Génération d'images par entrainement de réseaux adversaires

PEYRESQ SUMMER SCHOOL

Camille Couprie, Facebook AI research

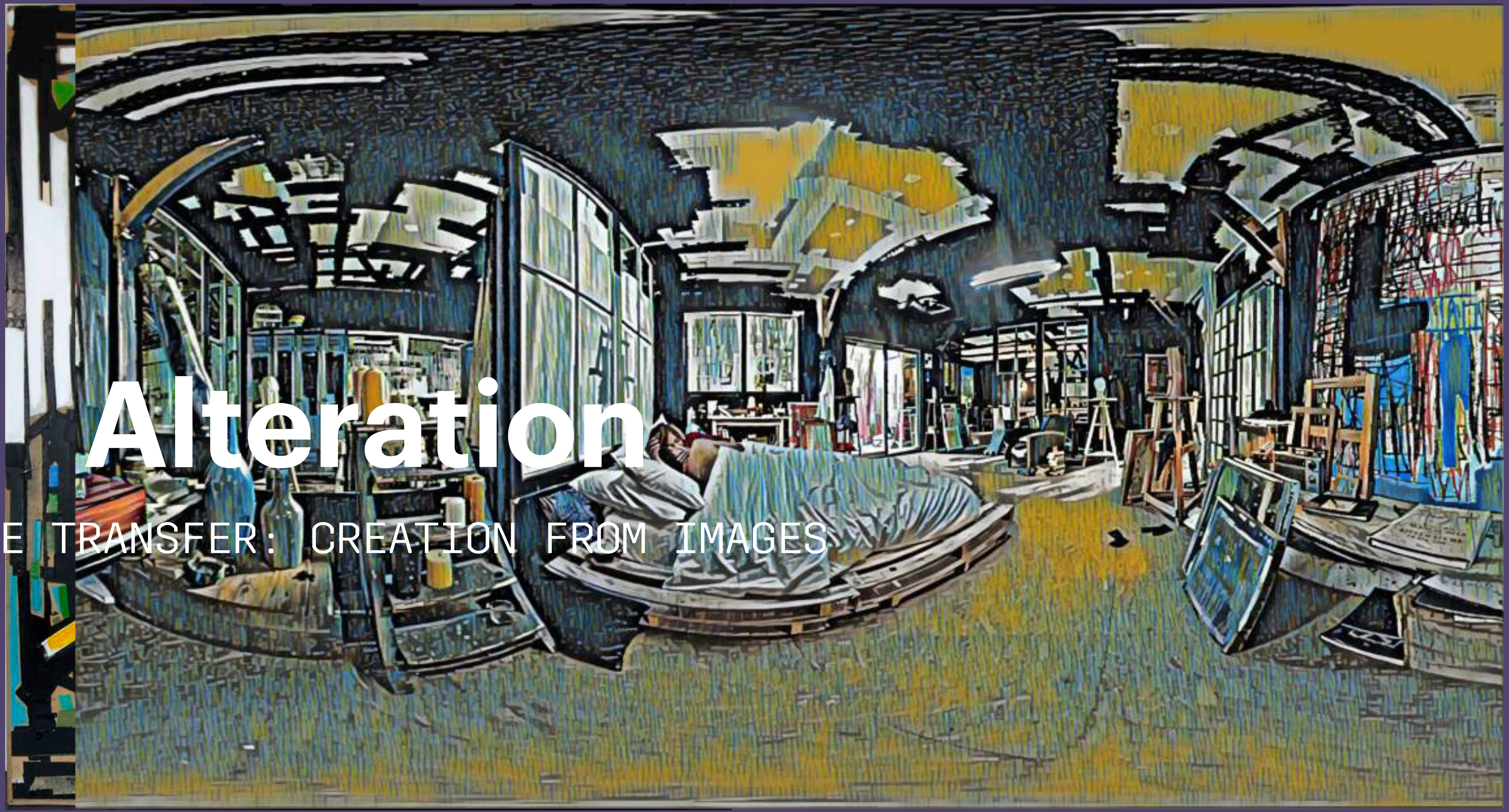
Introduction

F8 2018

DesIGN: Design Inspiration from Generative Networks

AI AND CREATIVITY

Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, Camille Couprie



Alteration

STYLE TRANSFER: CREATION FROM IMAGES

Source Image

Agenda

1 Introduction

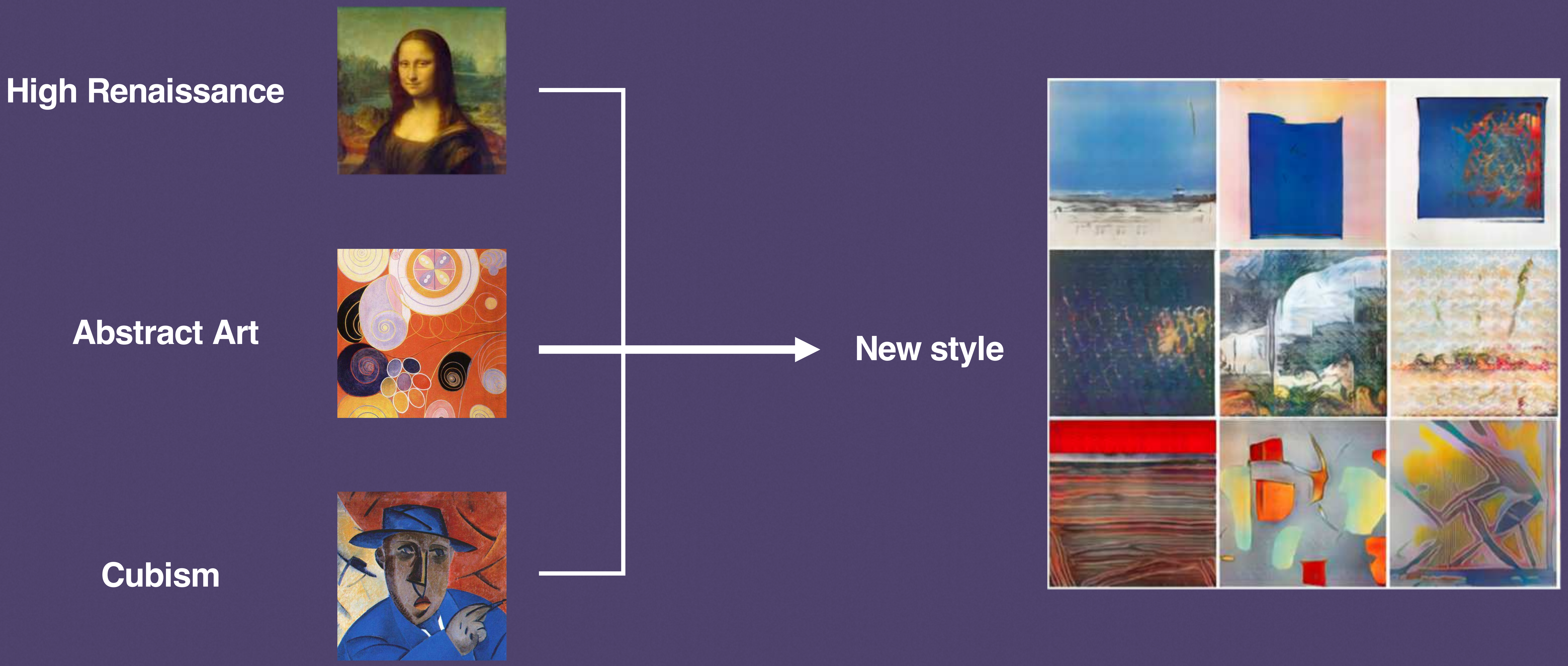
2 Shape and Texture Creativity

3 Conditioning on Shapes

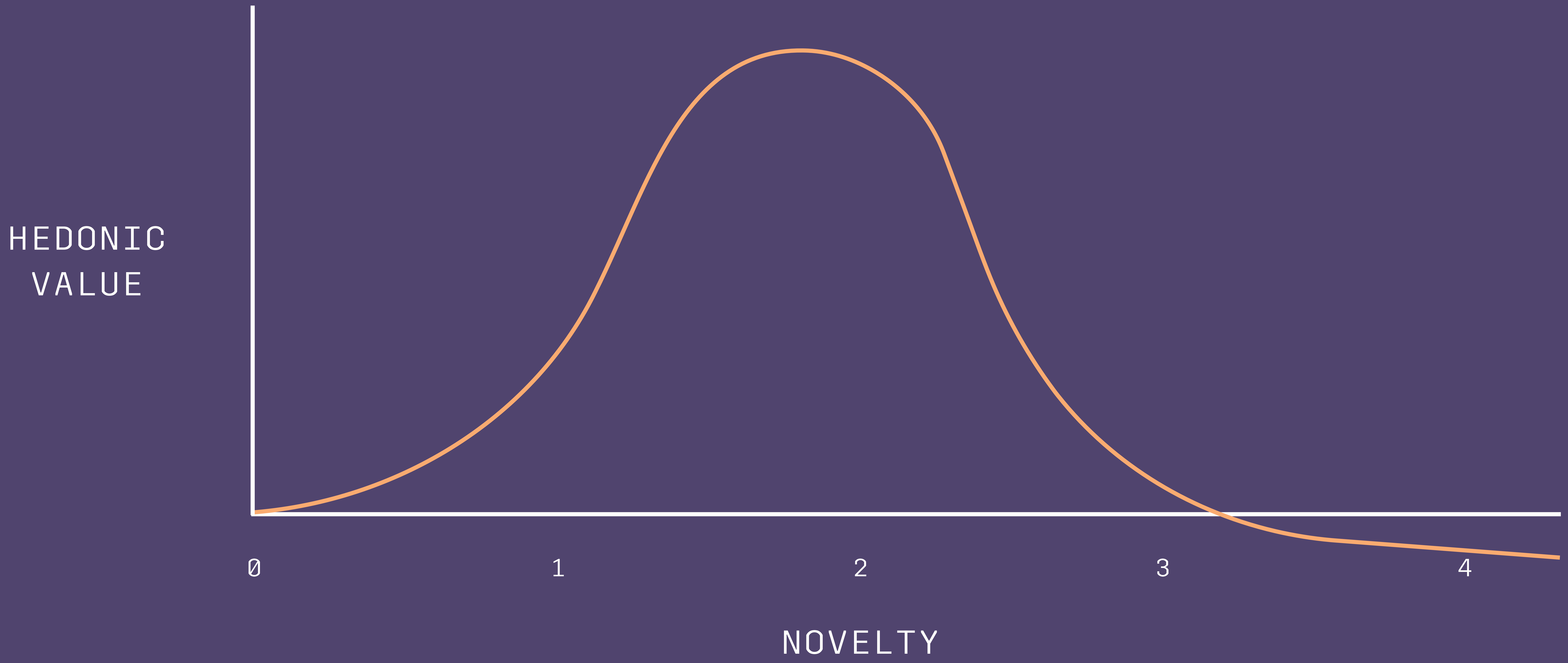
4 Evaluation

Creative Adversarial Networks (CAN)

CREATION FROM RANDOM NUMBERS



Principle of least effort: Wundt curve



When AI meets fashion...

Motivations for this project

FACEBOOK

- FAIR: Advance state of the art in machine intelligence
- Unlocking ways for AI to enhance creativity could enable new ways for people to express themselves creatively

FASHION BRANDS

- Create unexpected products
- Exploit data library of past collections to propose new products consistent with the brand DNA.
- Acquire new expertise

RELATED WORK 1

A GENERATIVE MODEL OF PEOPLE IN CLOTHING, CHRISTOPH LASSNER ET AL.

TOWARD BETTER RECONSTRUCTION OF STYLE IMAGES WITH GANS. ALEXANDER LORBERT ET AL.

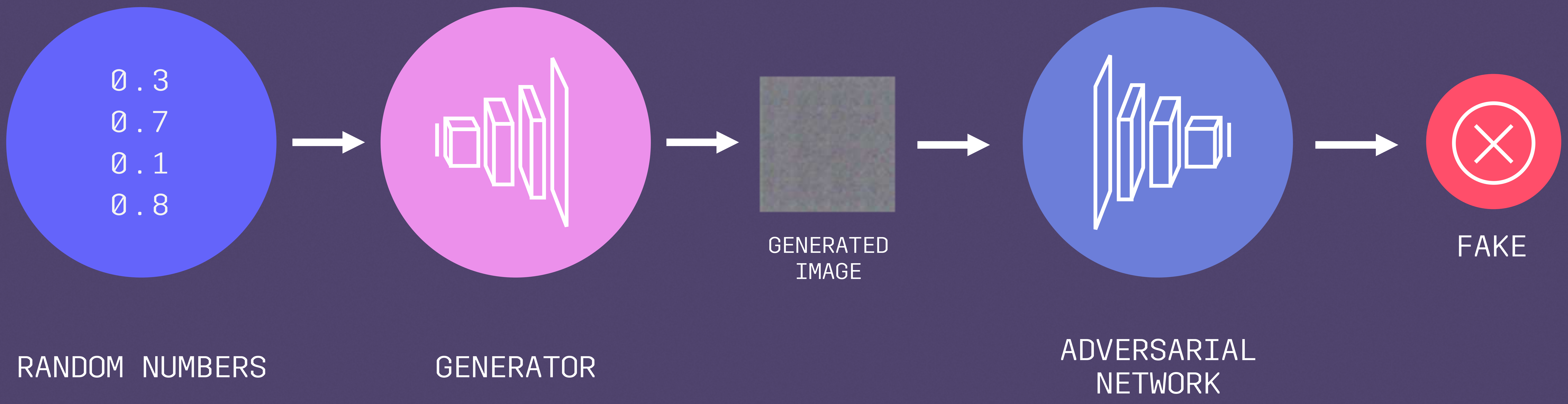
RELATED WORK 2

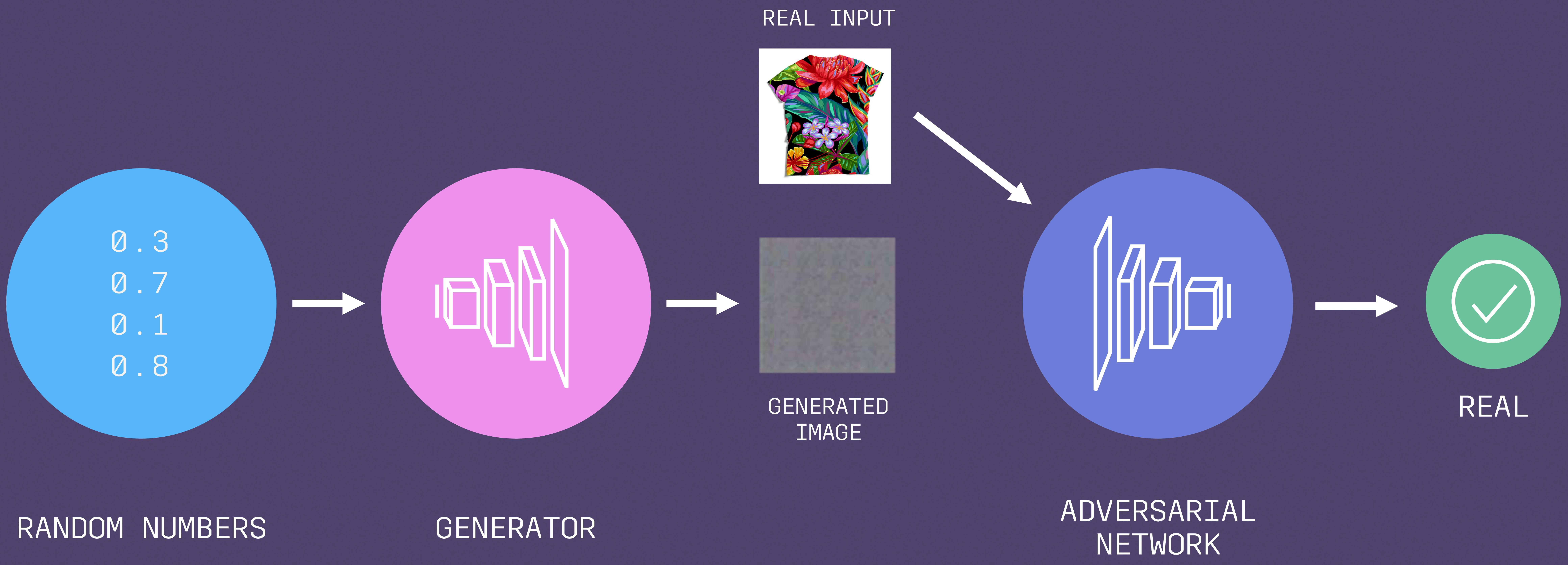
BE YOUR OWN PRADA: FASHION SYNTHESIS WITH STRUCTURAL COHERENCE. SHIZHAN ZHU ET AL.

PIX2PIX: **IMAGE-TO-IMAGE TRANSLATION**. PHILLIP ISOLA ET AL.,

DETERMINISTIC IMAGE TO IMAGE MAPPING

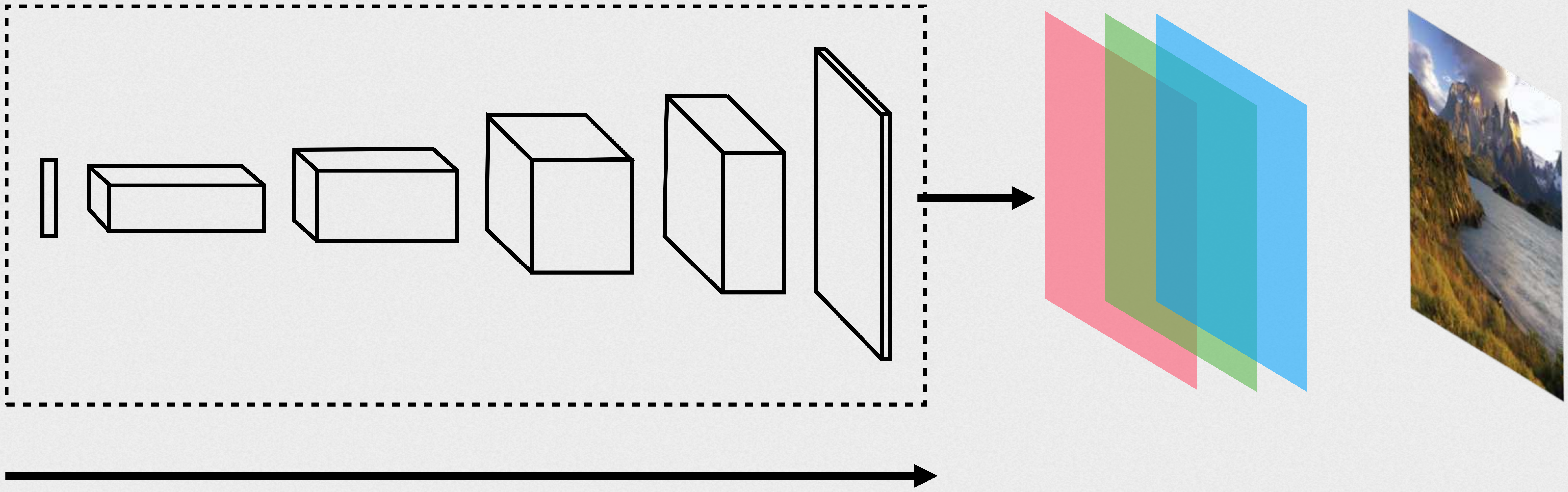
Generative Adversarial networks (GAN)

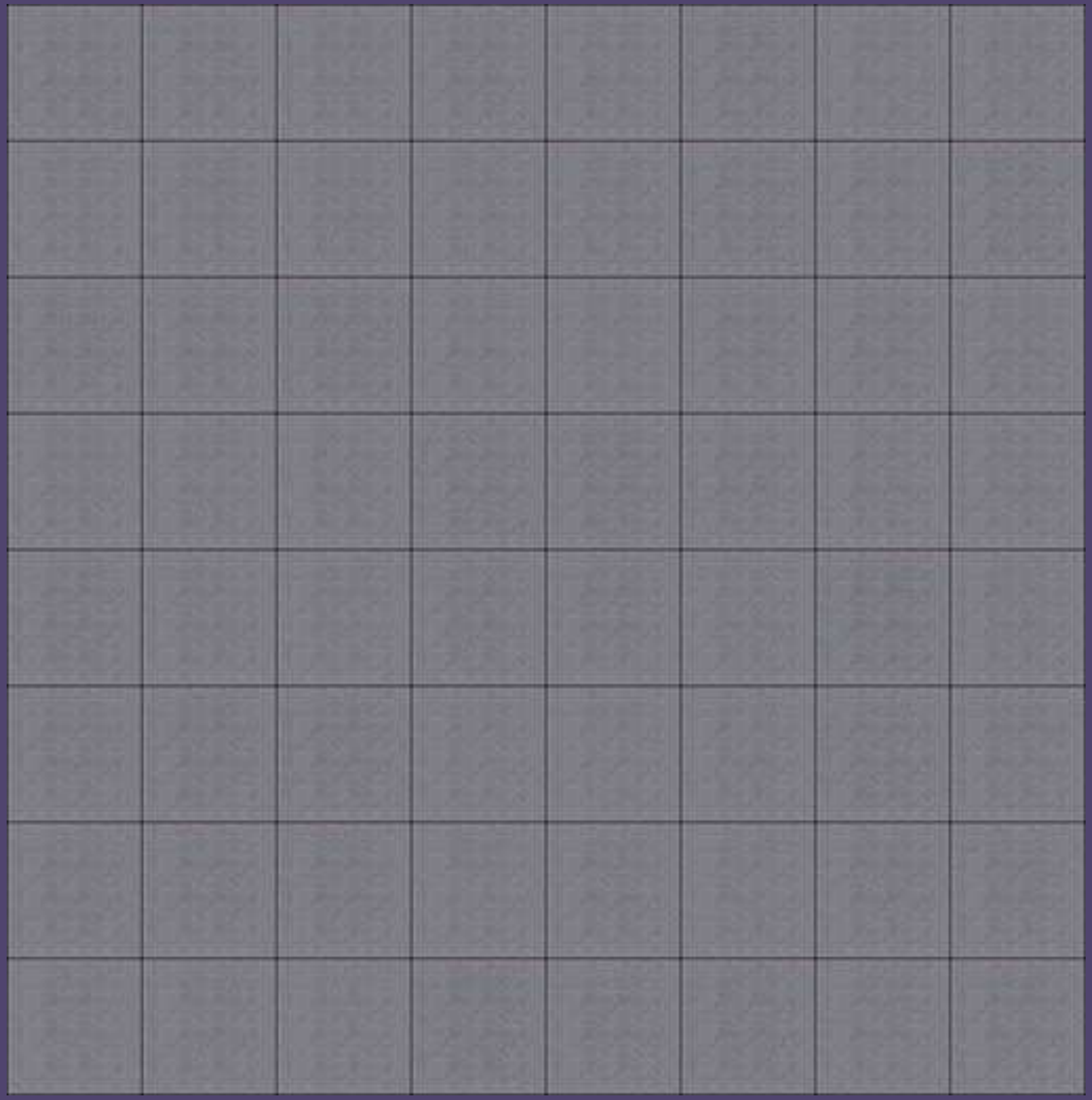

0.3
0.7
0.1
0.8
GENERATED IMAGE
FAKE
RANDOM NUMBERS
GENERATOR
ADVERSARIAL NETWORK


REAL INPUT
0.3
0.7
0.1
0.8
GENERATED
IMAGE
RANDOM NUMBERS
GENERATOR
ADVERSARIAL
NETWORK
REAL

Deep convolutional GANs

RADFORD ET AL : ICLR 2015

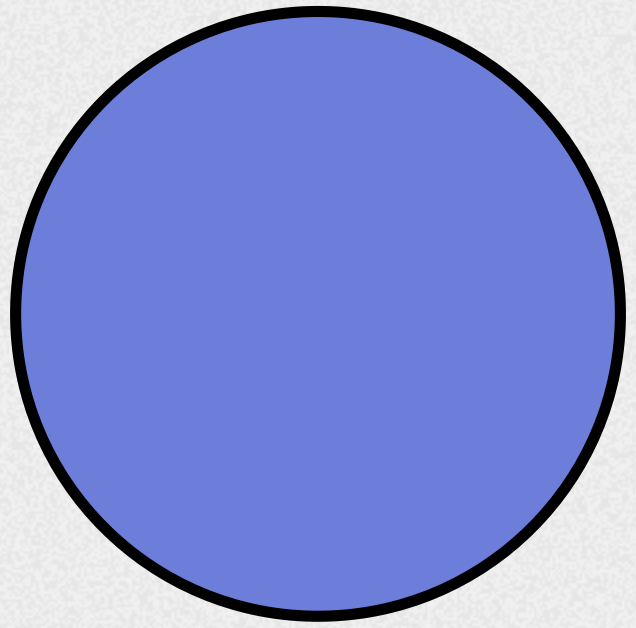




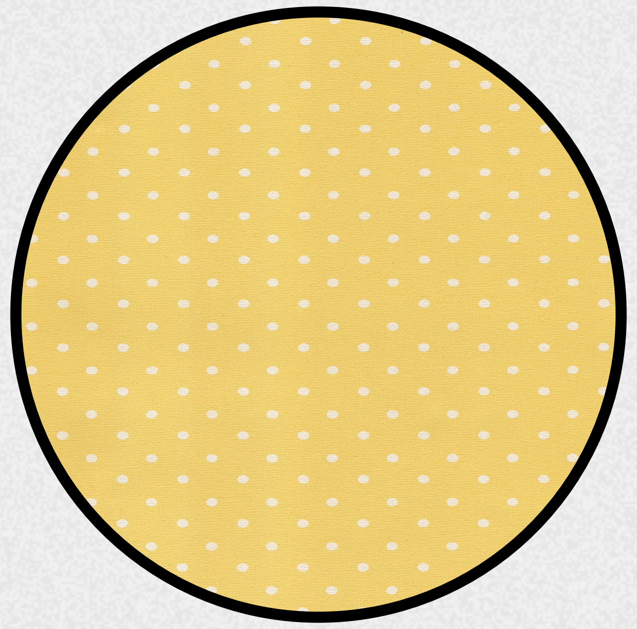
Training with pictures of about 2000 Clothing items

Shape and texture creativity

Texture classification



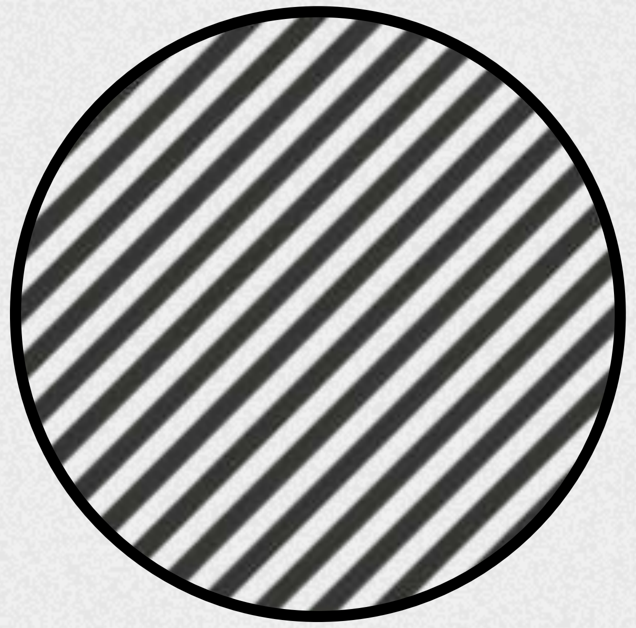
UNIFORM



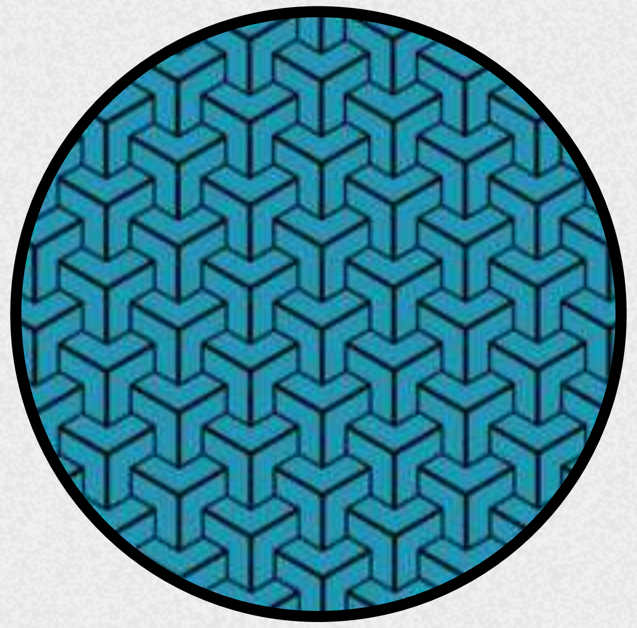
DOTTED

ANIMAL PRINT

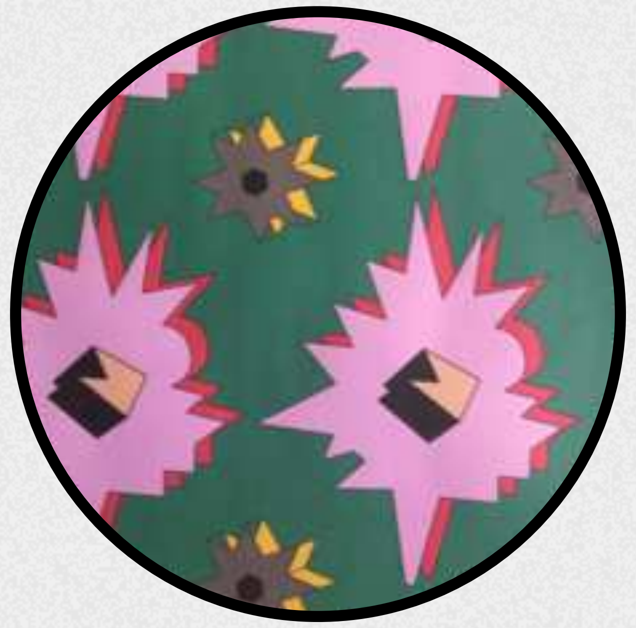
FLORAL



STRIPED



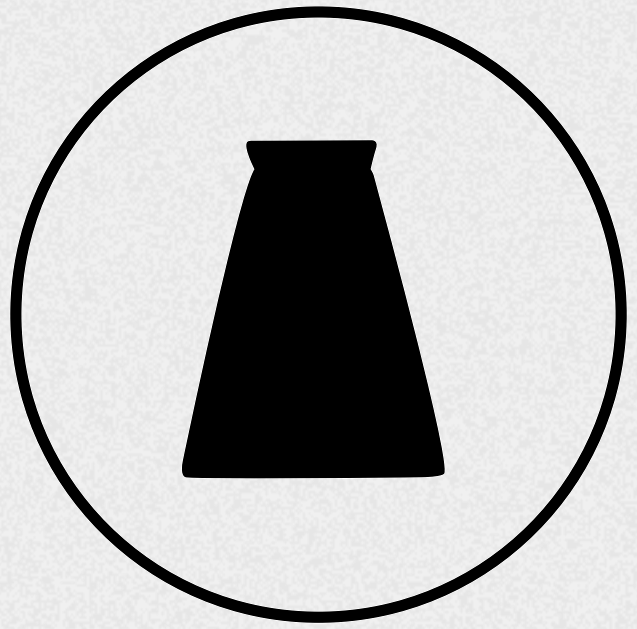
TILED



GRAPHICAL

Shape classification

DRESS



SKIRT

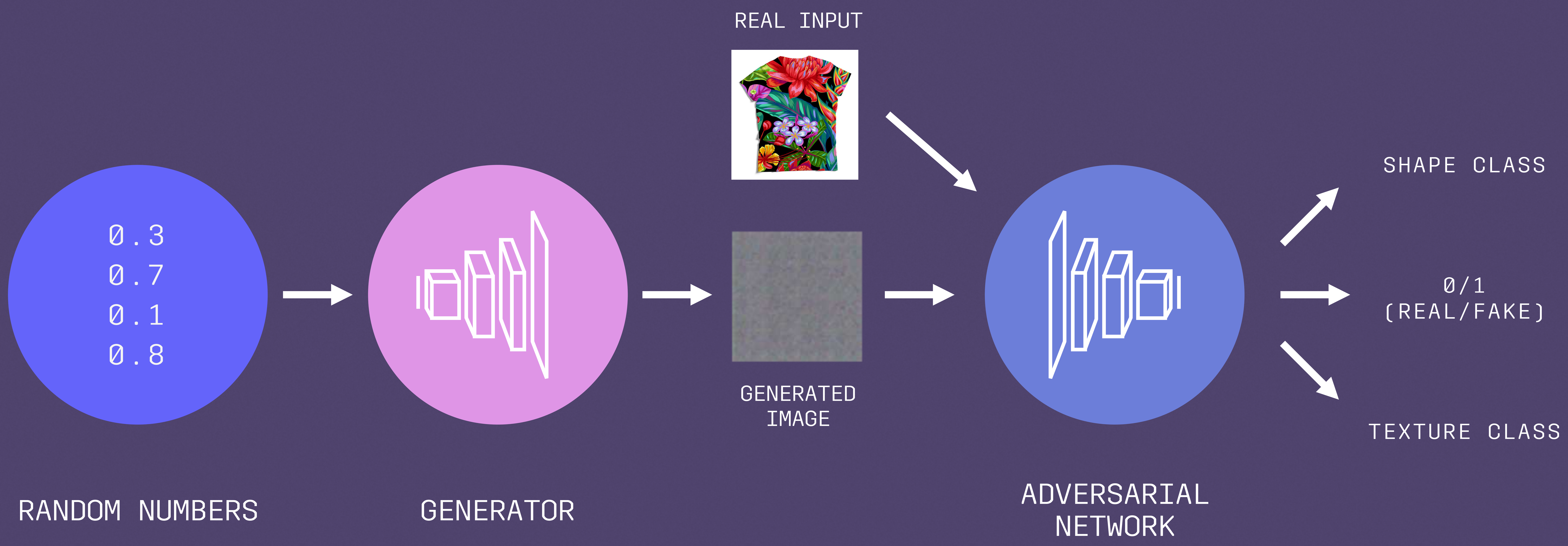
JACKET

PULLOVER

T-SHIRT

COAT

TOP



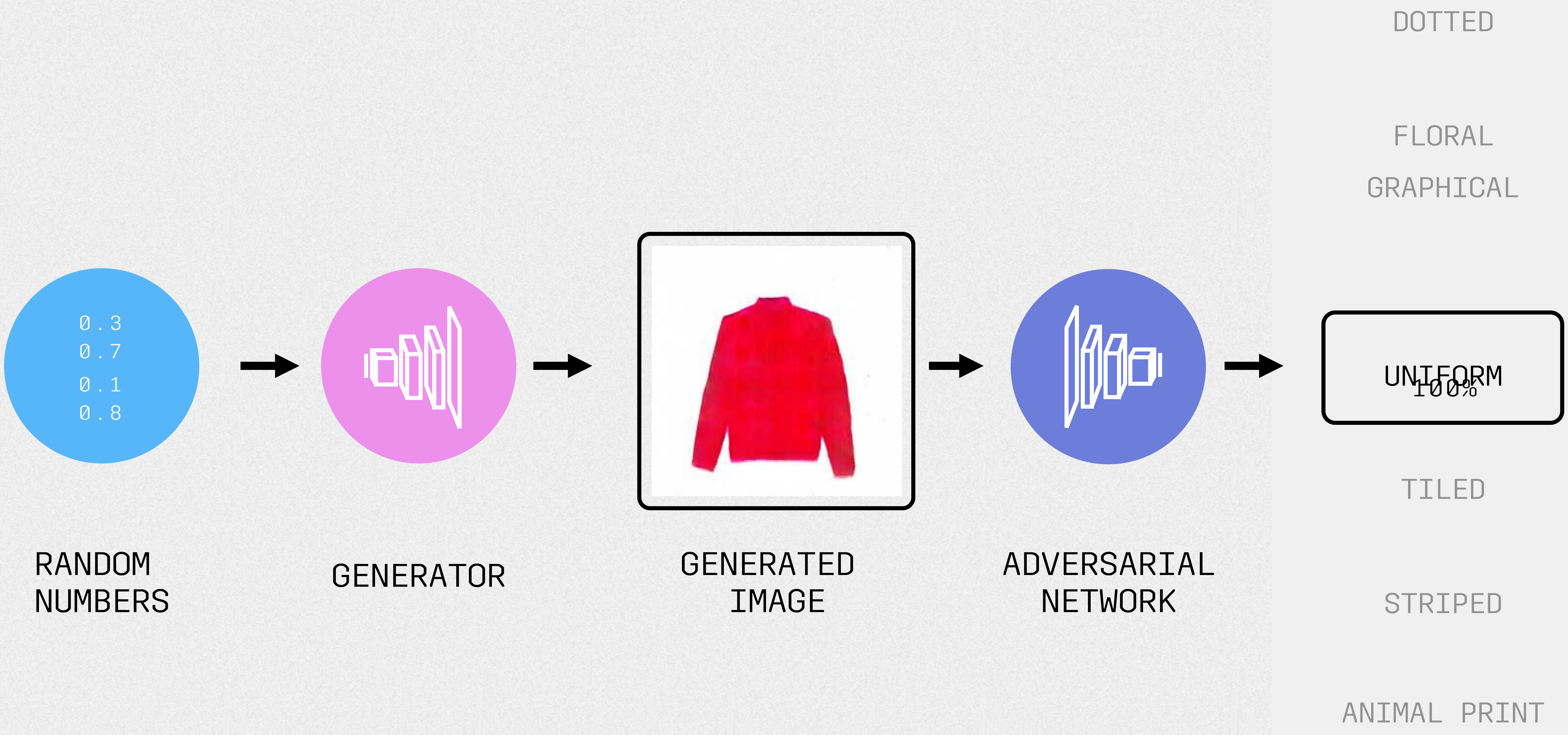
REAL INPUT
0.3
0.7
0.1
0.8
SHAPE CLASS
0/1
(REAL/FAKE)
TEXTURE CLASS
GENERATED
IMAGE
RANDOM NUMBERS
GENERATOR
ADVERSARIAL
NETWORK

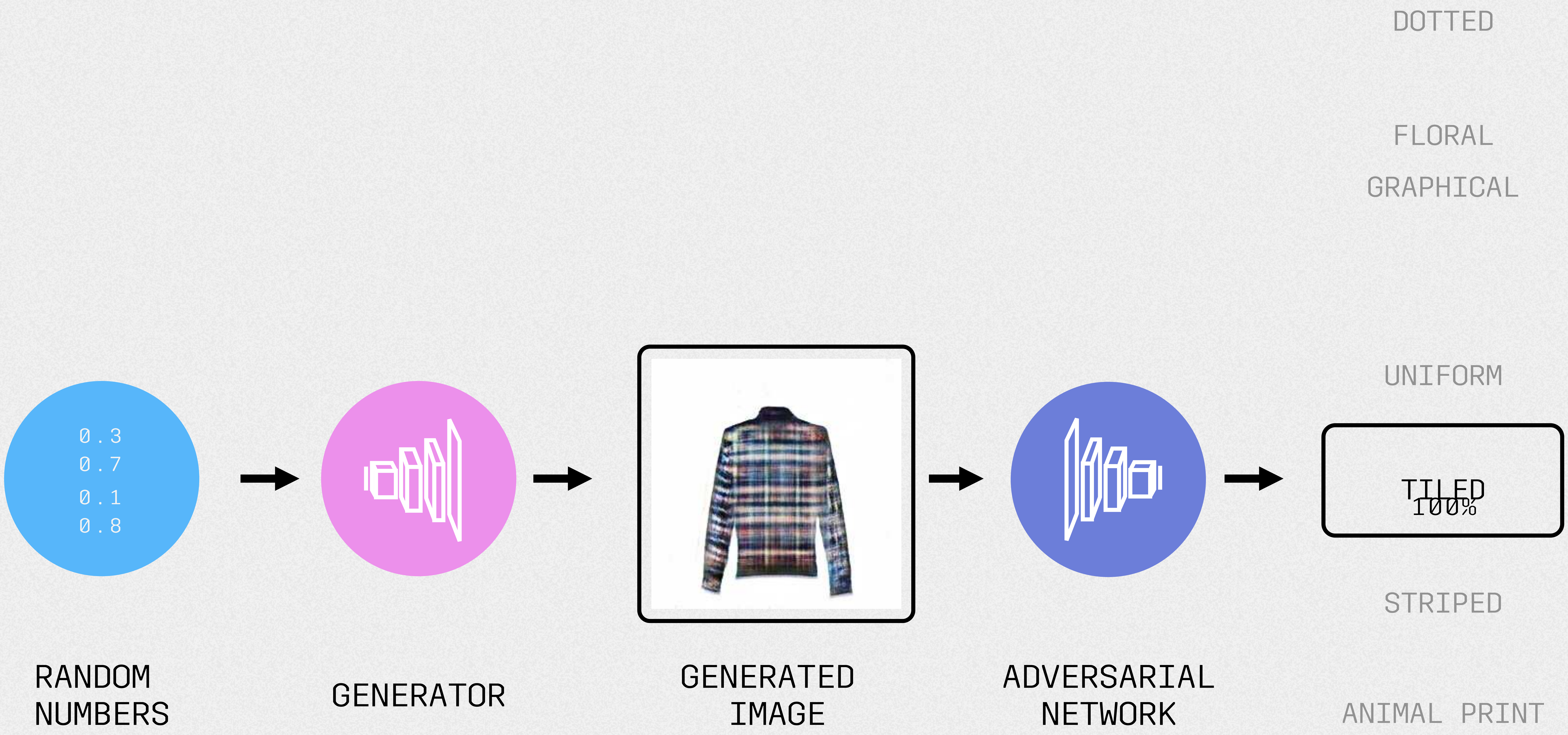
Before

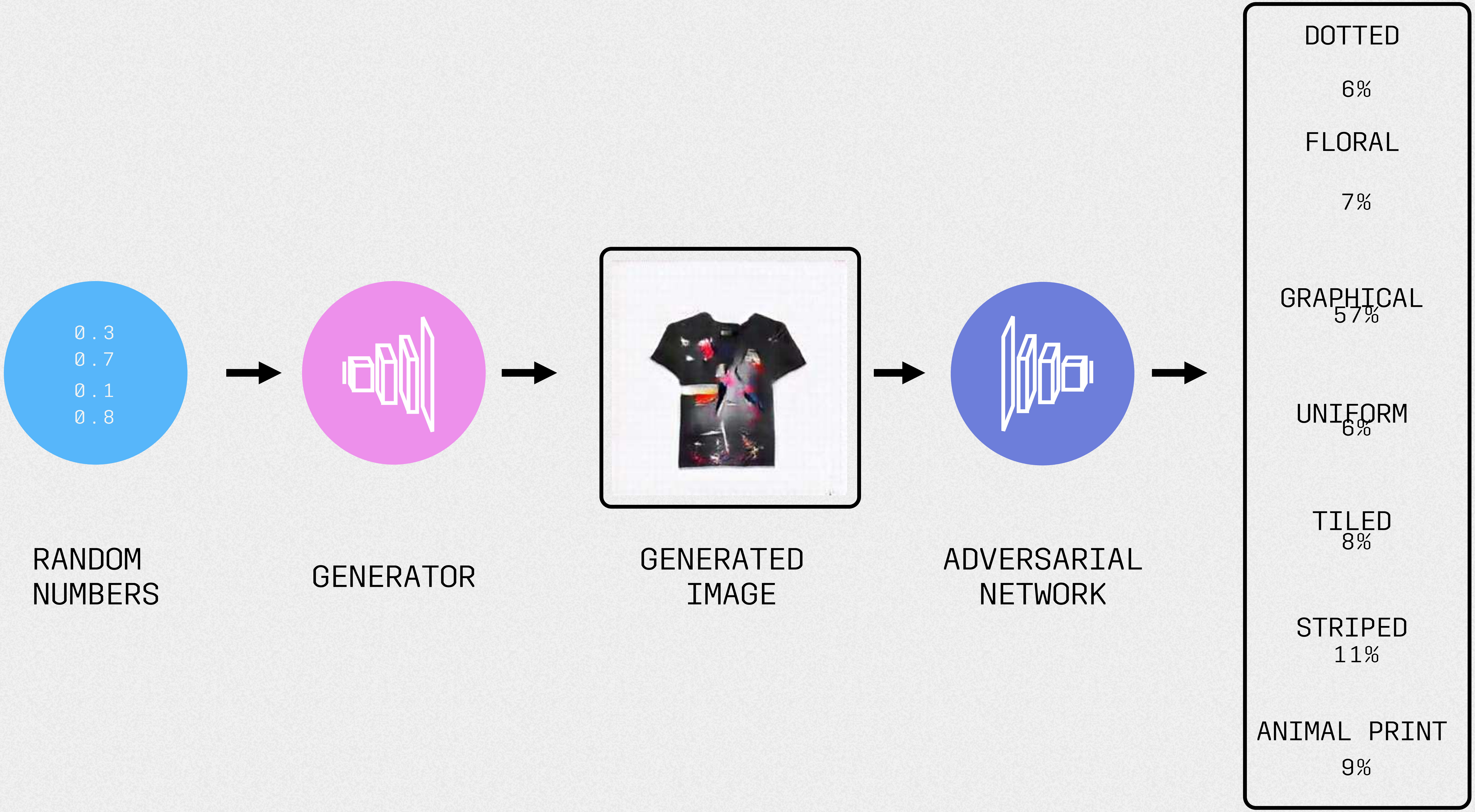
After

A holistic creativity Criterion

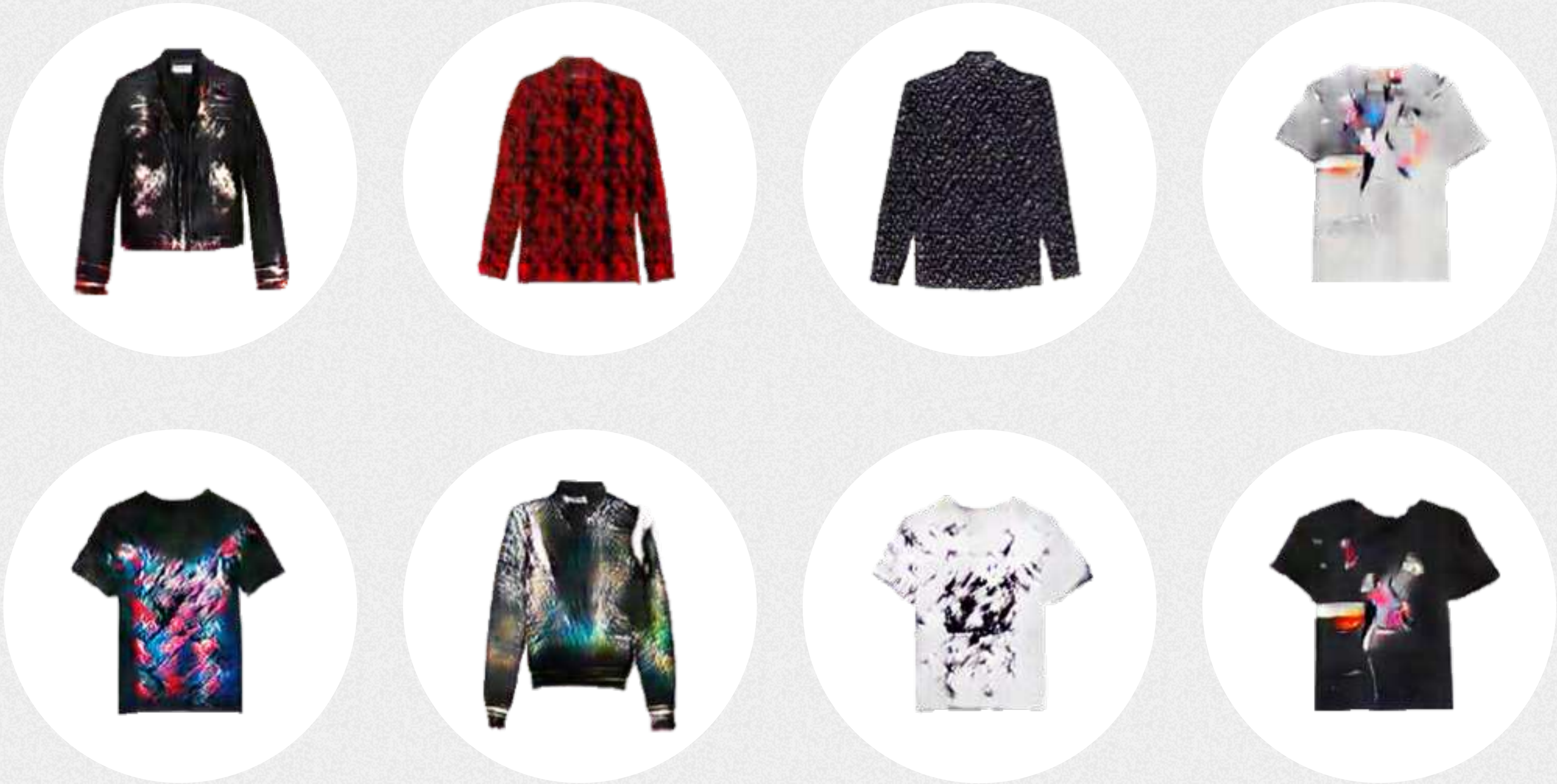
TO DEVIATE FROM EXISTING SHAPES AND TEXTURES


0.3
0.7
0.1
0.8
RANDOM NUMBERS
GENERATOR
GENERATED IMAGE
ADVERSARIAL NETWORK
DOTTED
FLORAL
GRAPHICAL
UNIFORM
100%
TILED
STRIPED
ANIMAL PRINT


DOTTED
FLORAL
GRAPHICAL
UNIFORM
0.3
0.7
0.1
0.8
TILED
100%
STRIPED
RANDOM NUMBERS
GENERATOR
GENERATED IMAGE
ADVERSARIAL NETWORK
ANIMAL PRINT


0.3
0.7
0.1
0.8
RANDOM NUMBERS
GENERATOR
GENERATED IMAGE
ADVERSARIAL NETWORK
DOTTED
6%
FLORAL
7%
GRAPHICAL
57%
UNIFORM
6%
TILED
8%
STRIPED
11%
ANIMAL PRINT
9%

With Creativity: CAN(H)



Optimization objectives

- **Generator's loss**

$$\min_{\theta_G} \mathcal{L}_{G\ \text{real/fake}} = \min_{\theta_G} \sum_{z_i \in \mathbb{R}^n} \log(1 - D(G(z_i)))$$

- **Discriminator's loss**

$$\min_{\theta_D} \mathcal{L}_{D\ \text{real/fake}} = \min_{\theta_D} \sum_{x_i \in \mathcal{D} z_i \in \mathbb{R}^n} -\log D(x_i) - \log(1 - D(G(z_i)))$$

- **auxiliary classifier discriminator:**

$$\mathcal{L}_D = \lambda_{D_r} \mathcal{L}_{D\ \text{real/fake}} + \lambda_{D_b} \mathcal{L}_{D\text{classif}}$$

- **Additional loss for the generator:**

$$\mathcal{L}_G = \lambda_{G_r} \mathcal{L}_{G\ \text{real/fake}} + \lambda_{G_e} \mathcal{L}_{G\ \text{creativity}}$$

CAN and CAN(H) losses

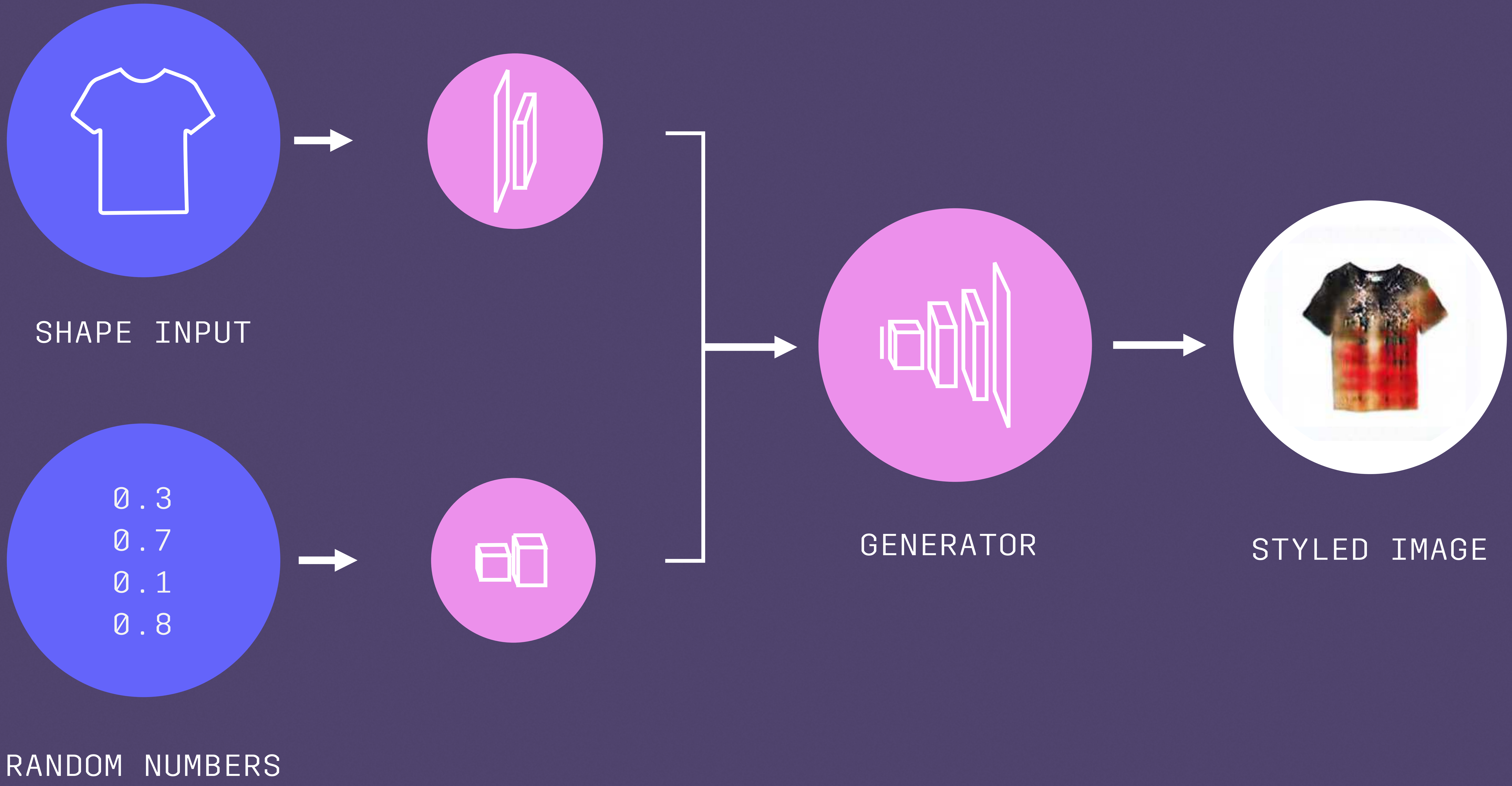
Binary cross entropy loss :

$$\mathcal{L}_{\text{CAN}} = -\sum_{k=1}^{K}\left[\frac{1}{K}\log(\sigma(D_{b,k}(x_i))) + (1-\frac{1}{K})\log(1-\sigma(D_{b,k}(x_i)))\right]$$

Multi-class cross entropy loss:

$$\begin{aligned}\mathcal{L}_{\text{CAN(H)}} &= -\sum_{x_i\in\mathcal{D}}\frac{1}{K}\log\operatorname{softmax}(D_b(x_i))\\ &= -\sum_{x_i\in\mathcal{D}}\frac{1}{K}\log\left(\frac{e^{D_{b,\hat{c}_i}(x_i)}}{\sum_{k=1}^{K}e^{D_{b,k}(x_i)}}\right)\end{aligned}$$

Conditioning on shapes



SHAPE INPUT
0.3
0.7
0.1
0.8
RANDOM NUMBERS
GENERATOR
STYLED IMAGE

Style GAN results

0.3	0.1	0.3
0.7	0.7	0.2
0.5	0.1	0.1
0.8	0.3	0.8



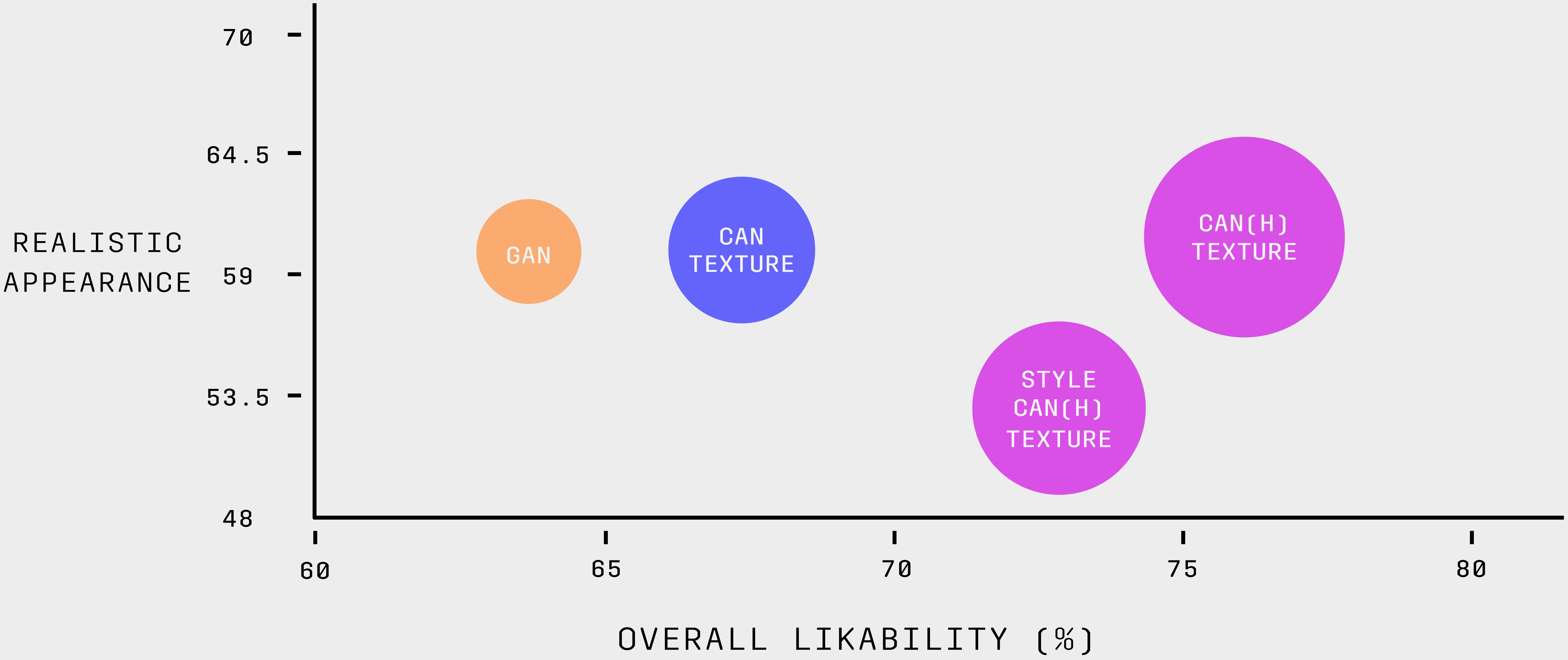






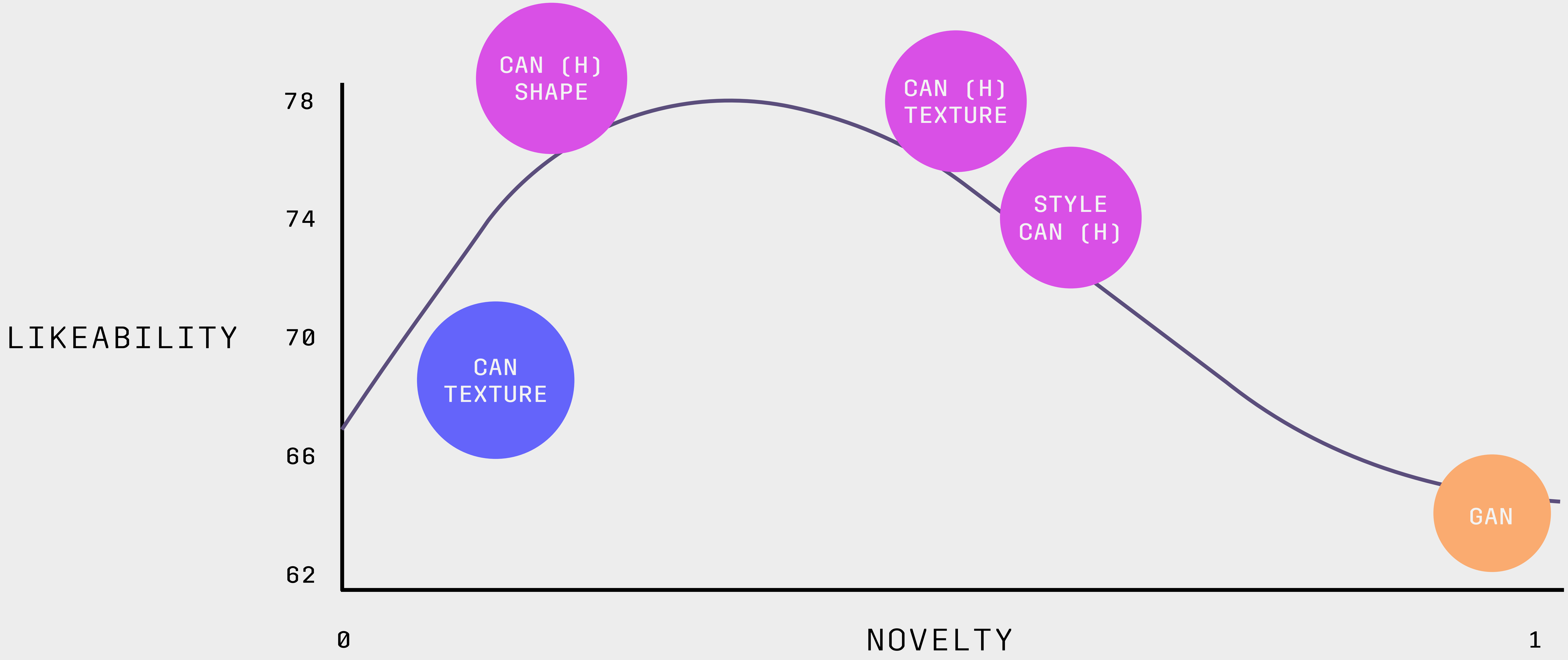

Evaluation

Human Evaluation Study



CAN: GAN WITH CREATIVITY LOSS, (H) STANDS FOR THE USE OF A HOLISTIC LOSS.

Creative Models are Most Popular



JUDGED BY HUMANS AND MEASURED AS A DISTANCE TO SIMILAR TRAINING IMAGES






"interesting" Shapes

Takeaways

Introduced Creative Image Modeling for a Non-Abstract Artistic Task

Creativity Criterion Lead to More Popular Results

Modeled Multiple Design Elements: Shape and Texture

What's next

Improve Stability of Generative Networks

Evaluation Remains an Open Research Problem

Higher Resolution

Factorization of Elements of Designs

Sbai, Elhoseiny, Bordes, LeCun, Couprie: ``DeSIGN: Design Inspiration from Generative Networks''

http://arxiv.org/abs/1804.00921

DesIGN: Design Inspiration from Generative Networks

Othman Sbai[1,2] Mohamed Elhoseiny[1] Antoine Bordes[1] Yann LeCun[1,3] Camille Couprie[1]

[1] Facebook AI Research
[2] École des Ponts, UPE
[3] New York University

Abstract

Can an algorithm create original and compelling fashion designs to serve as an inspirational assistant? To help answer this question, we design and investigate different image generation models associated with different loss functions to boost creativity in fashion generation. The dimensions of our explorations include: (i) different Generative Adversarial Networks architectures that start from noise vectors to generate fashion items, (ii) a new loss function that encourages creativity, and (iii) a generation process following the key elements of fashion design (disentangling shape and texture makers). A key challenge of this study is the evaluation of generated designs and the retrieval of best ones, hence we put together an evaluation protocol associating automatic metrics and human experimental studies that we hope will help ease future research. We show that our proposed creativity loss yields better overall appreciation than the one employed in Creative Adversarial Networks. In the end, about 61% of our images are thought to be created by human designers rather than by a computer while also being considered original per our human subject experiments, and our proposed loss scores the highest compared to existing losses in both novelty and likability.

1. Introduction

Artificial Intelligence (AI) research has been making huge progress in the machine's capability of human level understanding across the spectrum of perception, reasoning and planning [14, 1, 28]. Another key yet still relatively understudied direction is creativity where the goal is for machines to generate original items with realistic, aesthetic and/or thoughtful attributes, usually in artistic contexts. We can indeed imagine AI to serve as inspiration for humans in the creative process and also to act as a sort of creative

Figure 1: Training generative adversarial models with appropriate losses leads to realistic and creative 512 × 512 fashion images.

assistant able to help with more mundane tasks, especially in the digital domain. Previous work has explored writing pop songs [3], imitating the styles of great painters [9, 7] or doodling sketches [12] for instance. However, it is not clear how *creative* such attempts can be considered since most of them mainly tend to mimic training samples without expressing much originality.

Creativity is a subjective notion that is hard to define and evaluate, and even harder for an artificial system to optimize for. Colin Martindale put down a psychology based theory that explains human creativity in art [22] by connecting creativity or acceptability of an art piece to novelty with *"the principle of least effort"*. As originality increases, people like the work more and more until it becomes too novel and too far from standards to be understood. When this happens, people do not find the work appealing anymore because a lack of understanding and of realism leads to a lack of appreciation. This behavior can be illustrated by the Wundt curve that correlates the arousal potential (i.e. novelty) to hedonic responses (e.g. likability of the work)

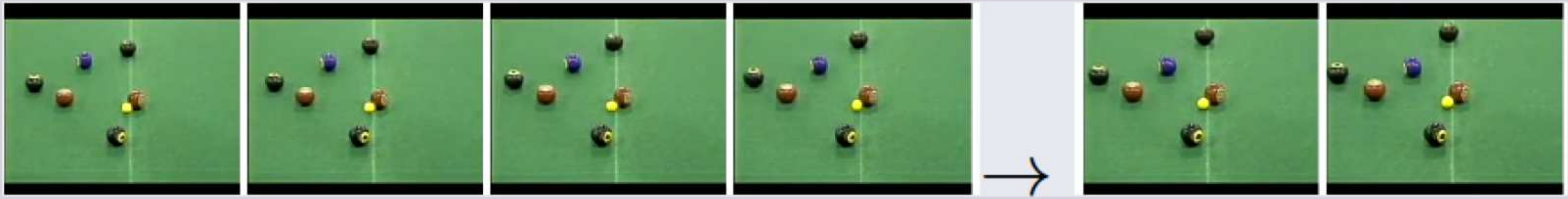
Future video prediction

With Pauline Luc, Michael Mathieu, Natalia Neverova,

Yann LeCun and Jakob Verbeek (INRIA)

Motivation

MOTIVATION



- Building internal representations that model the image evolution accurately, its content and dynamics.
- We postulate that the better the predictions of such system are, the better the feature representation should be.
- Representations learned through prediction of future sequences have been shown to lead to improvements in weakly supervised and even fully supervised tasks (e.g. [Srivastava et al. ICML'15])

Agenda

1 Future image prediction

2 Future semantic segmentation prediction

3 Future instance prediction

4 Joint future instance and semantic segmentation prediction

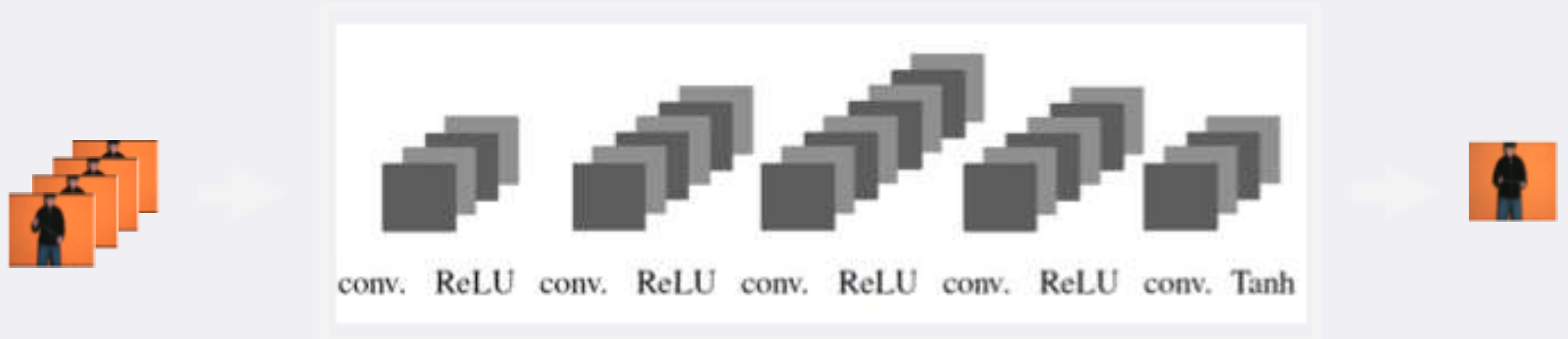
1) Predicting next frames in videos

MICHAEL MATHIEU, CAMILLE COUPRIE, YANN LECUN, ICLR16

4 INPUT IMAGES **OUR 2 PREDICTIONS**

Our contributions

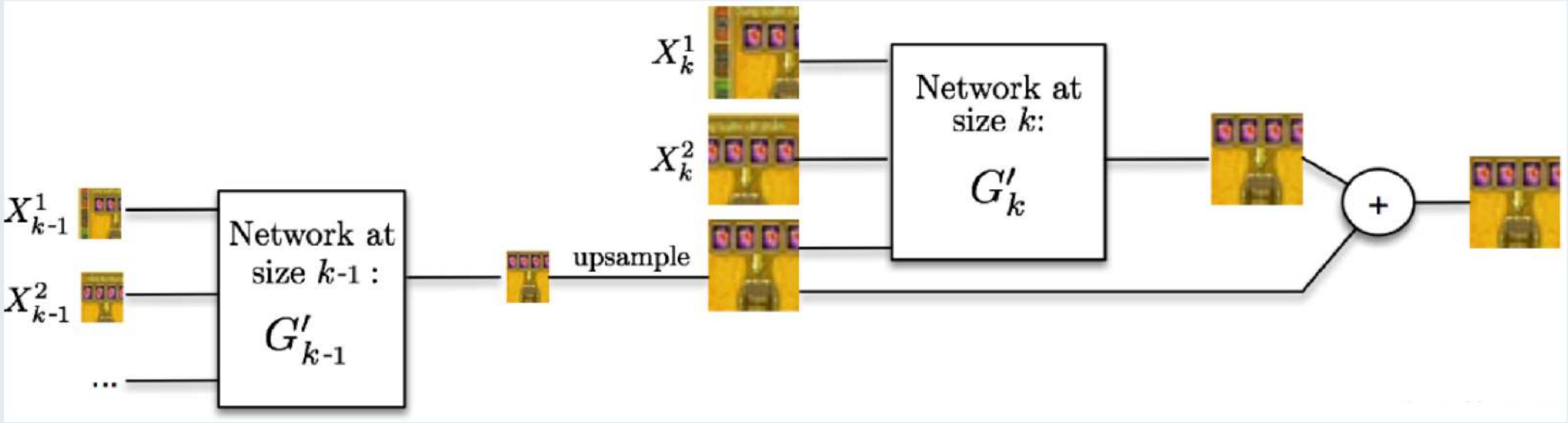
- Result with a simple convolutional network trained minimizing an l2 loss



- **OUR RESULT USING**
  - **A MULTISCALE ARCHITECTURE**
  - **AN IMAGE GRADIENT DIFFERENT LOSS**
  - **USE ADVERSARIAL TRAINING [GOODFELLOW ET AL'14]**

MULTISCALE ARCHITECTURE

$$\hat{Y}_k = G_k(X) = u_k(\hat{Y}_{k/2}) + G'_k\left(X_k, u_k(\hat{Y}_{k/2})\right).$$



ADVERSARIAL TRAINING

- Two models are trained simultaneously : the generative model and a discriminative model that estimates the probability that the predicted frame belongs to a real video sequence.
- **Training D:** Perform a SGD step to minimize

$$\mathcal{L}_D(X, Y) = \sum_{k=1}^{N_{\text{scales}}} \left(\mathcal{L}_{BCE}(D_k(X_k, Y_k), 1) + \mathcal{L}_{BCE}(D_k(X_k, G_k(X)), 0)\right)$$

  where $\mathcal{L}_{BCE}$ is the binary cross-entropy loss.
- **Training G:** Perform a SGD step to minimize

$$\mathcal{L}_G(X, Y) = \sum_{k=1}^{N_{\text{scales}}} \left(\lambda_D \mathcal{L}_{BCE}(D_k(X_k, G_k(X_k)), 1) + \lambda_G L_G(\hat{Y}_k, Y_k)\right)$$

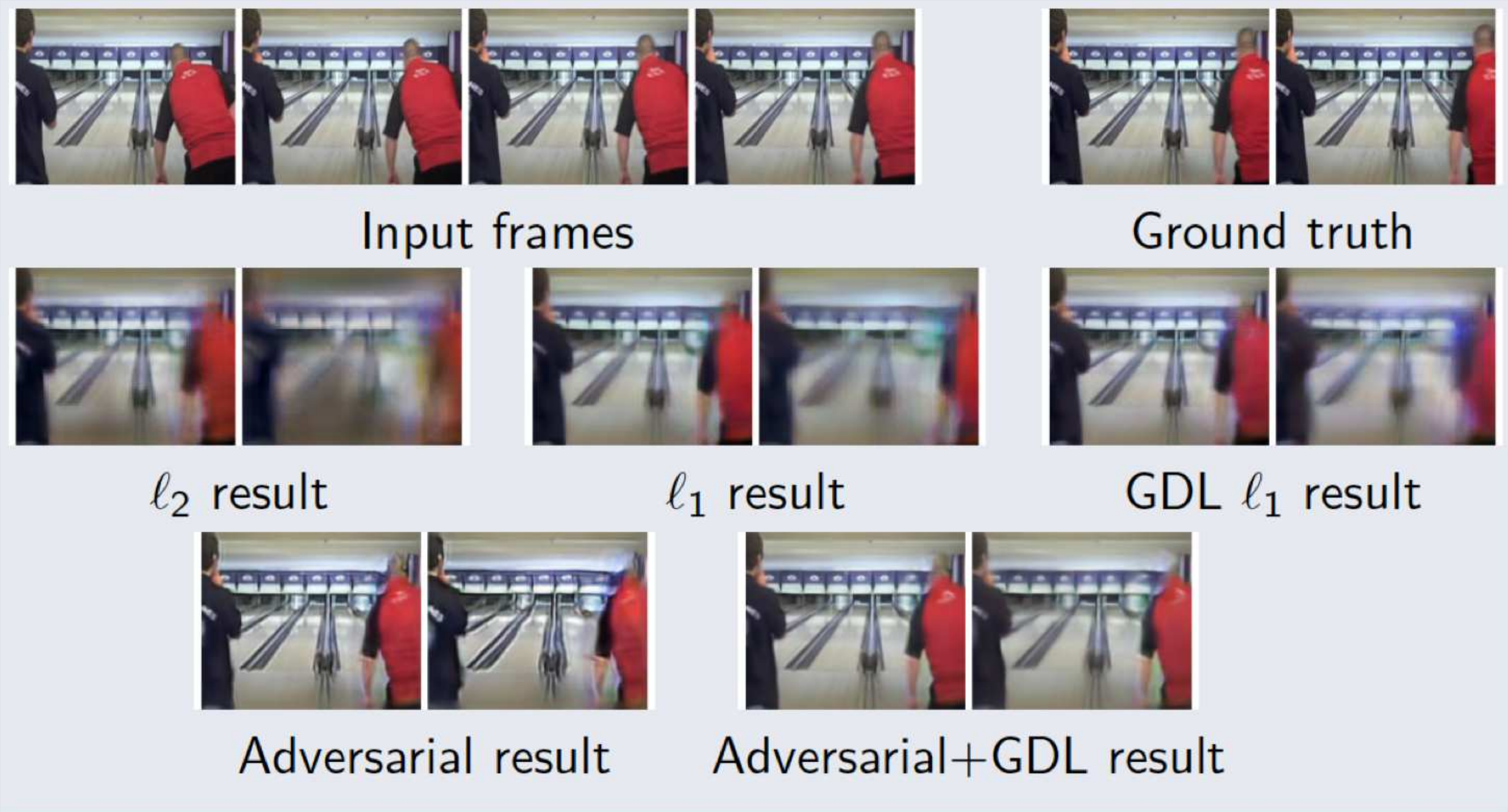
GRADIENT DIFFERENCE LOSS

Another way to avoid blurry predictions is to minimize the local image gradient of the true frame Y and the prediction $\hat{Y}$ at every pixel:

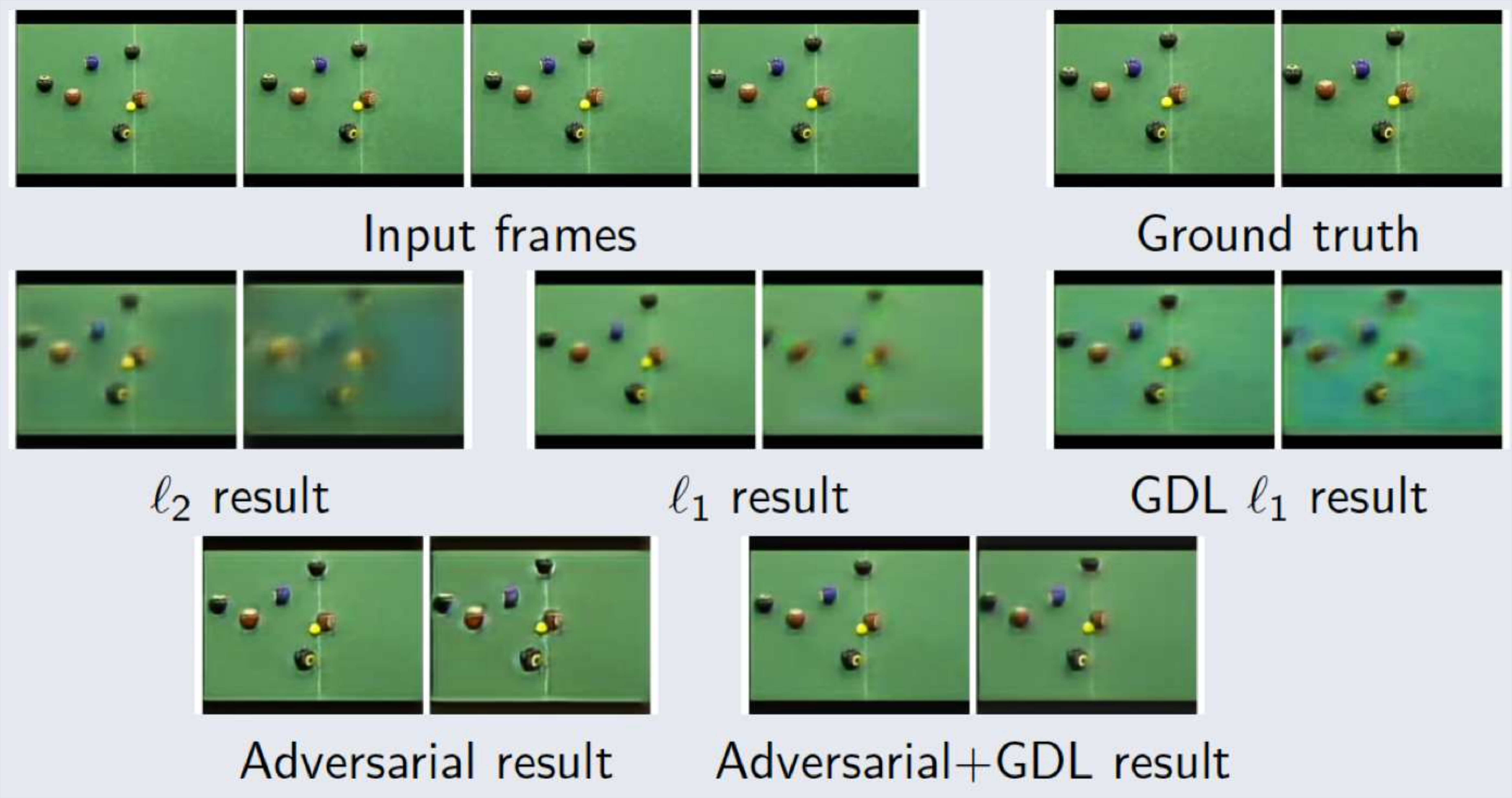
$$GDL(Y, \hat{Y}) = \sum_{i,j} ||Y_{i,j} - Y_{i-1,j}| - |\hat{Y}_{i,j} - \hat{Y}_{i-1,j}||^{\alpha} + ||Y_{i,j-1} - Y_{i,j}| - |\hat{Y}_{i,j-1} - \hat{Y}_{i,j}||^{\alpha},$$

where α is an integer greater or equal to 1.

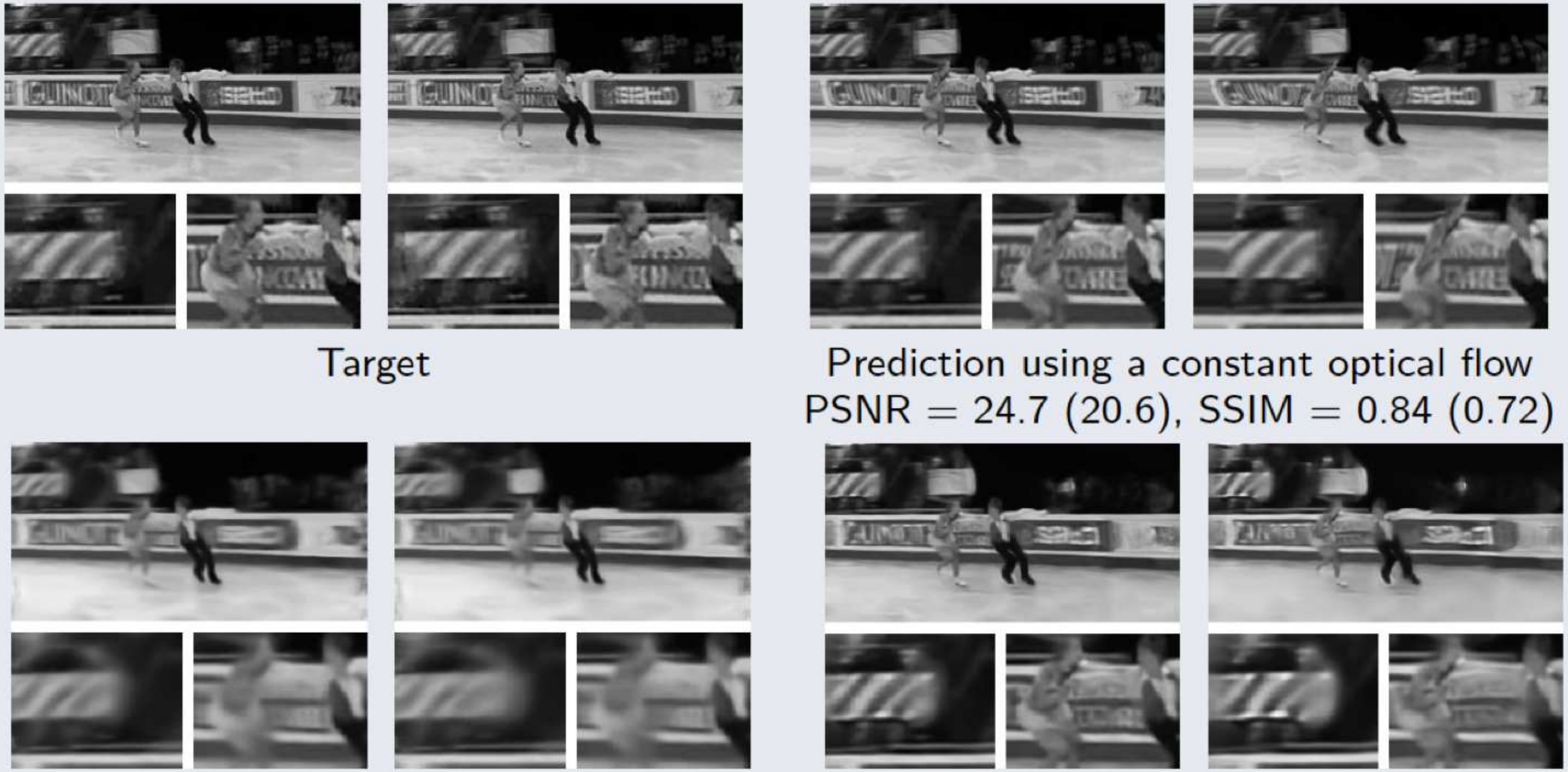
RESULTS ON THE UCF101 DATASET



RESULTS ON THE UCF101 DATASET



COMPARISONS WITH BASELINES



Target

Prediction using a constant optical flow
PSNR = 24.7 (20.6), SSIM = 0.84 (0.72)

Ranzato et al. result
PSNR = 20.1 (17.8), SSIM = 0.72 (0.65)

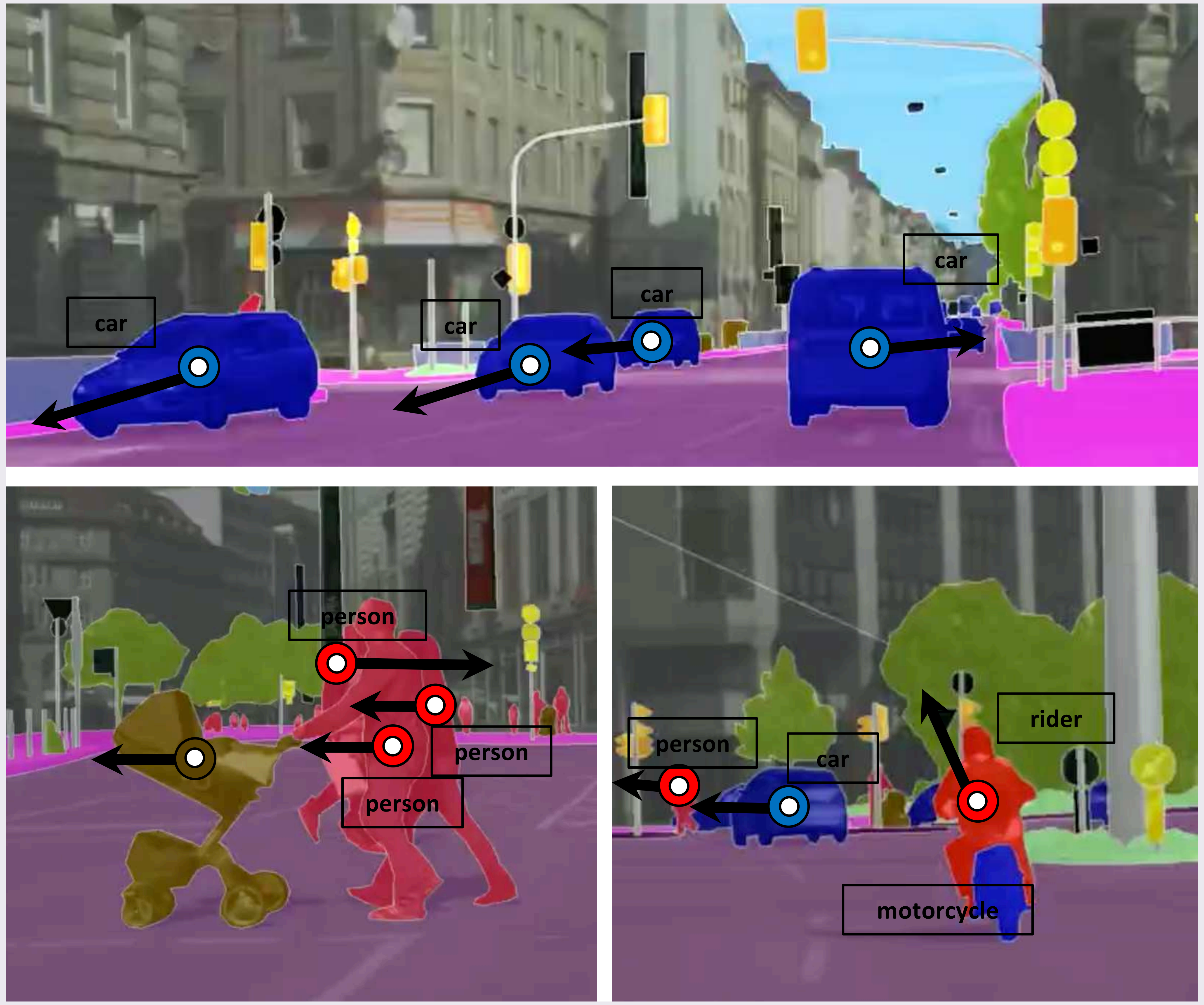
Adv GDL ℓ_1 result
PSNR = 24.6 (20.5), SSIM = 0.81 (0.69)

Towards longer term predictions

- Predictions in the RGB space quickly become blury despites previous attempts
- Idea: predict in the space of semantic segmentation

2) Predicting Deeper into the Future of Semantic Segmentation

P. LUC, N. NEVEROVA, C. COUPRIE, J. VERBEEK, Y. LECUN, ICCV'17

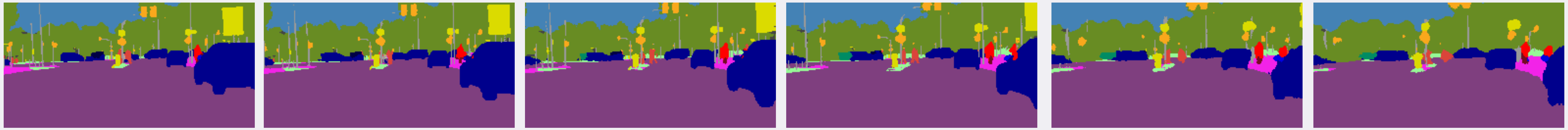


- Use a state-of-the-art semantic segmentation network to obtain densely segmented input / target sequences, e.g. Dilation10, [Yu et al.'16].
  Specifically, use the softmax pre-activations, i.e. the (continuous) outputs of the last convolutional layer, before the softmax

Setting and dataset presentation

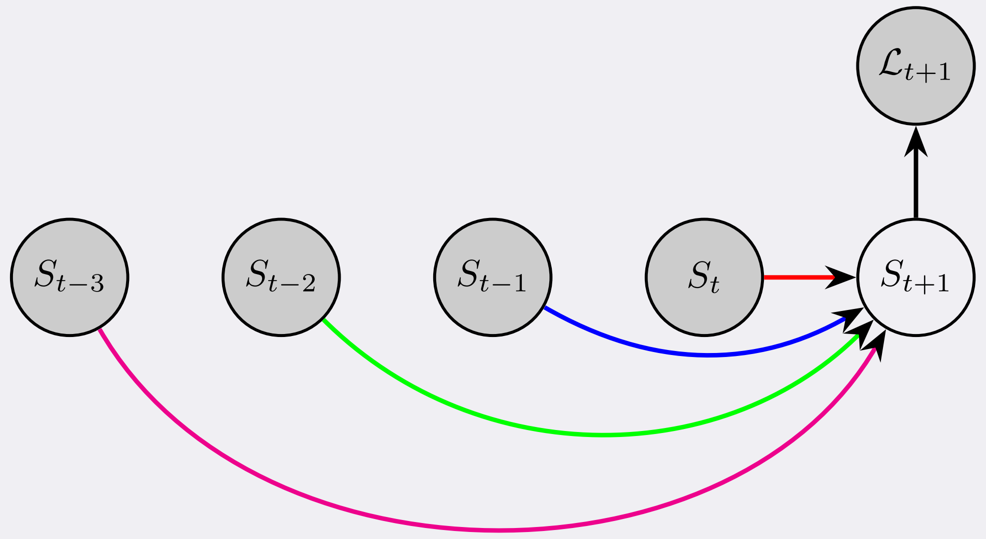
4 INPUT IMAGES

OUR 2 PREDICTIONS

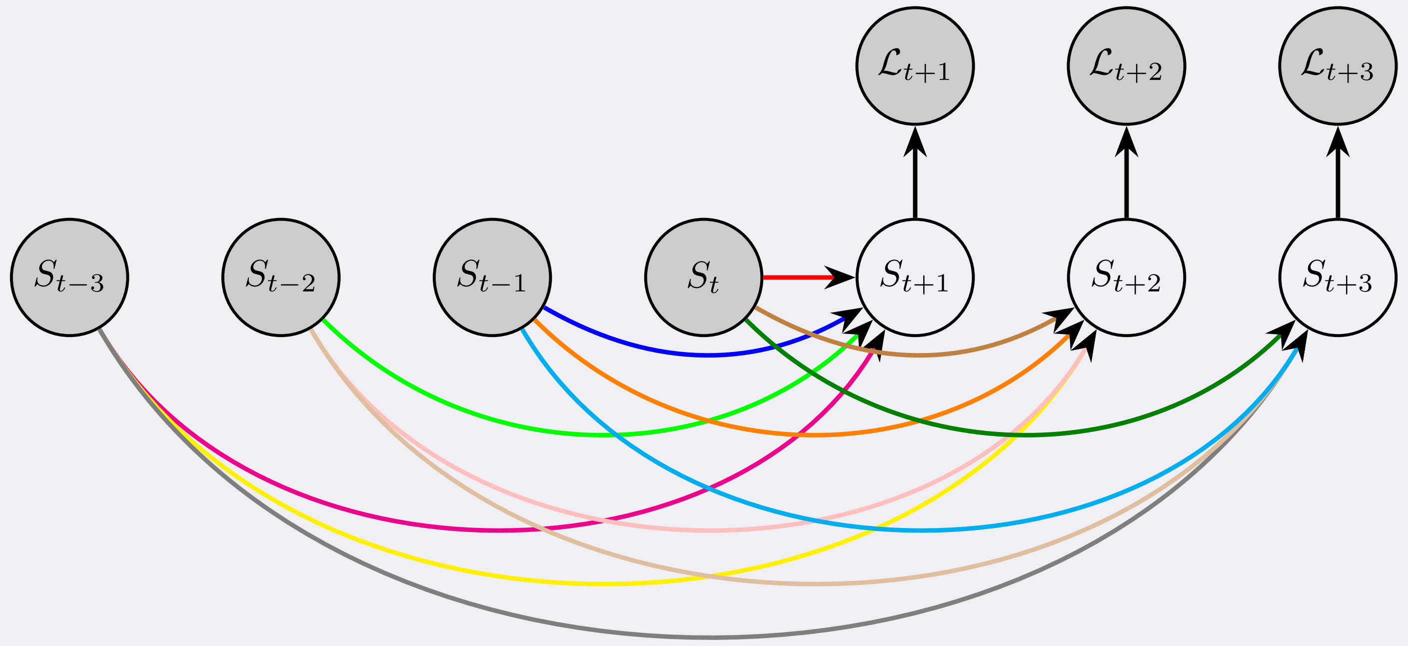


CITYSCAPES DATASET
[CORDTS ET AL.'16]

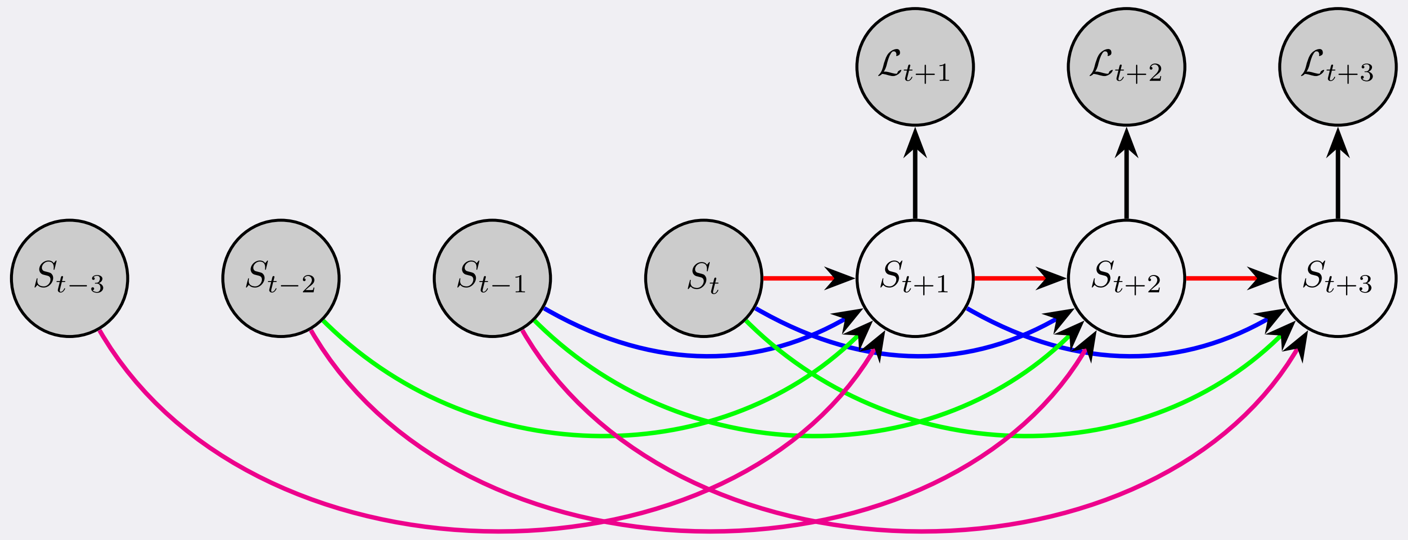
Approach – predicting deeper into the future



Single time-step



BATCH MODEL



AUTOREGRESSIVE MODEL

SAME COLOR = SHARED WEIGHTS

AUTOREGRESSIVE MODE IS ONLY POSSIBLE FOR X2X, S2S, XS2XS

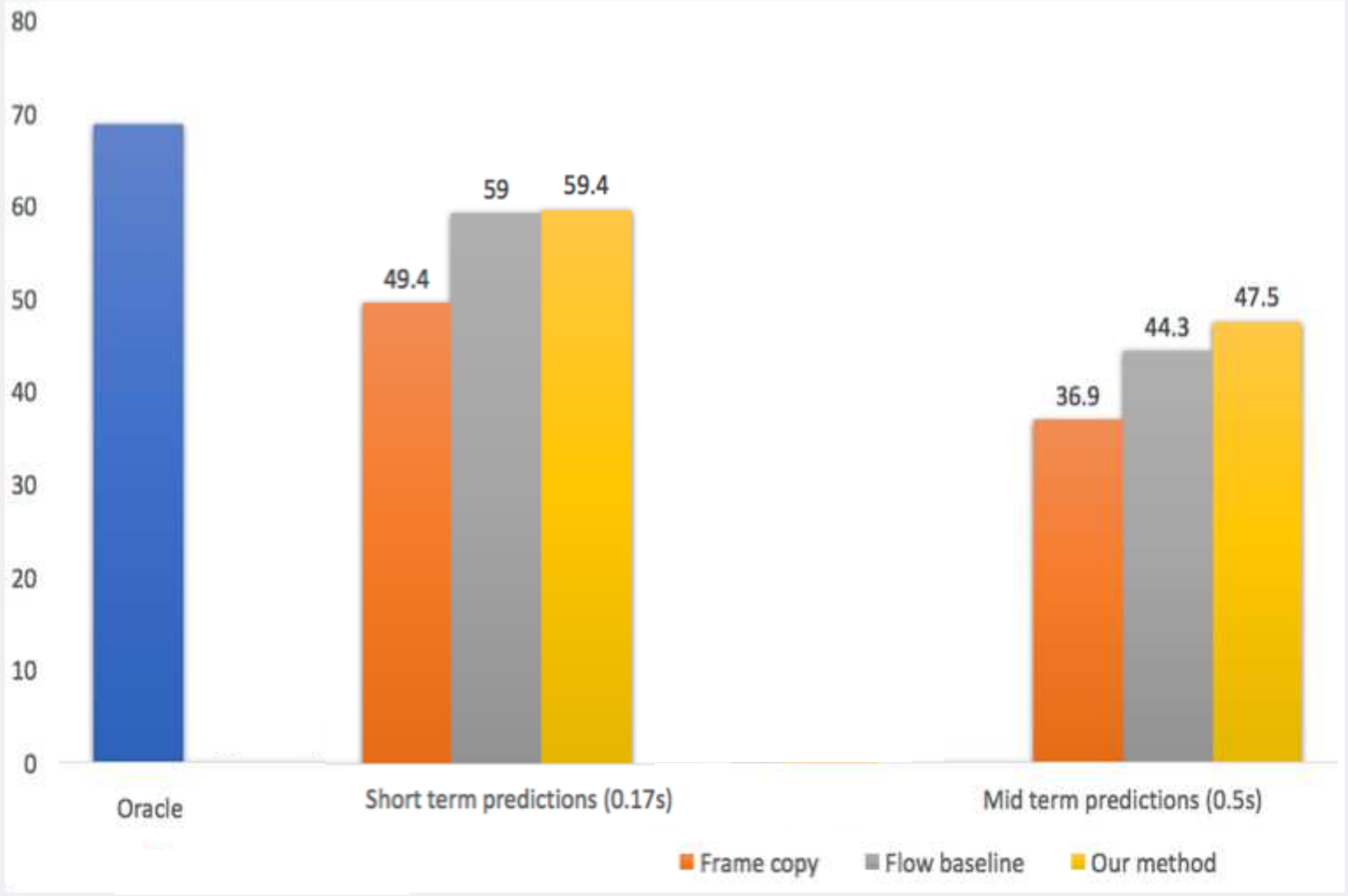
AUTOREGRESSIVE MODEL IS EITHER :

- USED FOR INFERENCE WITHOUT ADDITIONAL TRAINING (W.R.T. TO SINGLE TIME STEP MODEL) AR
- FINE-TUNED USING BPTT AR FINE-TUNE

Quantitative results

BASELINES :
- COPY THE LAST INPUT FRAME TO THE OUTPUT
- ESTIMATE FLOW BETWEEN THE TWO LAST INPUTS, AND PROJECT THE LAST INPUT FORWARD USING THE FLOW

PERFORMANCE MEASURE (MEAN IOU) OF OUR APPROACH AND BASELINES

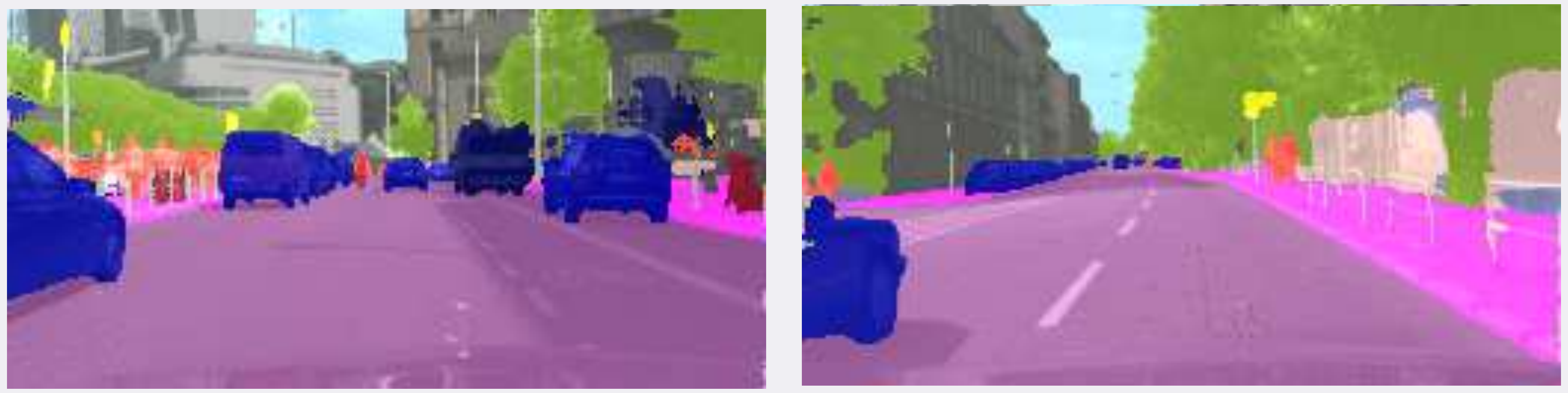


Showed experimentally that it is a better setting:

RGB	:	Autoregressive	<	Batch
Semantic segmentation	:	Autoregressive	>	Batch

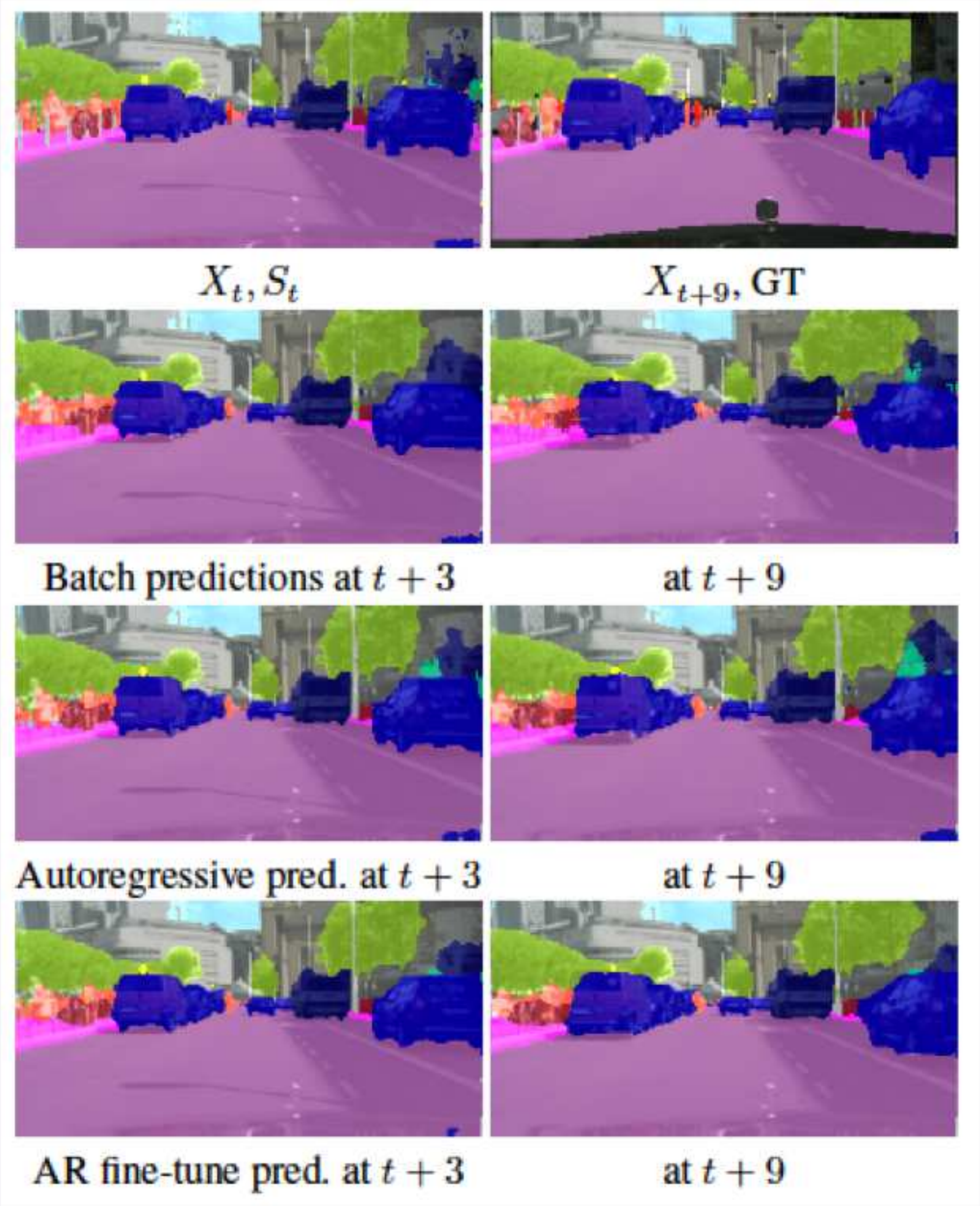
Mid term segmentation predictions (0.5 s)

FLOW BASELINE



OUR AUTOREGRESSIVE FINE-TUNE RESULT

LAST INPUT **GROUND TRUTH**



Mid term segmentation predictions (0.5 s)

FLOW BASELINE

OUR AUTOREGRESSIVE FINE-TUNE RESULT

Long term prediction (10 s) & going further

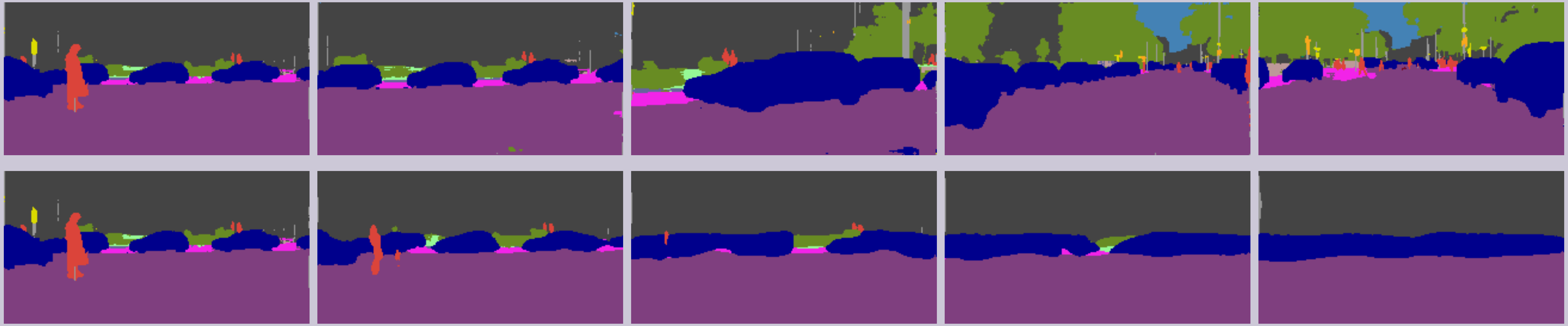
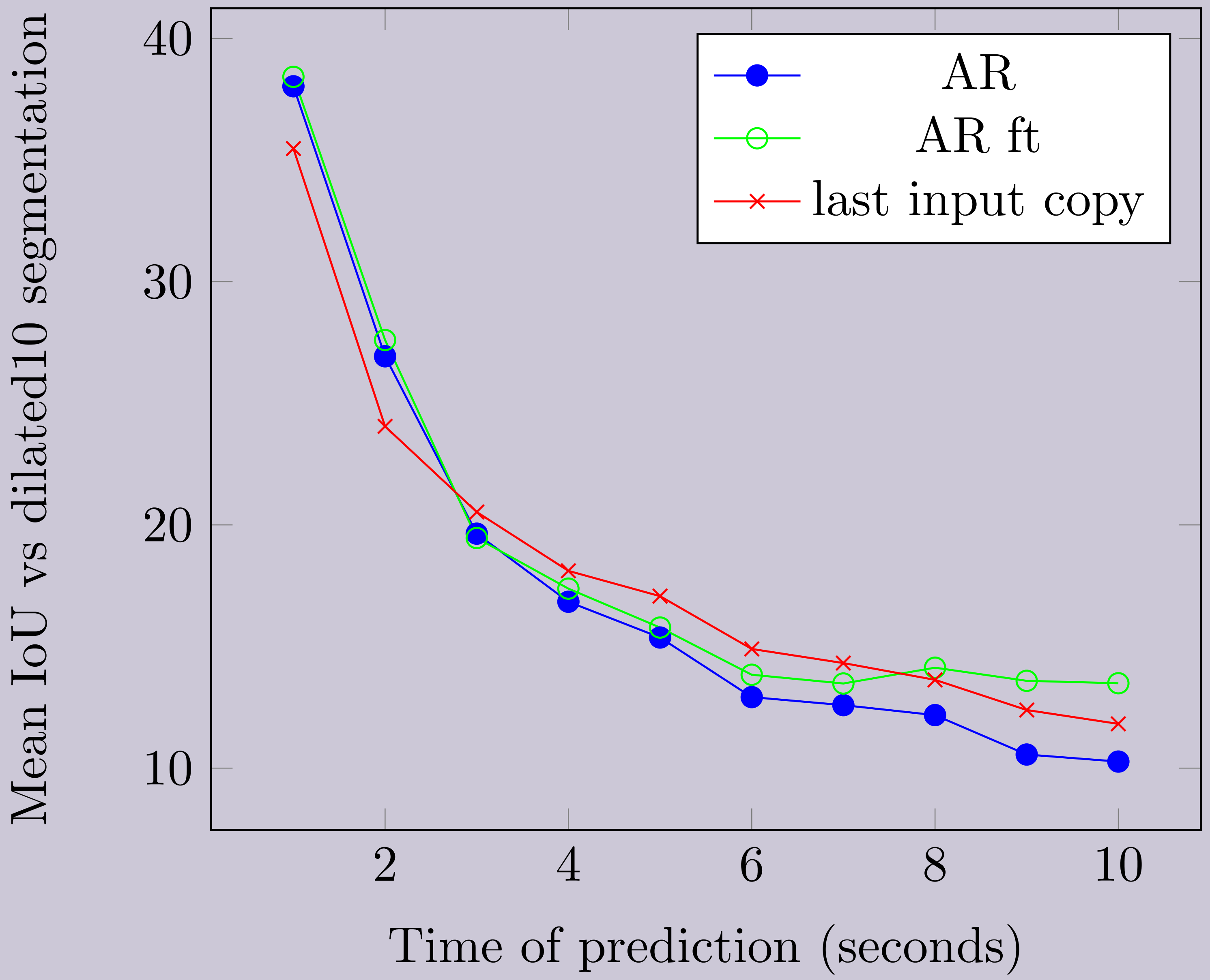


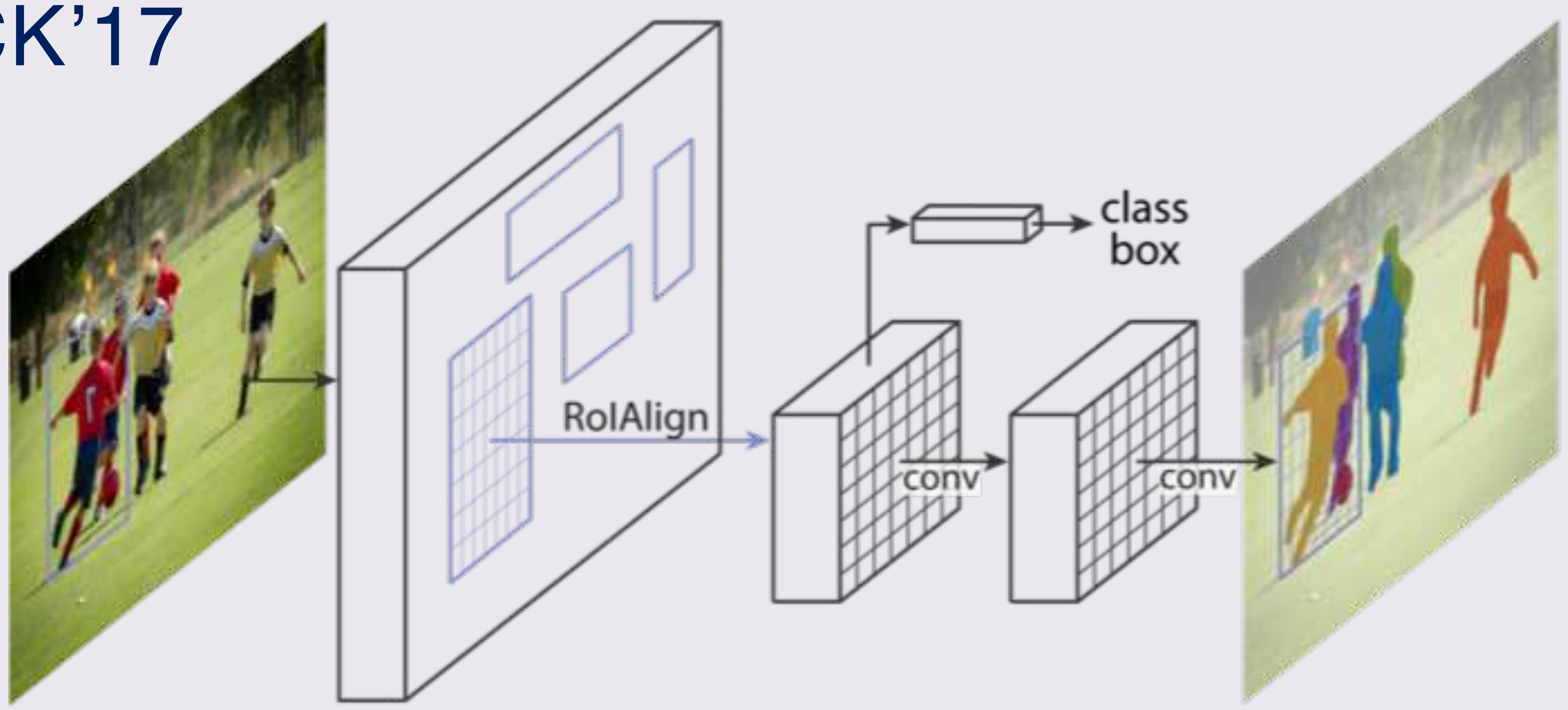
Figure 6: Last input segmentation, and ground truth segmentations at 1, 4, 7, and 10 seconds into the future (top row), and corresponding predictions of the autoregressive S2S model trained with fine-tuning (bottom row).

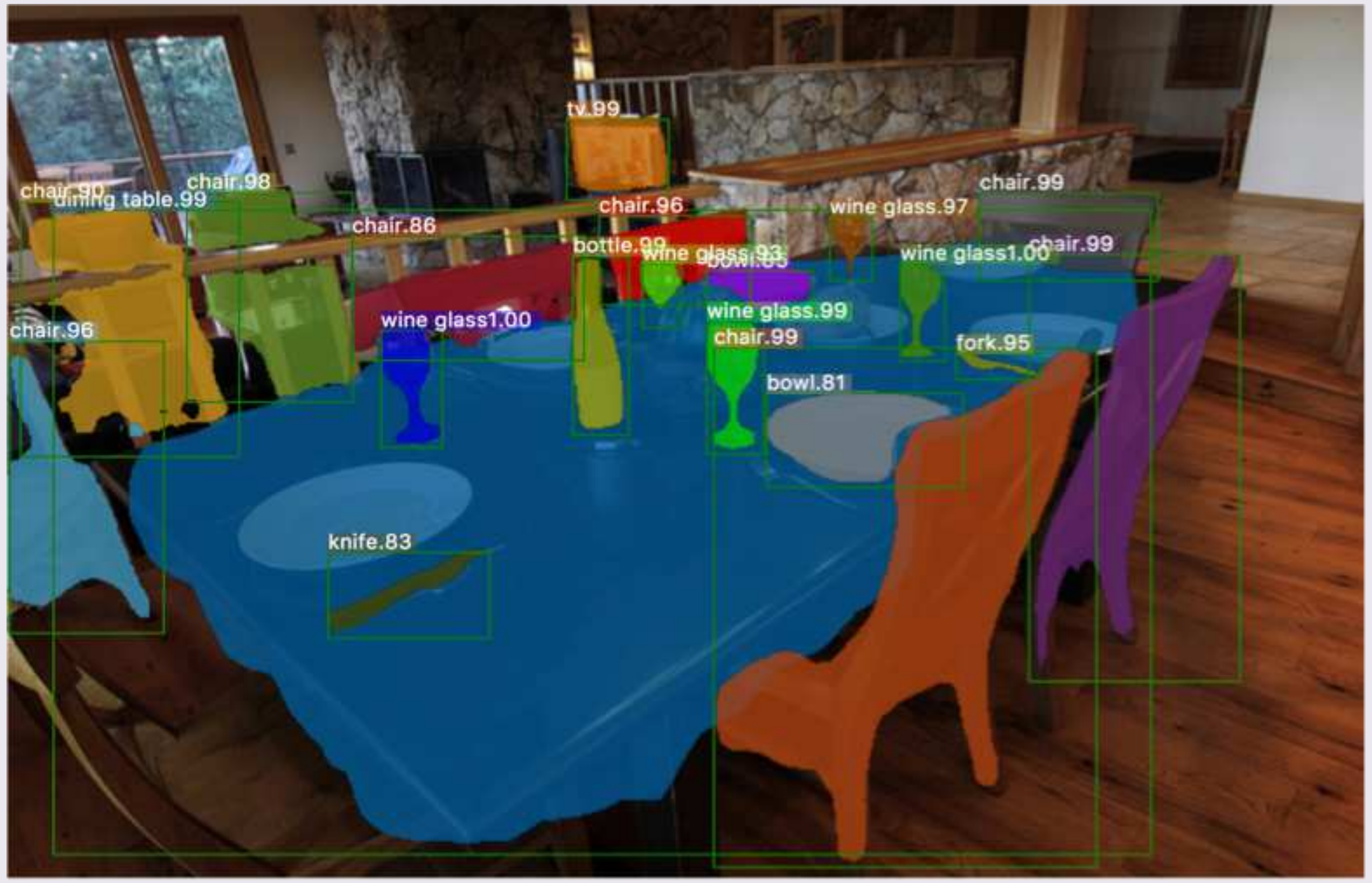


Instance level segmentation: Mask RCNN

K. HE G. GKIOXARI P. DOLLAR R. GIRSHICK'17

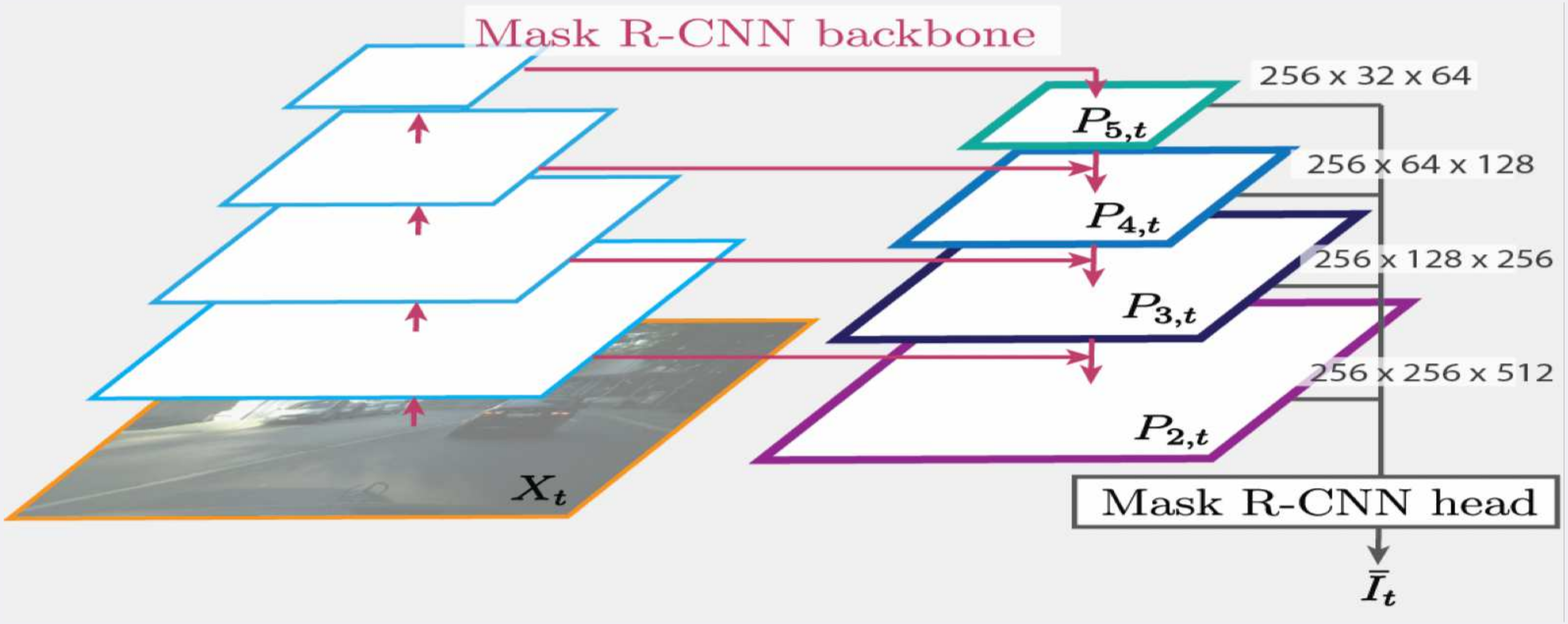
- Extends Faster RCNN [Ren et al.'15] by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition





Instance level segmentation: Mask RCNN

K. HE G. GKIOXARI P. DOLLAR R. GIRSHICK'17

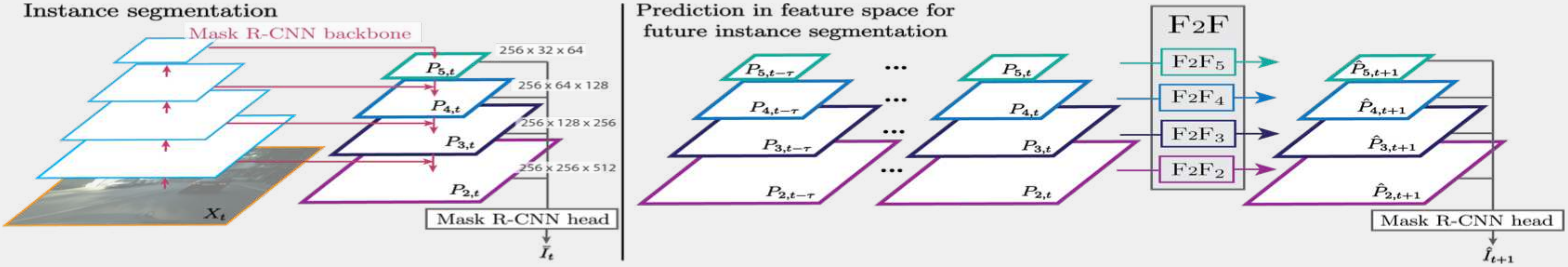


3) Predicting Future Instance Segmentations by Forecasting Convolutional Features

P. LUC, C. COUPRIE, Y. LECUN, J. VERBEEK, ARXIV 2018

LUC, NEVEROVA ET AL. ICCV17

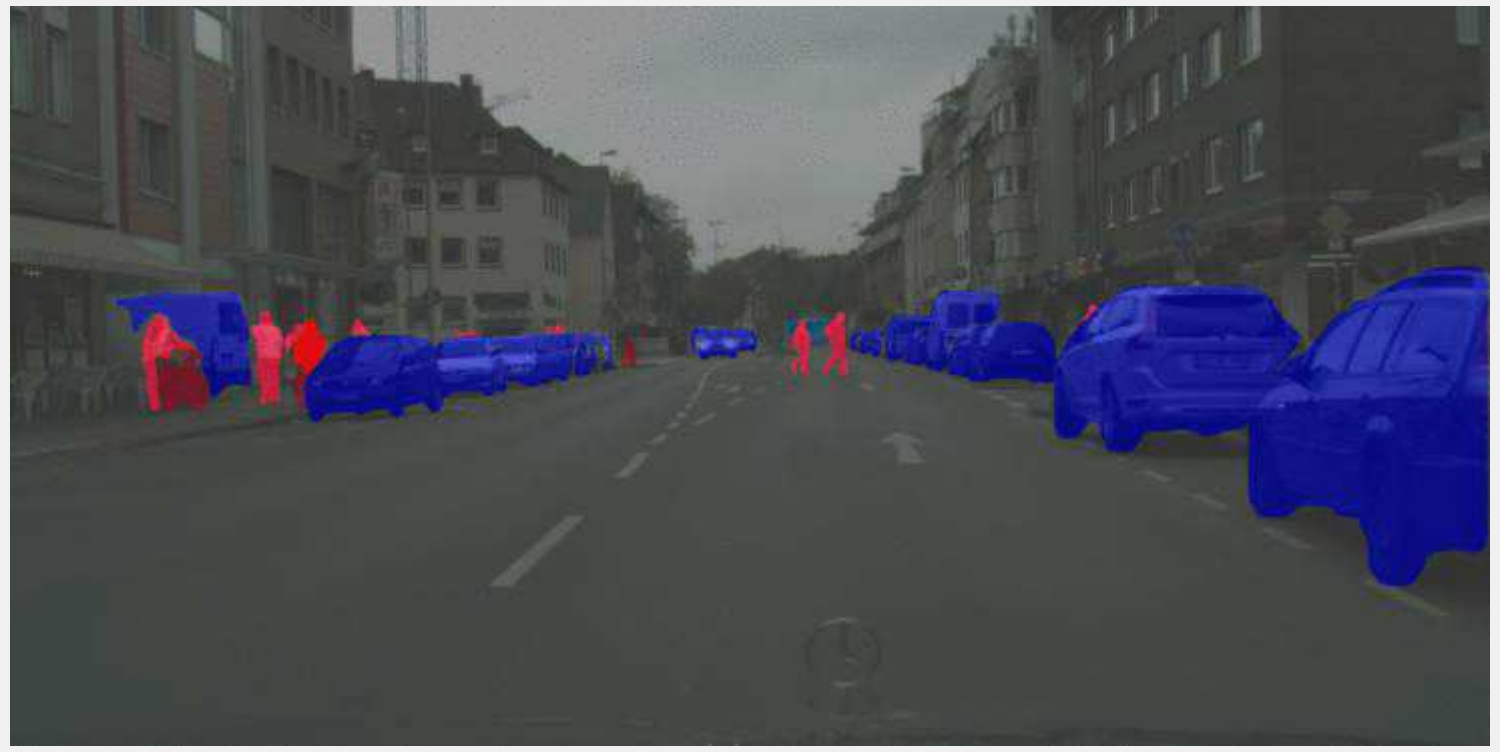
NEW ECCV SUBMISSION: F2F PREDICTIONS

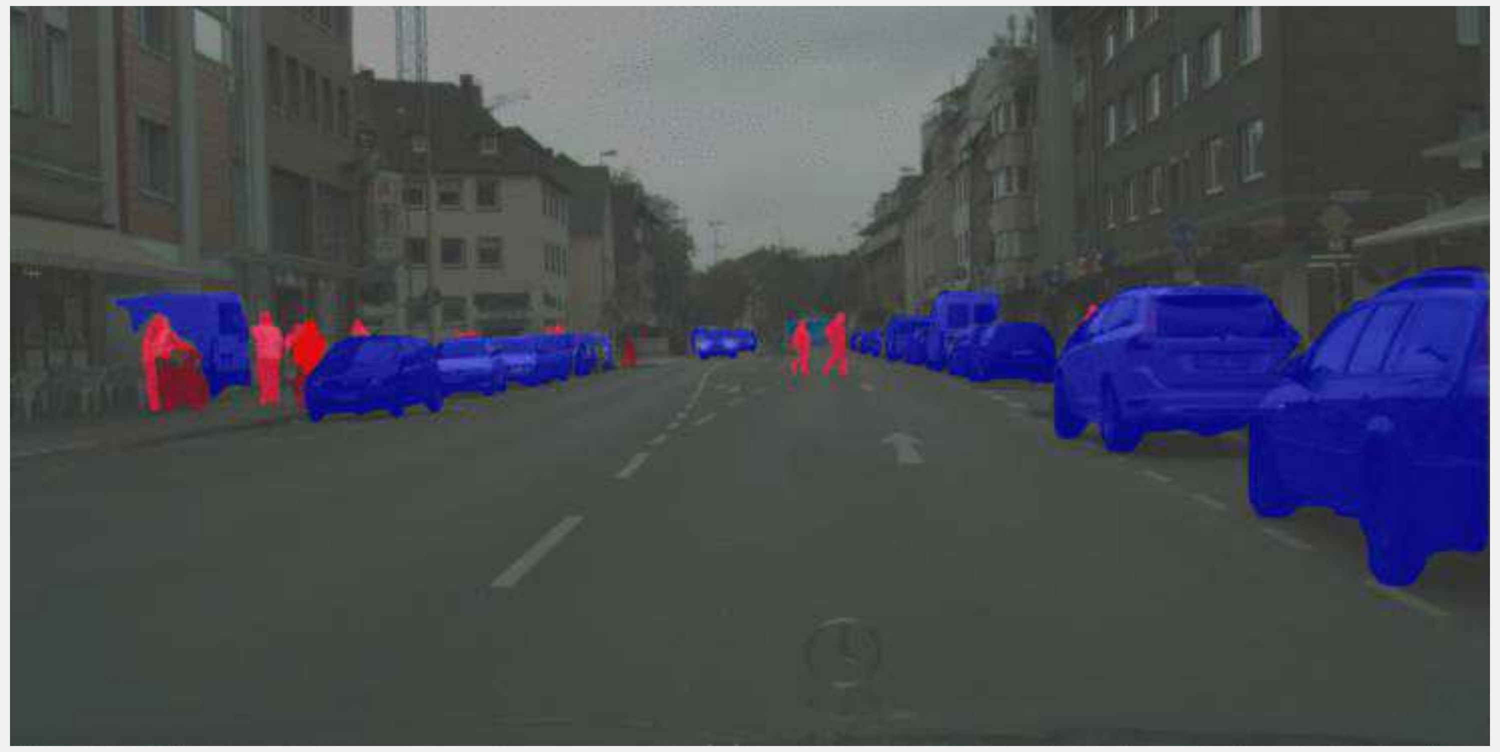


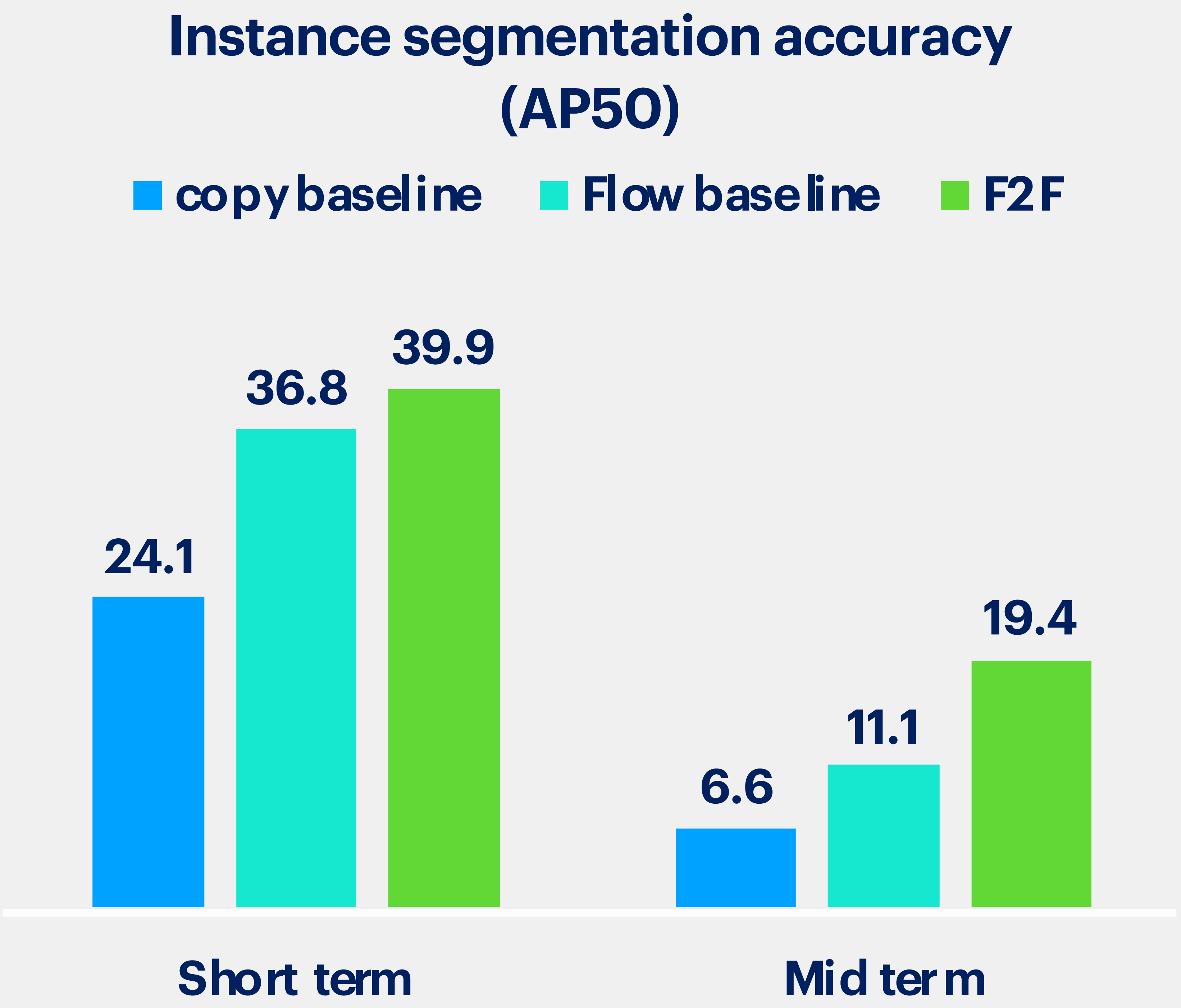
Results

OPTICAL FLOW BASELINE

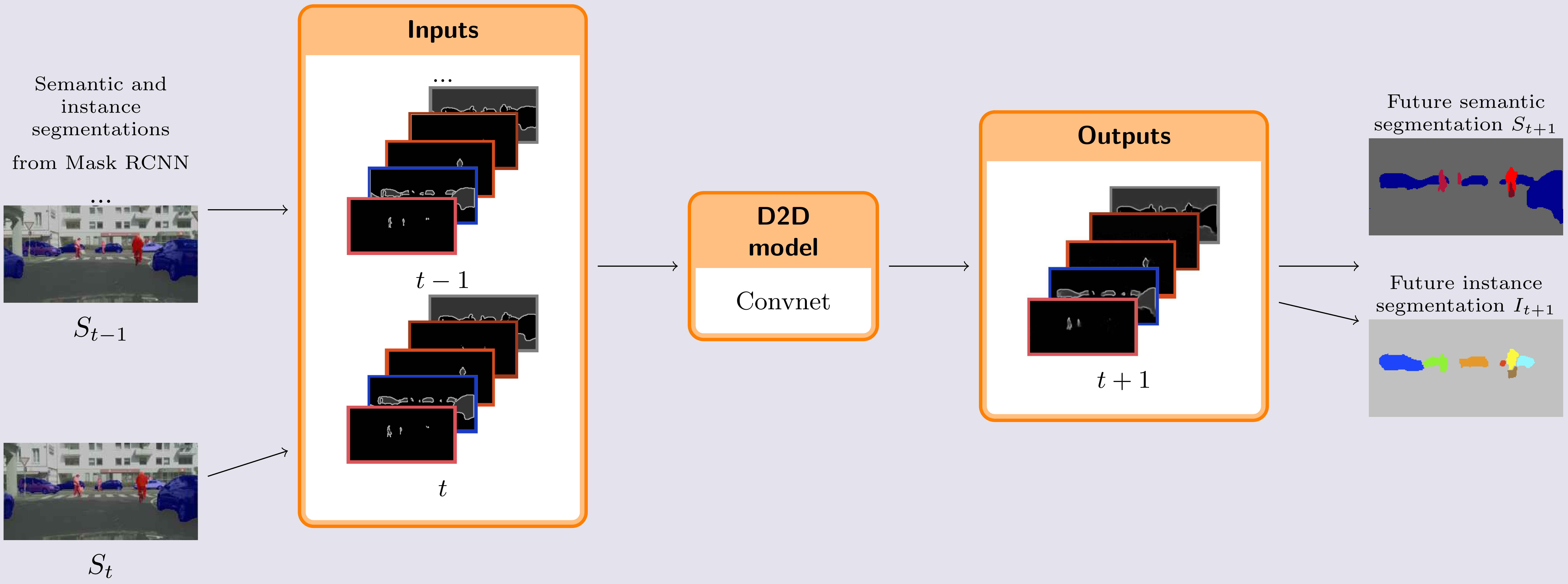
OUR F2F RESULTS





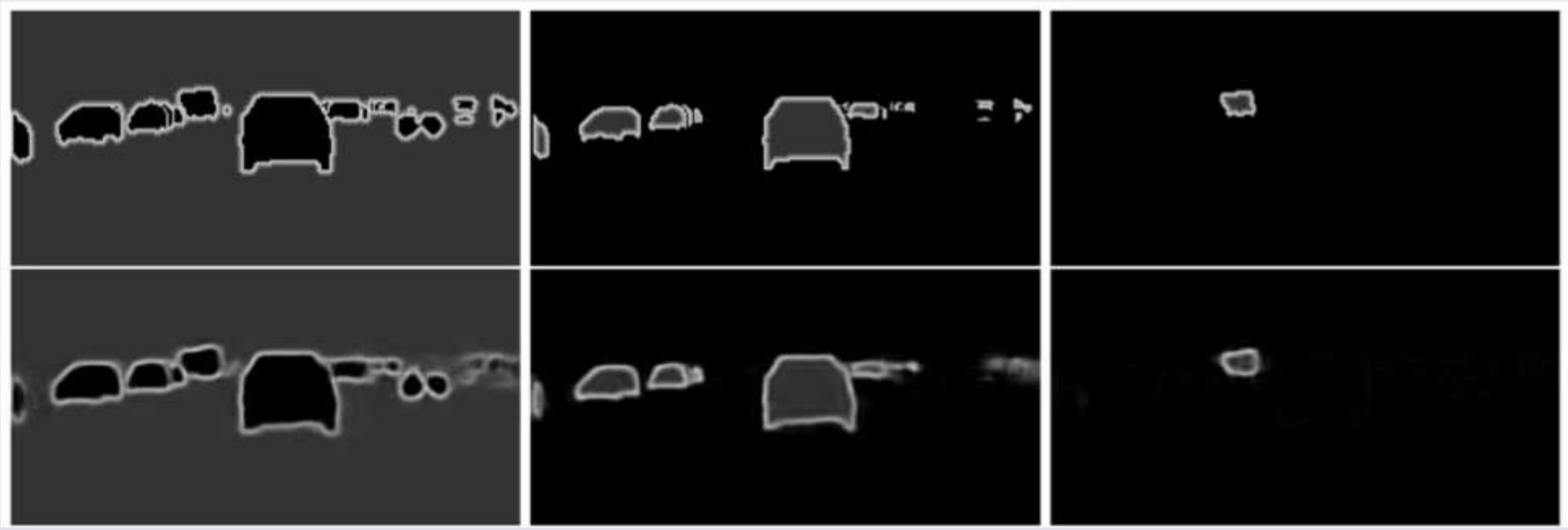


4) Joint semantic and instance segmentation



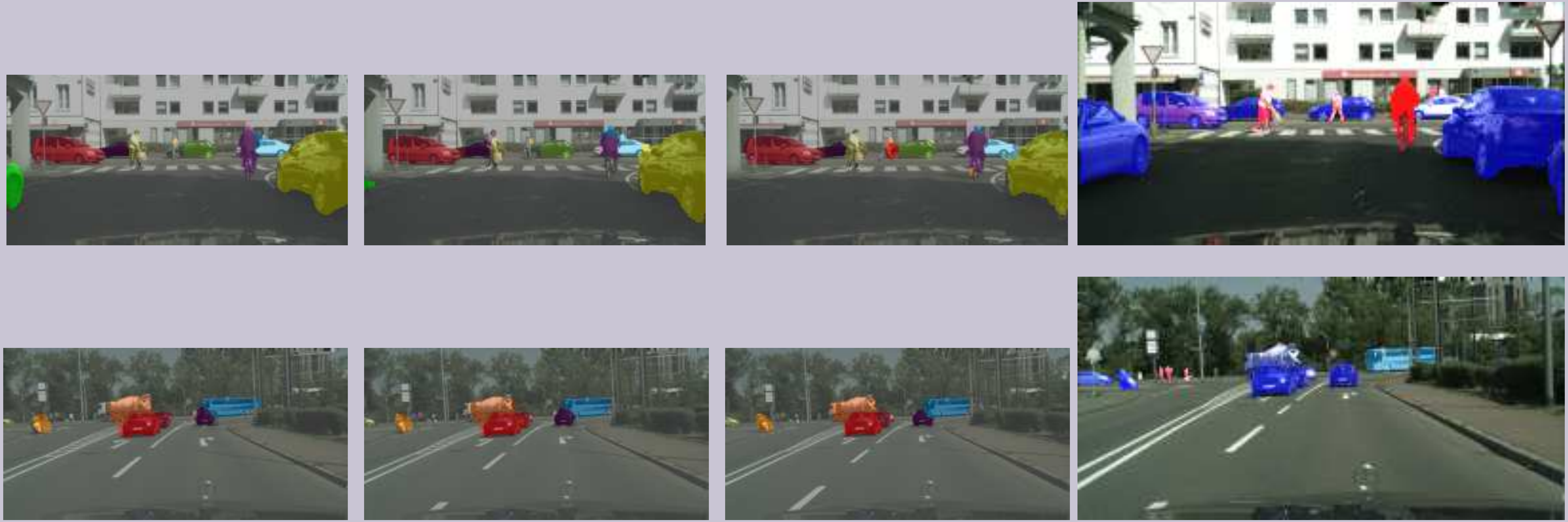
Overview of our approach

- 1) Computation of distance map based representations r(t), r(t-1), ...
- 2) Training a convnet to predict future representations r(t+1)
- 3) Object centroids extraction and linear extrapolation
- 4) Computing instance segmentations using centroids as seeds, and map of maxima of r(t+1) as weights



- 5) Computing the semantic segmentation map as the argmax of r(t+1)

Results



4) Joint semantic and instance segmentation

	Mask R-CNN Feature	Optical Flow	Distance based representation
Mid term sem. segm (IoU)	41.2	41.4	**43.0**
Mid term inst. segm (NO-AP50)	**16.1**	9.5	10.2
Tracking included	no	**yes**	**yes**
Training time	6 days	-	**1 day**
Network size	65M	-	0.8 M
Training hyperparam. to tune	8	-	2
Inference time	**some sec.**	2 min	**some sec.**
Post processing	threshold	hole filling, thres.	optimization

Conclusions

Introduced generic approaches for video prediction

Many problems remain, e.g. handling occlusions

Non deterministic models

Thank You.