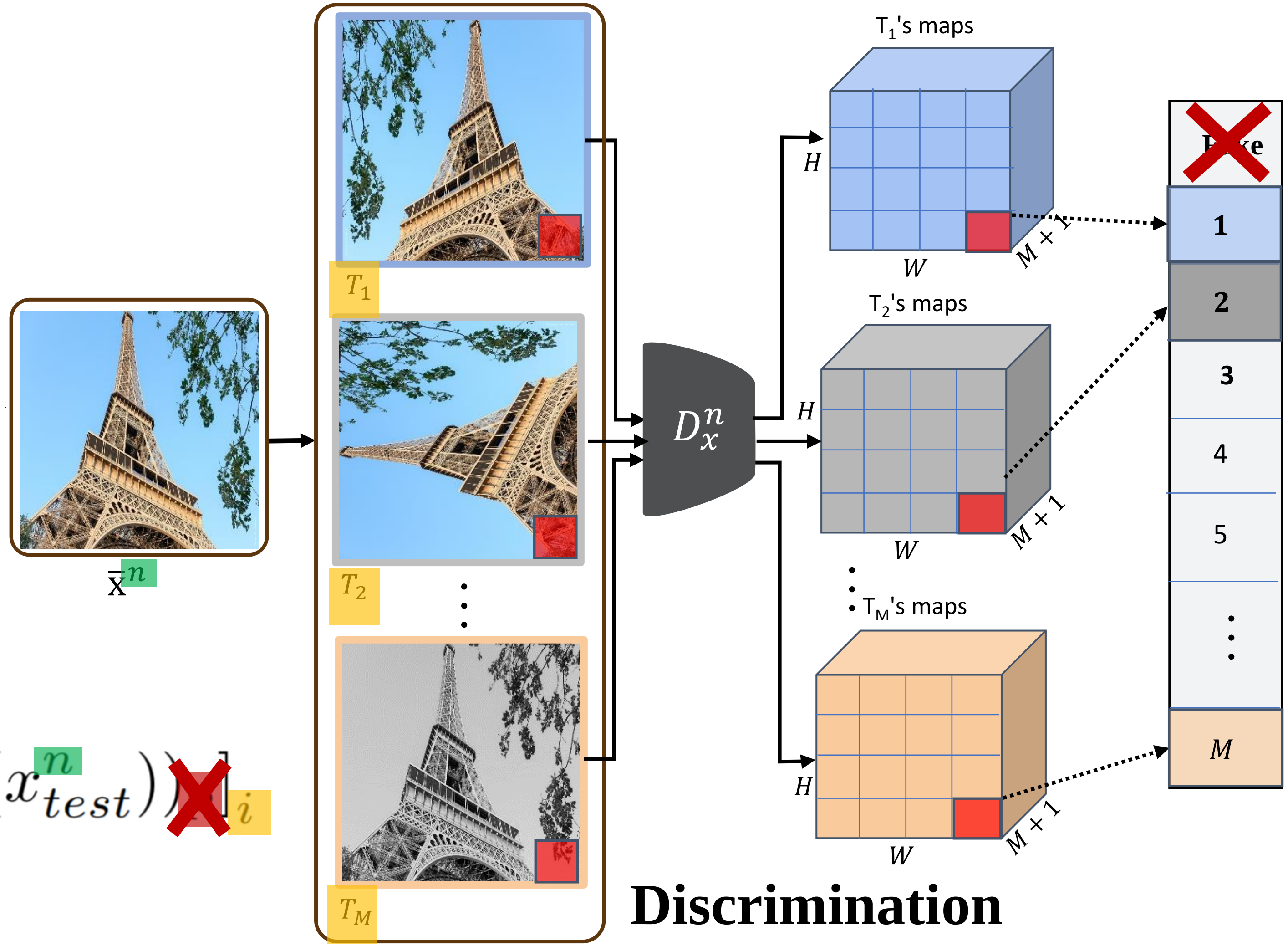
Predictions of our One-Shot Model



One Shot Defect Localization

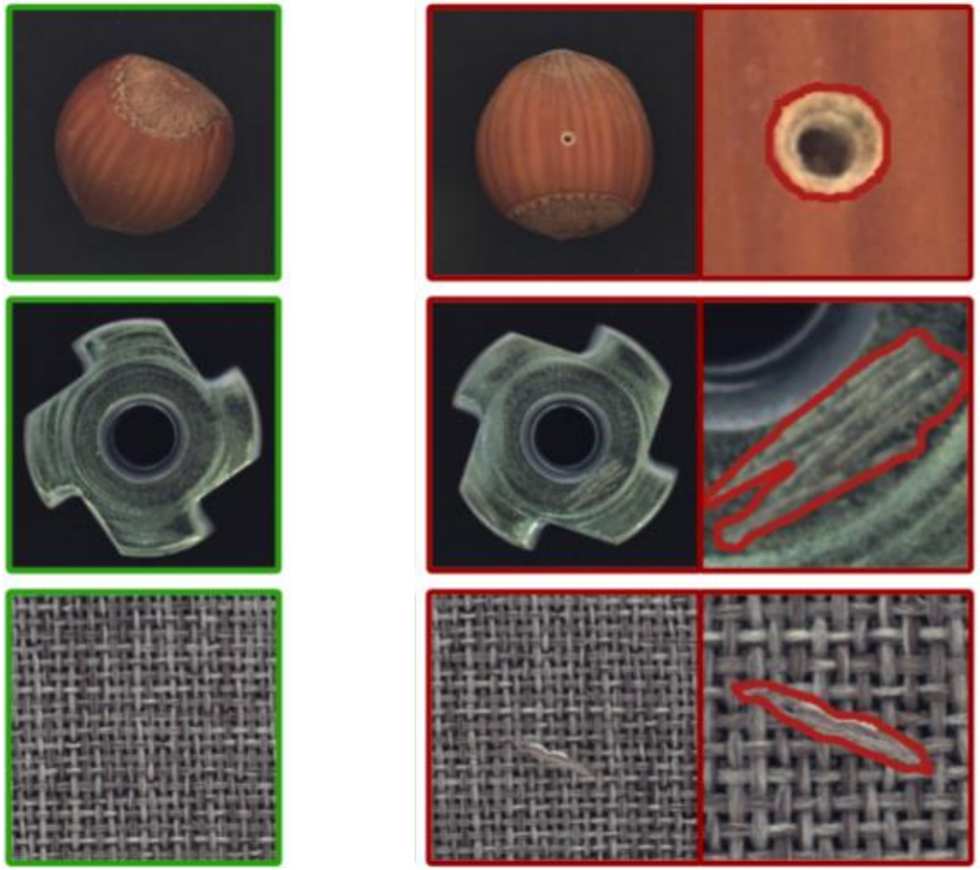


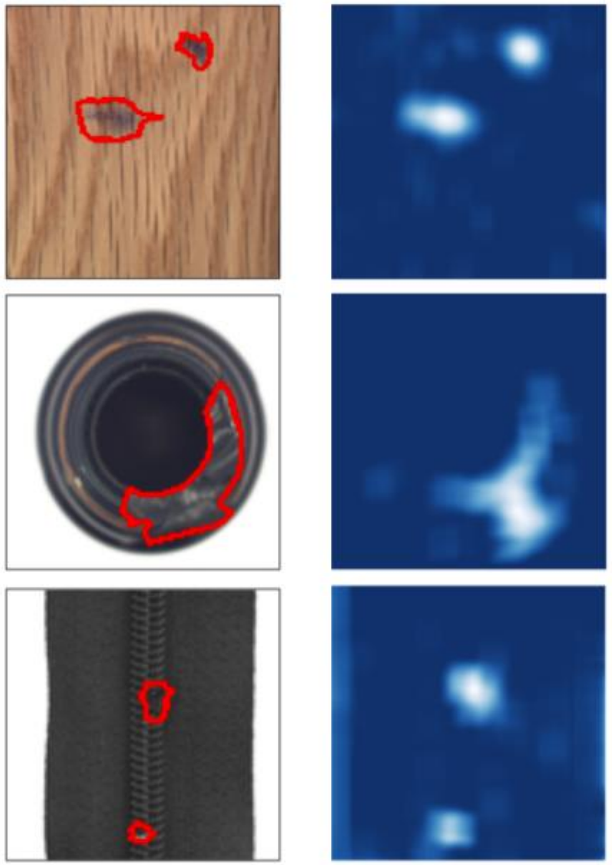
One Shot Defect Localization



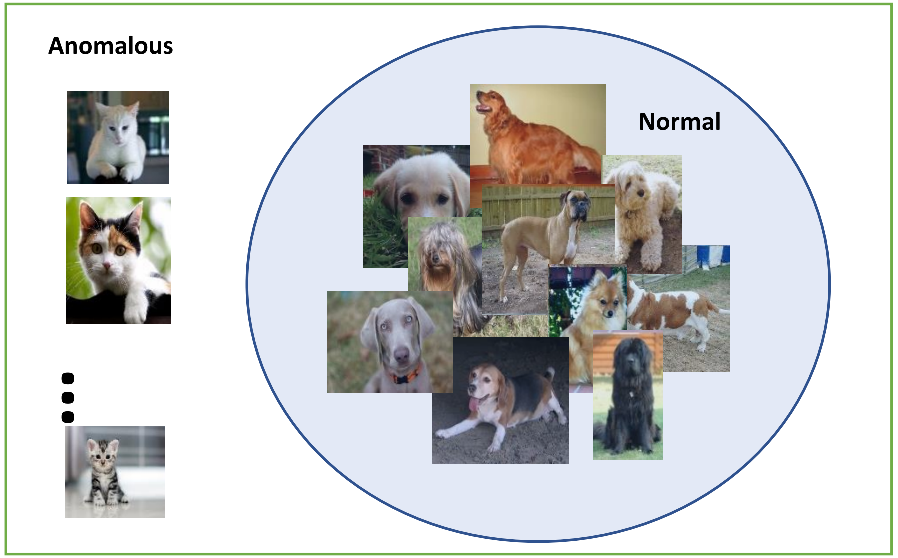
$$\sum_{n=0}^{N}\sum_{i=1}^{M}\cancel{\sum_{p\in H\,x\,W}}[D_x^{n*}(T_i(x_{test}^{n}))\cancel{_p}]_i$$

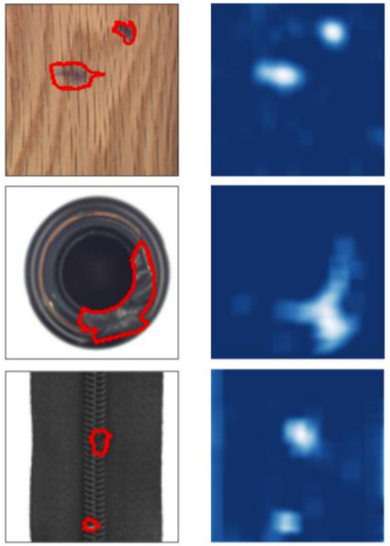
One Shot Defect Localization





Manipulating Structure by Understanding Structure





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

Automatically predict "speediness"

Uniform Speed Up (2x)

Adaptive speed up (2x)

Other Applications:

- Self-supervised action recognition
- Video retrieval

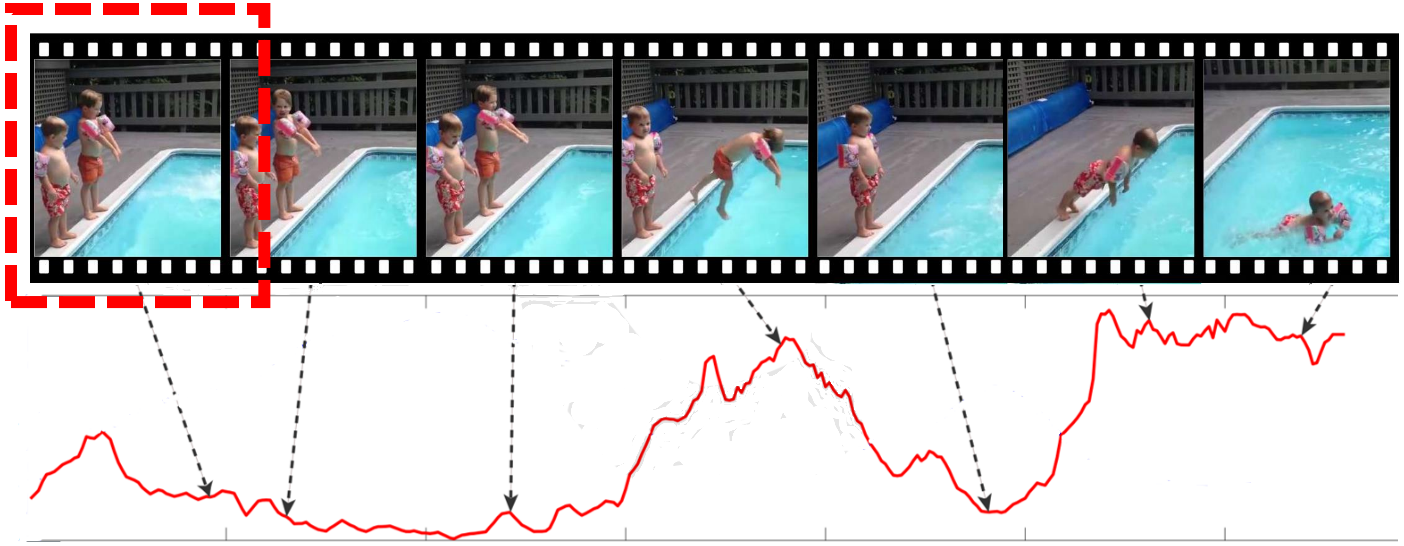
SpeedNet

Self-supervised training

Input video

Sped Up

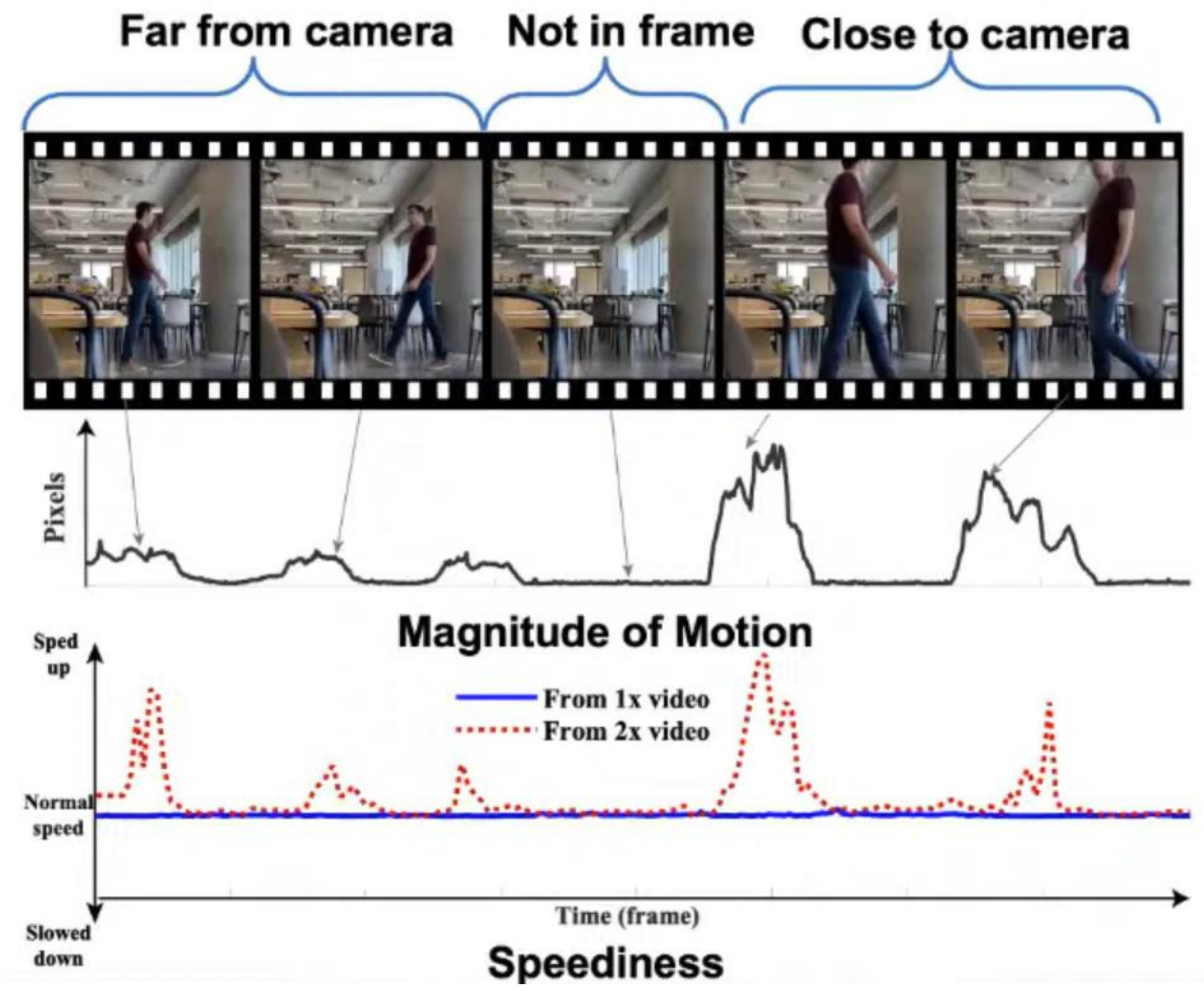
Inference on full **sped-up** video



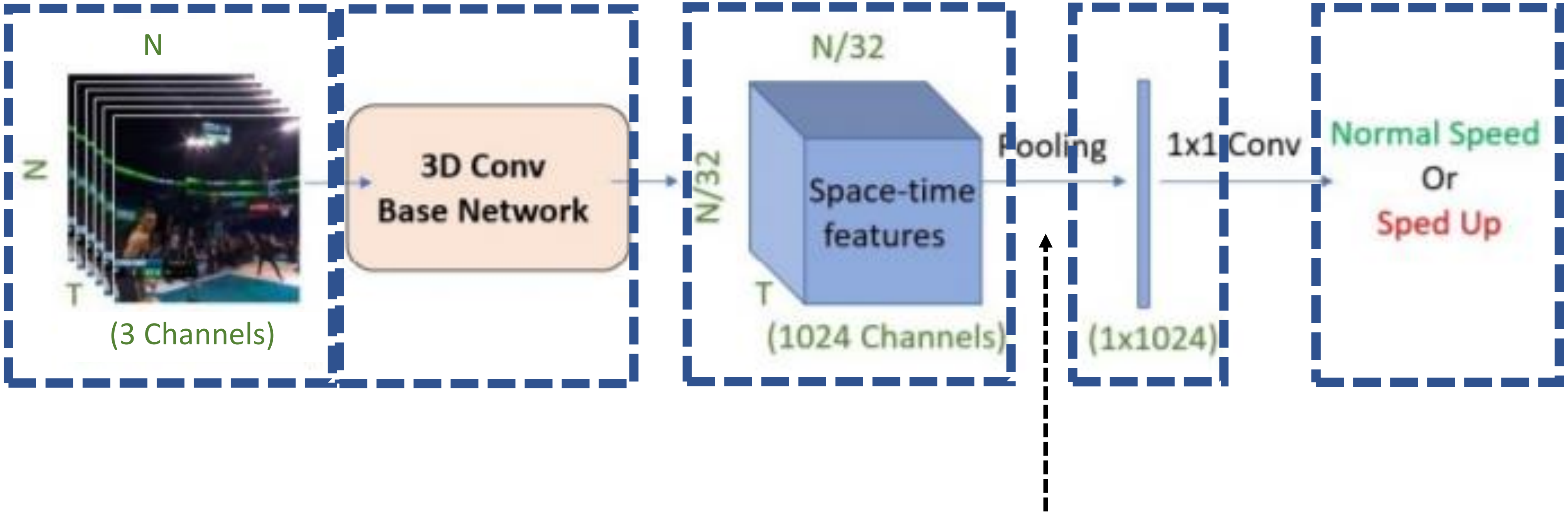
Sped-up

Normal speed

SpeedNet ≠ Motion Magnitude



Training SpeedNet



Spatial Max Pooling
Temporal Average Pooling

`

Training SpeedNet: Artificial Cues

- Spatial augmentations.
- Temporal augmentations
- Same-batch training.

Spatial Augmentations

- Fully convolutional network
- Random resize between 64 to 336
- Blurring helps mitigate potential pixel intensity jitter caused by MPEG or JPEG compression

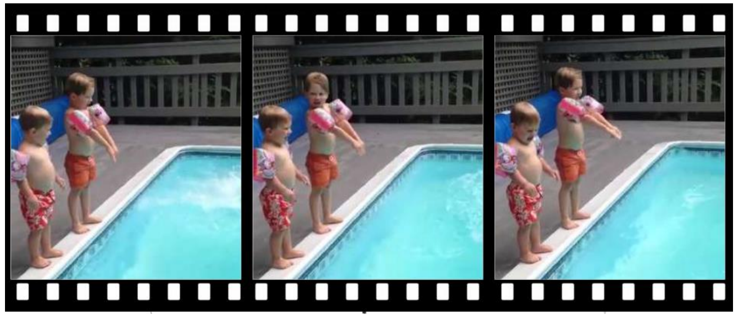
Temporal Augmentations

- 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

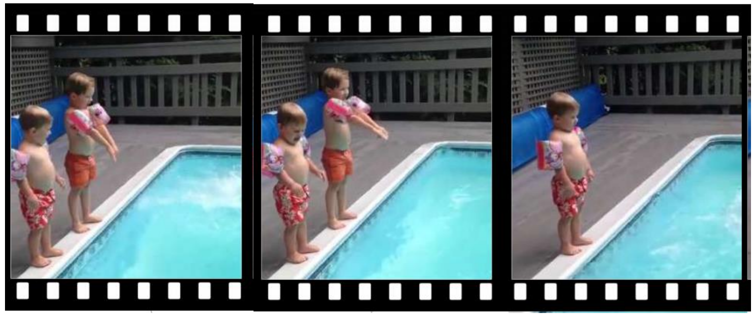
Same Batch

Normal speed



…

Speed up



…

`

Training SpeedNet: Artificial Cues

- NFS: Need For Speed dataset taken at 240 FPS

	Model Type		Accuracy	
Batch	**Temporal**	**Spatial**	**Kinetics**	**NFS**
Yes	Yes	Yes	75.6%	73.6%
No	Yes	Yes	88.2%	59.3%
No	No	Yes	90.0%	57.7%
No	No	No	96.9%	57.4%

No “Shortcuts” – A gap of 2%

“Shortcuts” – A gap of > 28%

From Speediness to Adaptive Speedup

Original 1x video

N videos of increasing speed

1x video (T frames)

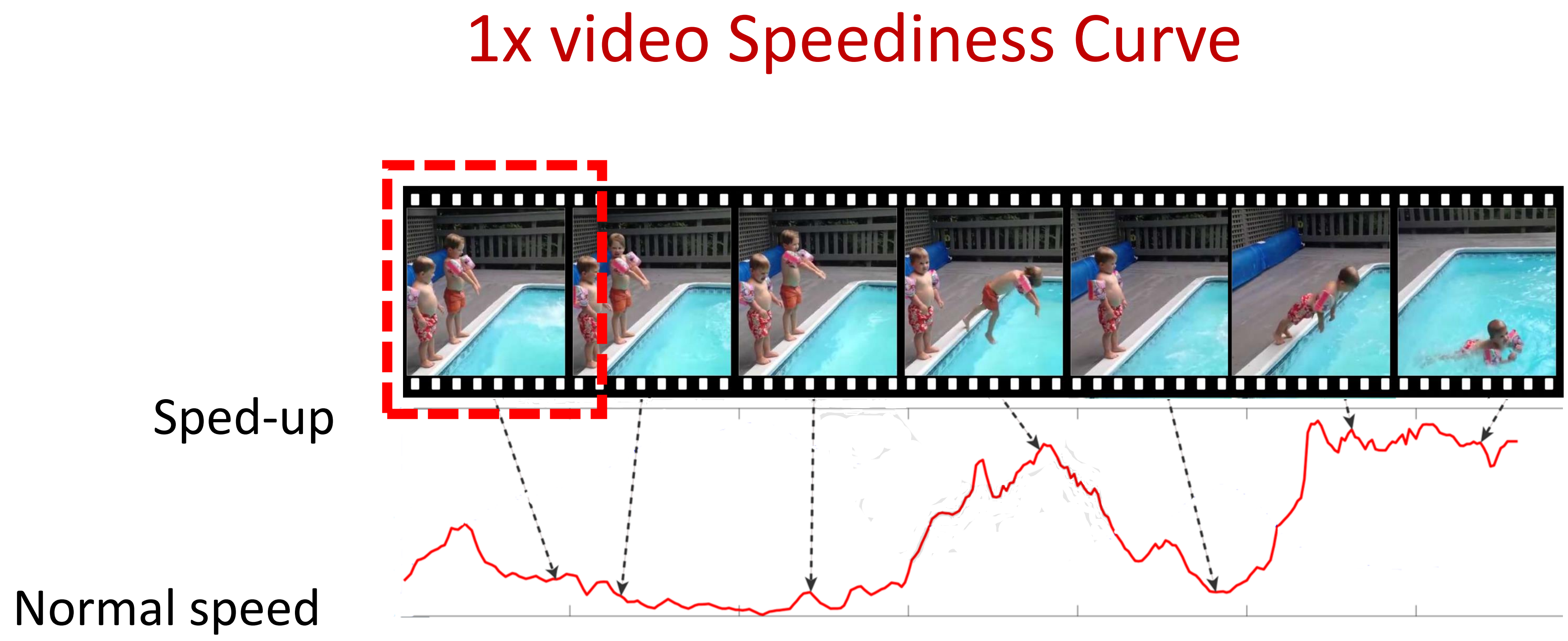
2x video (Interpolate to T Frames)

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

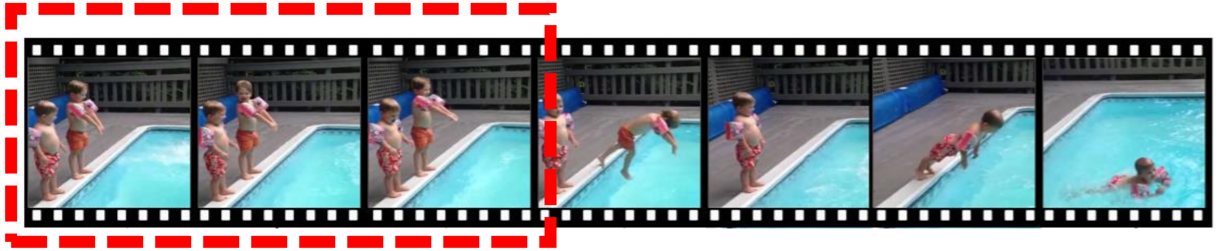
3x video Speediness Curve

...

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)

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

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

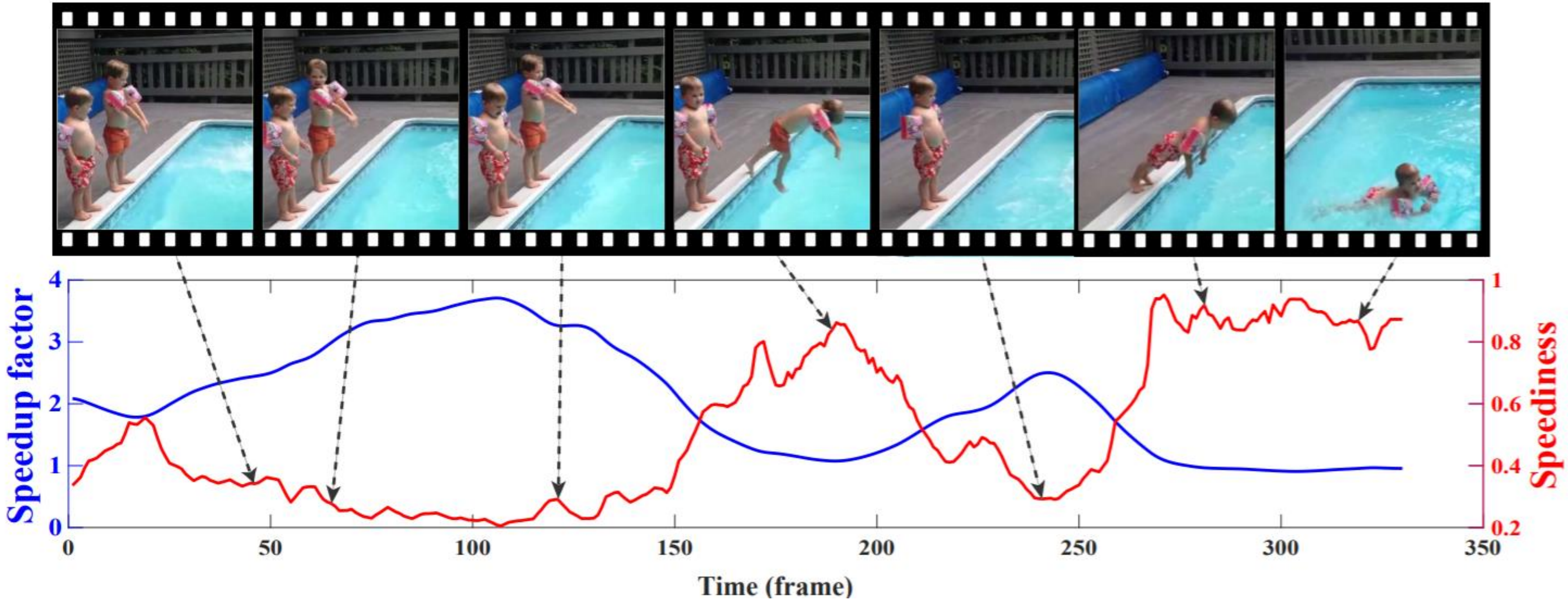
Final step: Estimate a smoothly varying speedup curve

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

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

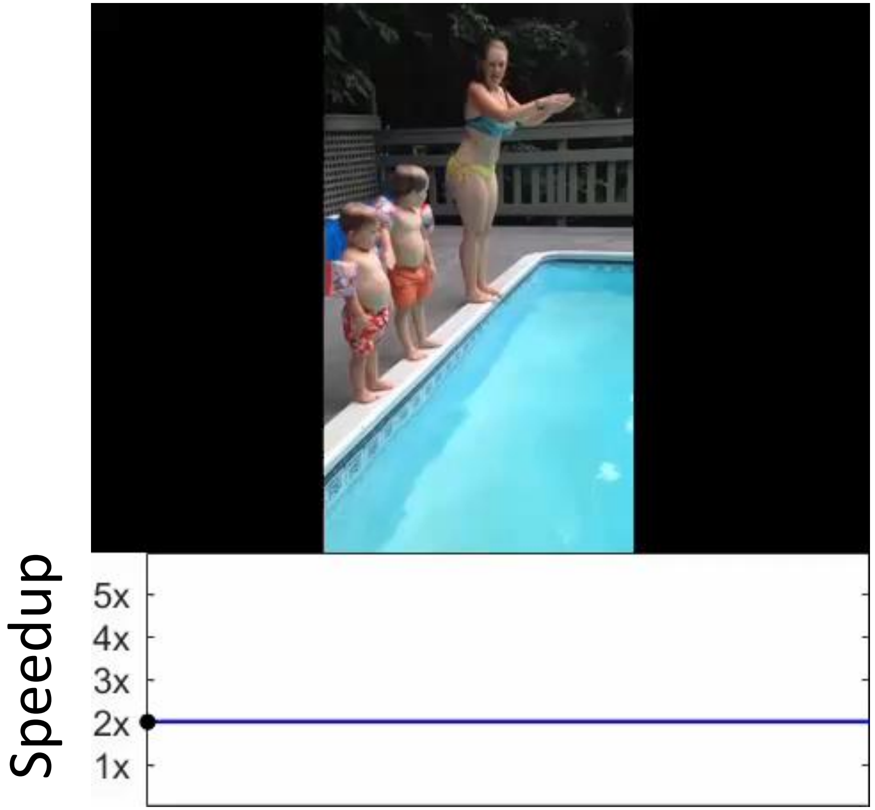
From Speediness to Adaptive Speedup

2x final "speediness curve" (blue):



Adaptive video speedup

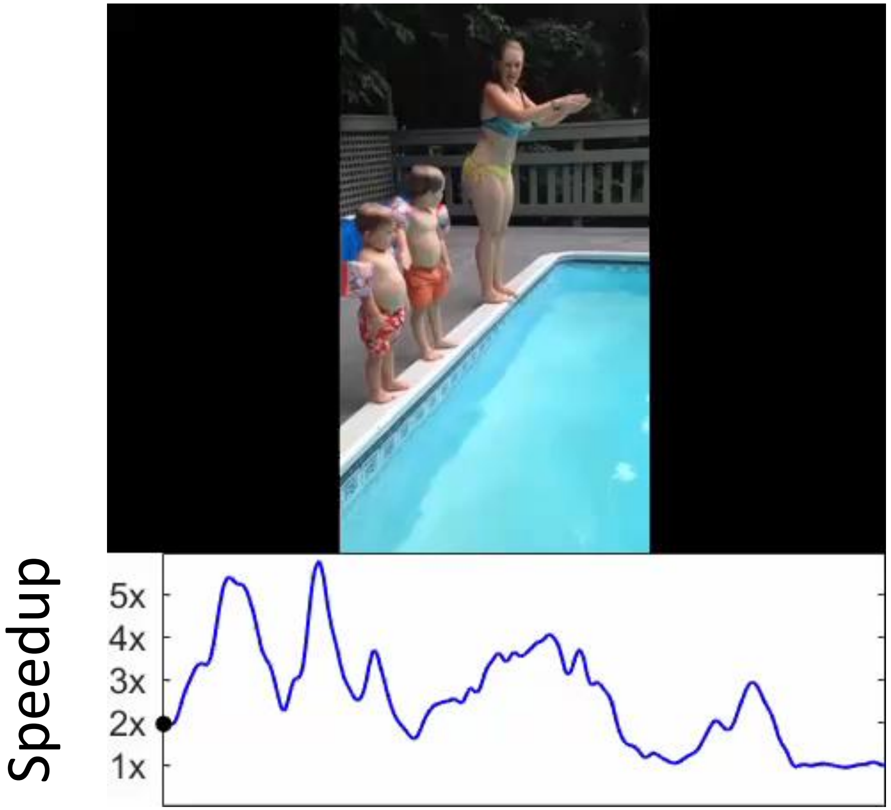
Total time = $\frac{1}{2}$ input time



Time

Uniform Speedup

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

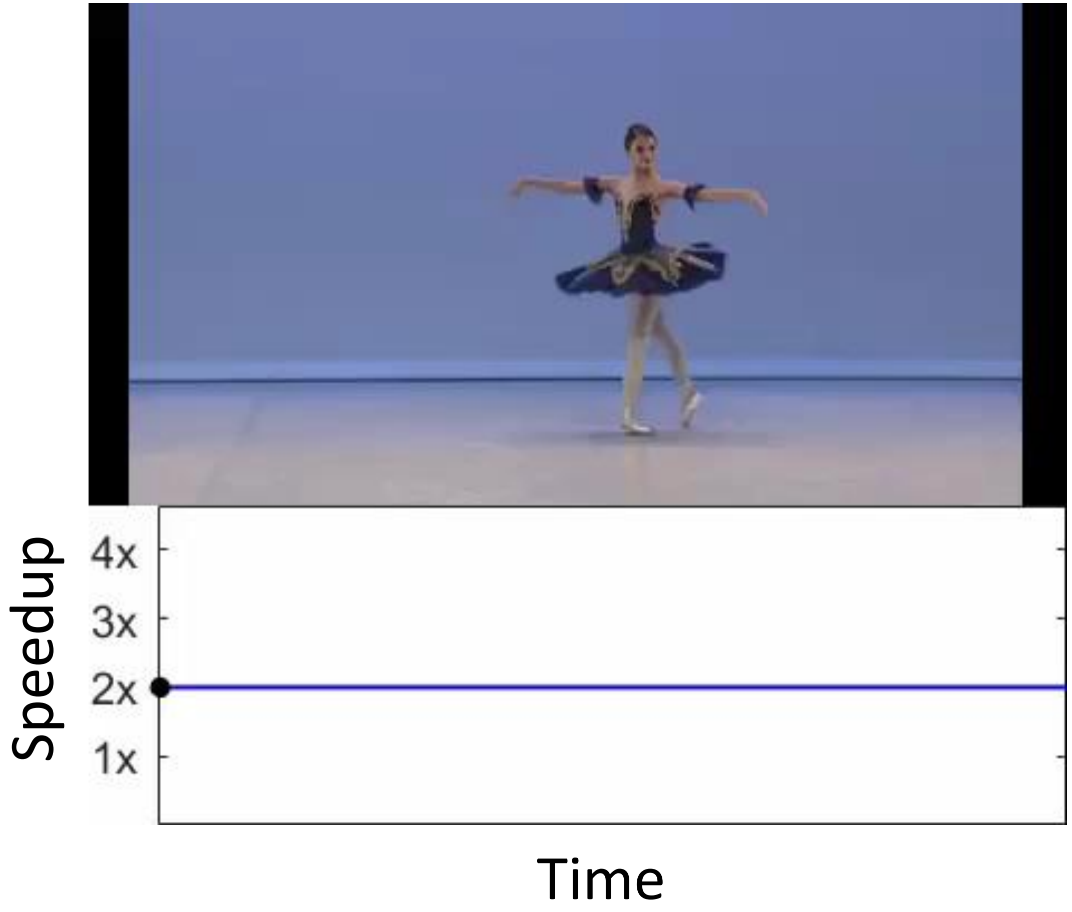


Time

Adaptive Speedup (ours)

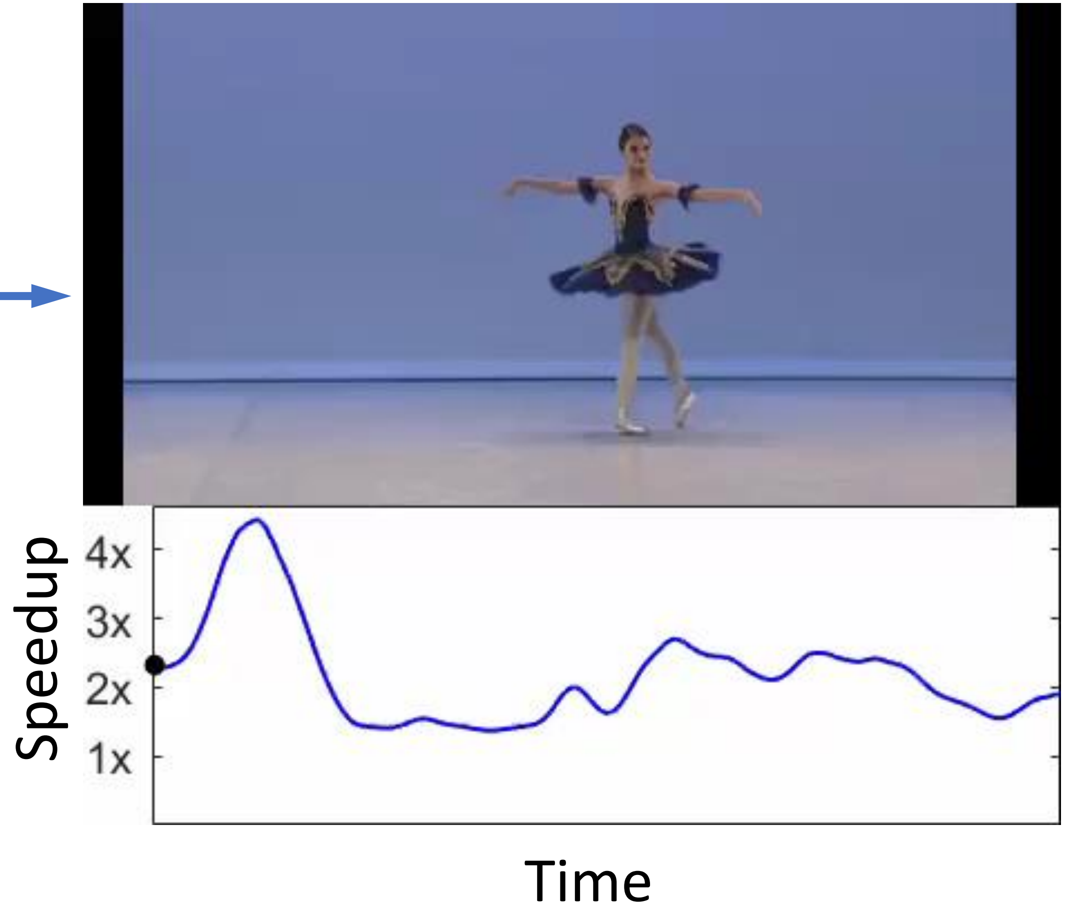
Adaptive video speedup

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



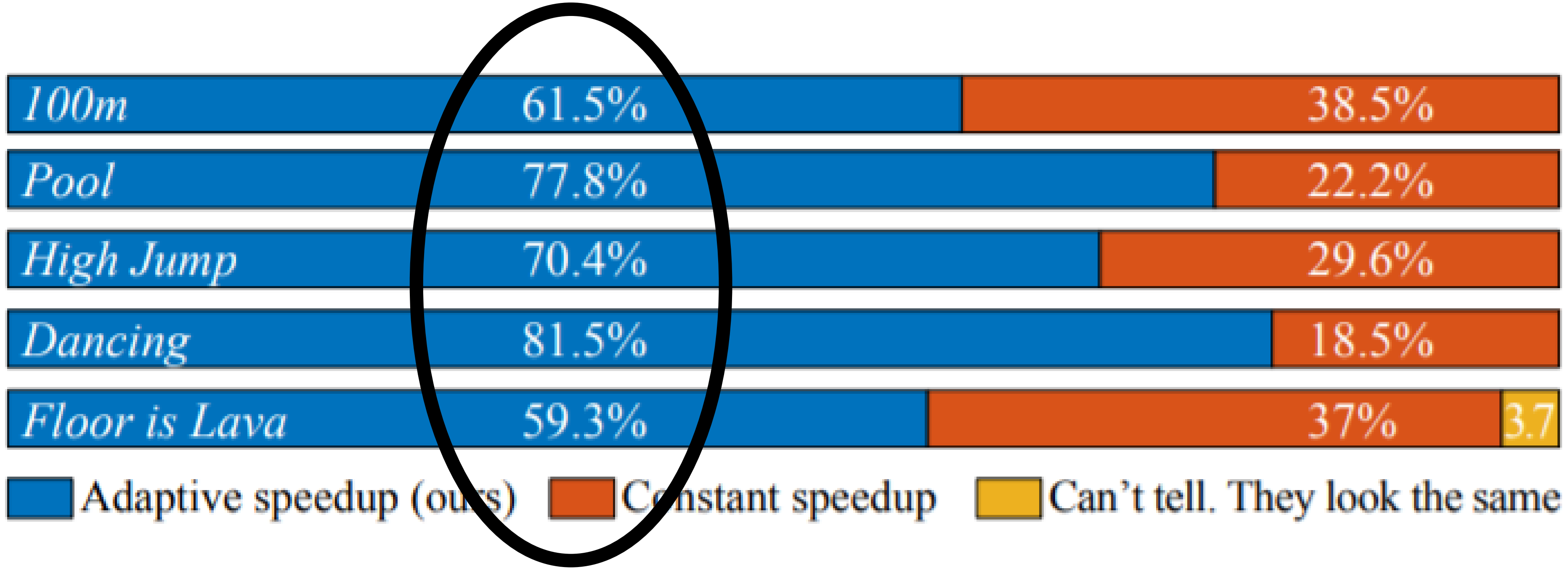
Uniform Speedup

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



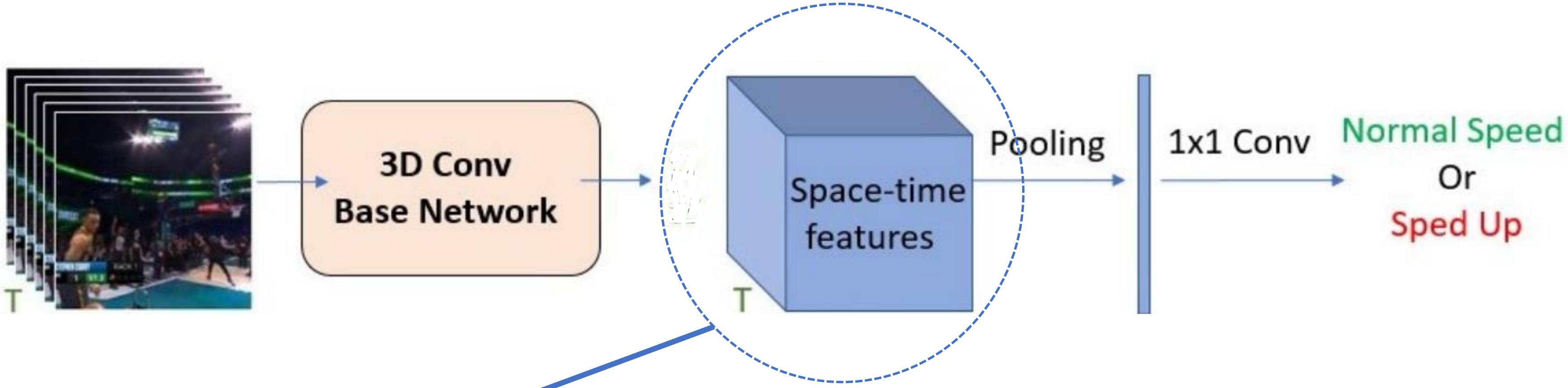
Adaptive Speedup (ours)

Adaptive Speedup Preferred in all videos of a user study



Other self supervised tasks

Train SpeedNet

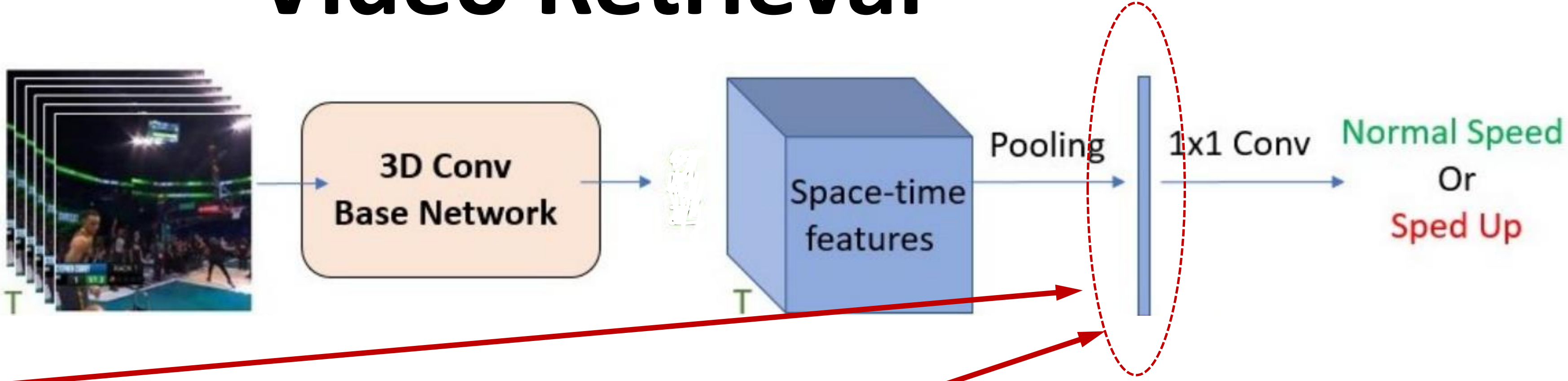


Self Supervised Action Recognition

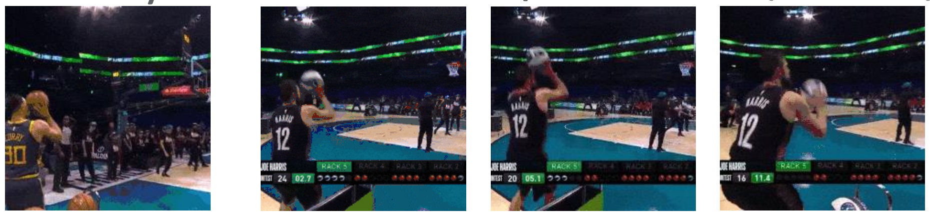
Initialization		Supervised accuracy	
Method	**Architecture**	**UCF101**	**HMDB51**
Random init	S3D-G	73.8	46.4
ImageNet inflated	S3D-G	86.6	57.7
Kinetics supervised	S3D-G	96.8	74.5
CubicPuzzle [19]	3D-ResNet18	65.8	33.7
Order [40]	R(2+1)D	72.4	30.9
DPC [13]	3D-ResNet34	75.7	35.7
AoT [38]	T-CAM	79.4	-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

Spatio-Temporal Visualizations

blue/green =
normal speed

yellow/orange =
slowed down

Manipulating Structure

- Multi-sample approaches
- Structural analogies
- Novel videos of similar structure
- Few shot anomaly detection

Manipulating by Understanding Structure

- Speed up videos "gracefully" using "speed" as supervision
- Image classification and domain adaptation by reducing bias towards global statistics (CVPR 2021)

Structure is Key to **Image Understanding**

Demonstrate using **Structure Aware Manipulation**

Next?

- 3D-aware structure manipulation
- Manipulating multiple objects from multiple scenes
- Functional relationships: A person riding a bike vs a person beside a bike

Thank You! Questions?