# New Perspectives for Processing and Synthesizing Images and Videos

Qifeng Chen
Assistant Professor, HKUST

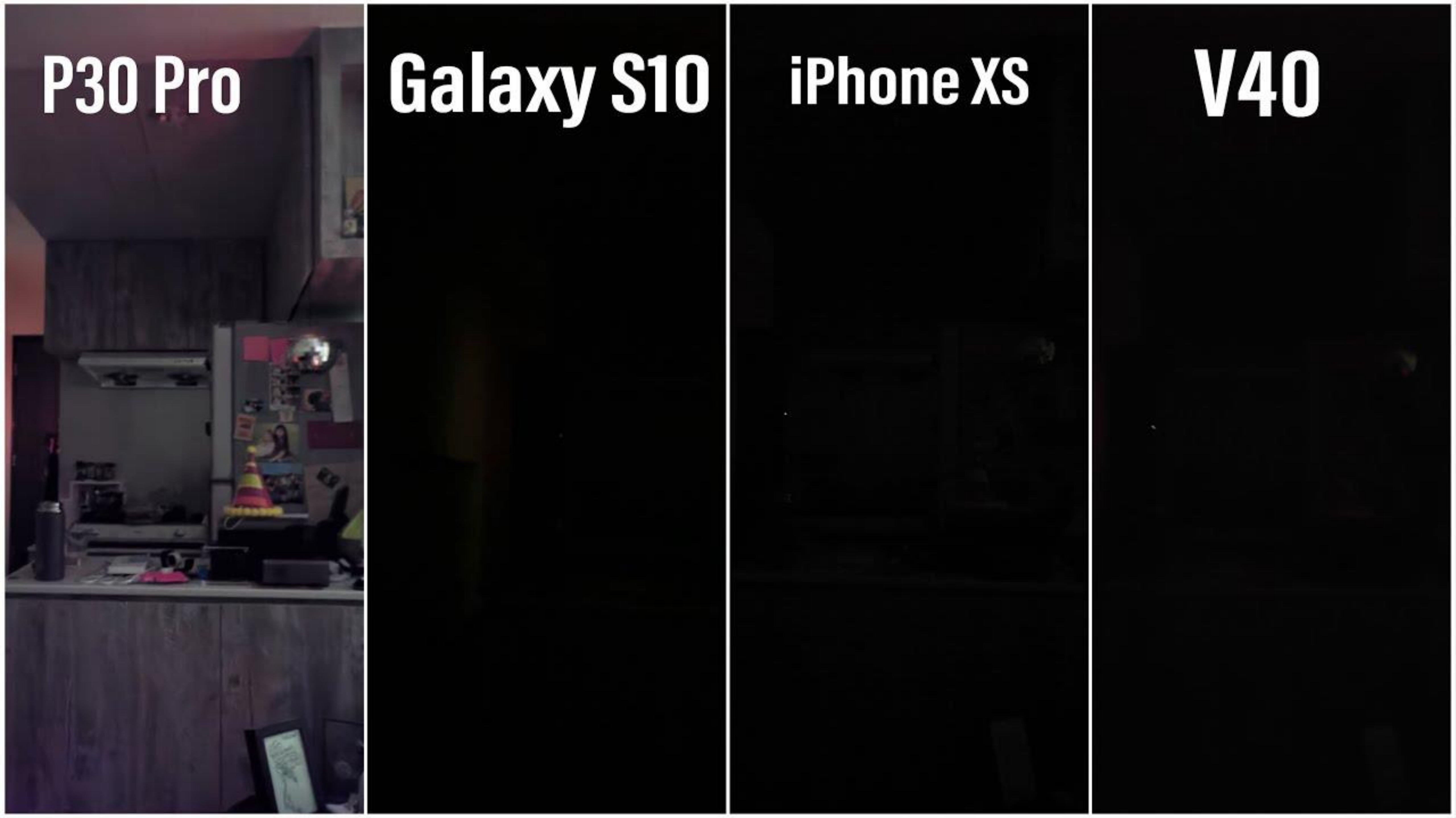
# Q&A

- Which company is the most valuable worldwide?
- Apple

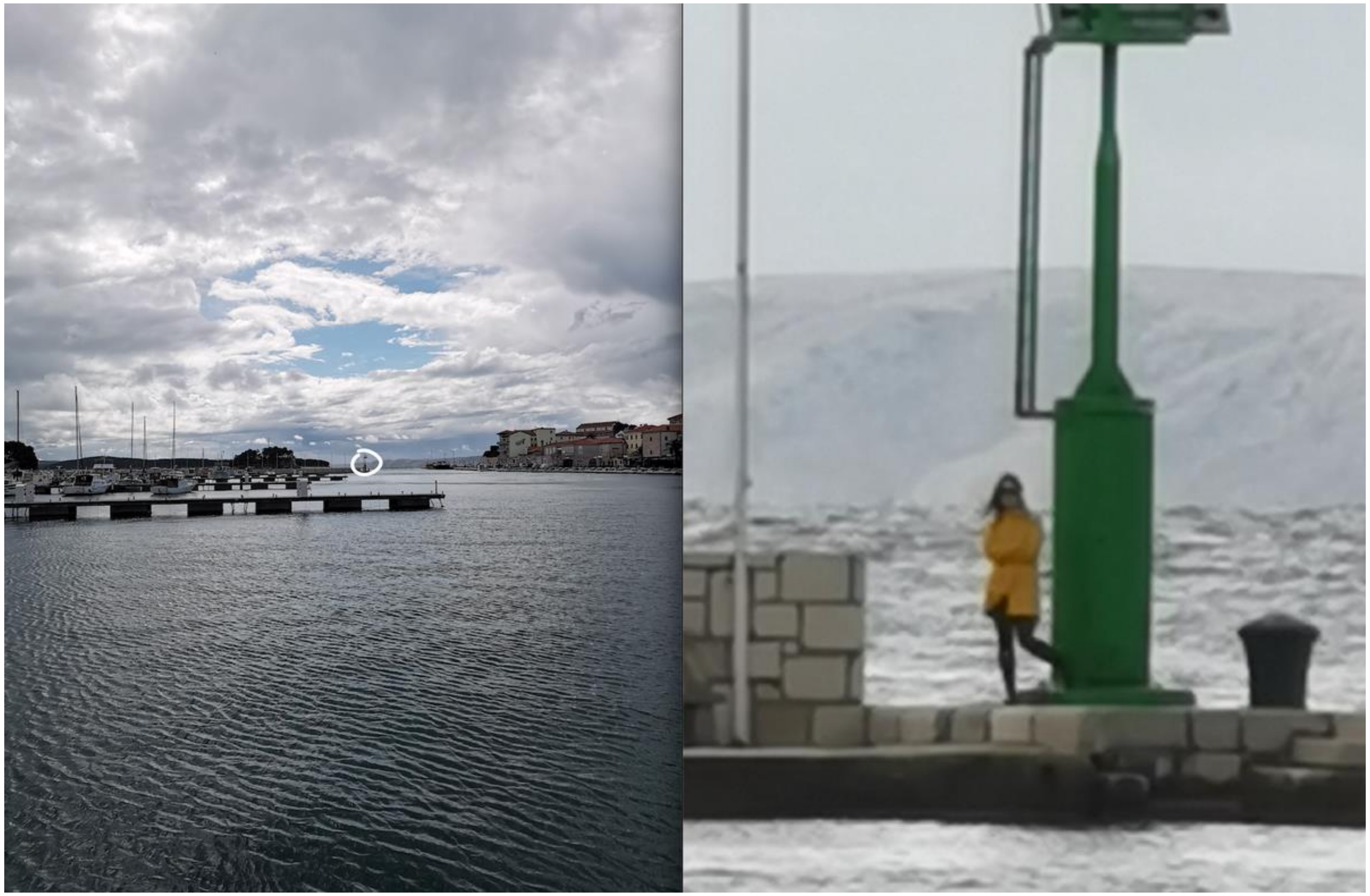
- What is the most important product of Apple?
- iPhone

- What is the most differentiable functionality of a smart phone today?
- Photography (arguably)

# Low-light Imaging



# Powerful Zoom



# Overview

- Image and Video Processing
  - Learning to See in the Dark
  - Zoom to Learn, Learn to Zoom
  - Fast Image and Video Processing
  - Reflection Removal
- Image and Video Synthesis
  - Photographic Image Synthesis
  - Semi-parametric Image Synthesis
  - RGBD Future Video Prediction
  - Fully Automatic Video Colorization

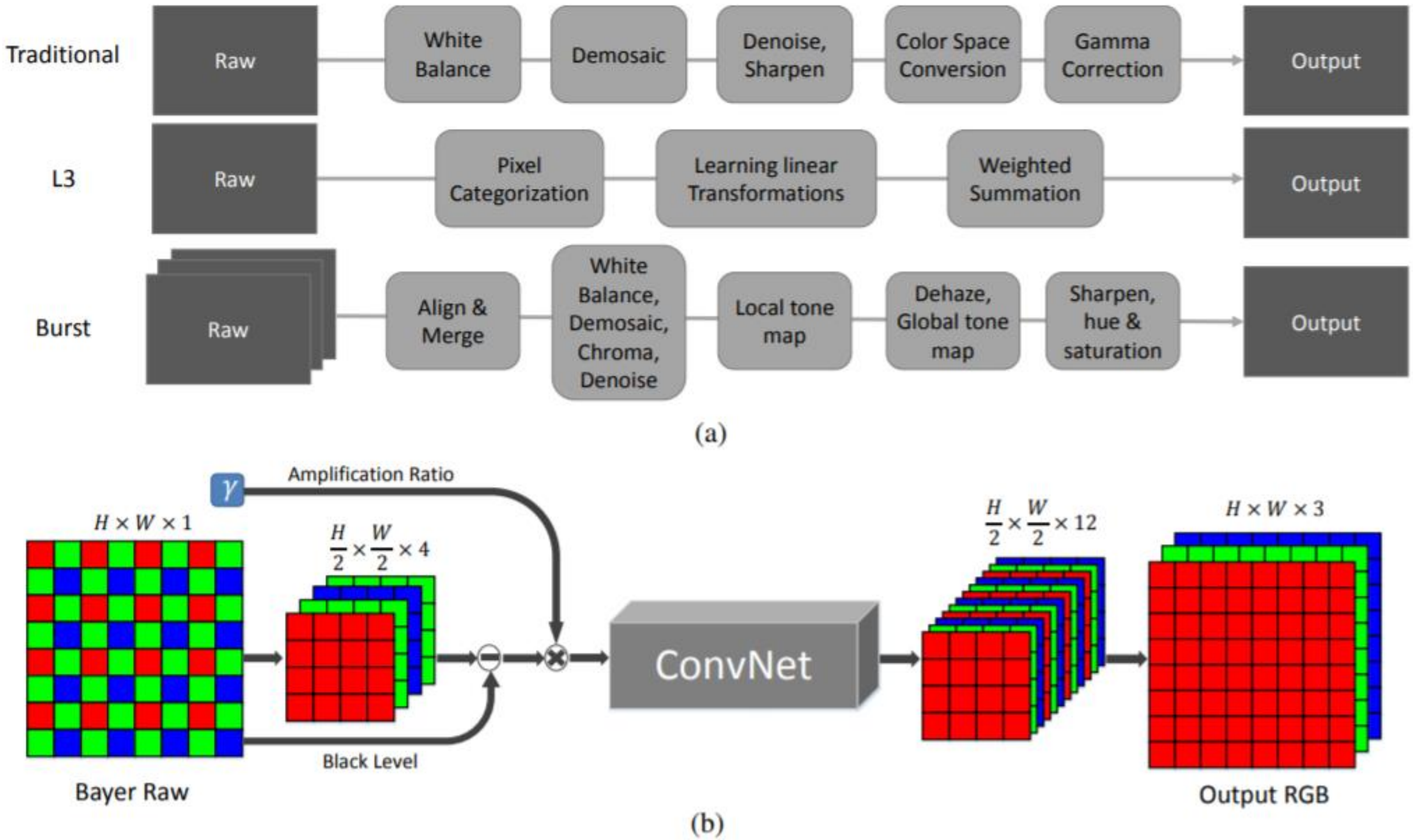
# Image and Video Processing

# Learning to See in the Dark


(a) Camera output with ISO 8,000 (b) Camera output with ISO 409,600 (c) Our result from the raw data of (a)

Figure 1. Extreme low-light imaging with a convolutional network. Dark indoor environment. The illuminance at the camera is $< 0.1$ lux. The Sony $\alpha$7S II sensor is exposed for 1/30 second. (a) Image produced by the camera with ISO 8,000. (b) Image produced by the camera with ISO 409,600. The image suffers from noise and color bias. (c) Image produced by our convolutional network applied to the raw sensor data from (a).

# Learning to See in the Dark




A deep learning based Image Signal Processor

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to See in the Dark, CVPR 2018

# Dataset

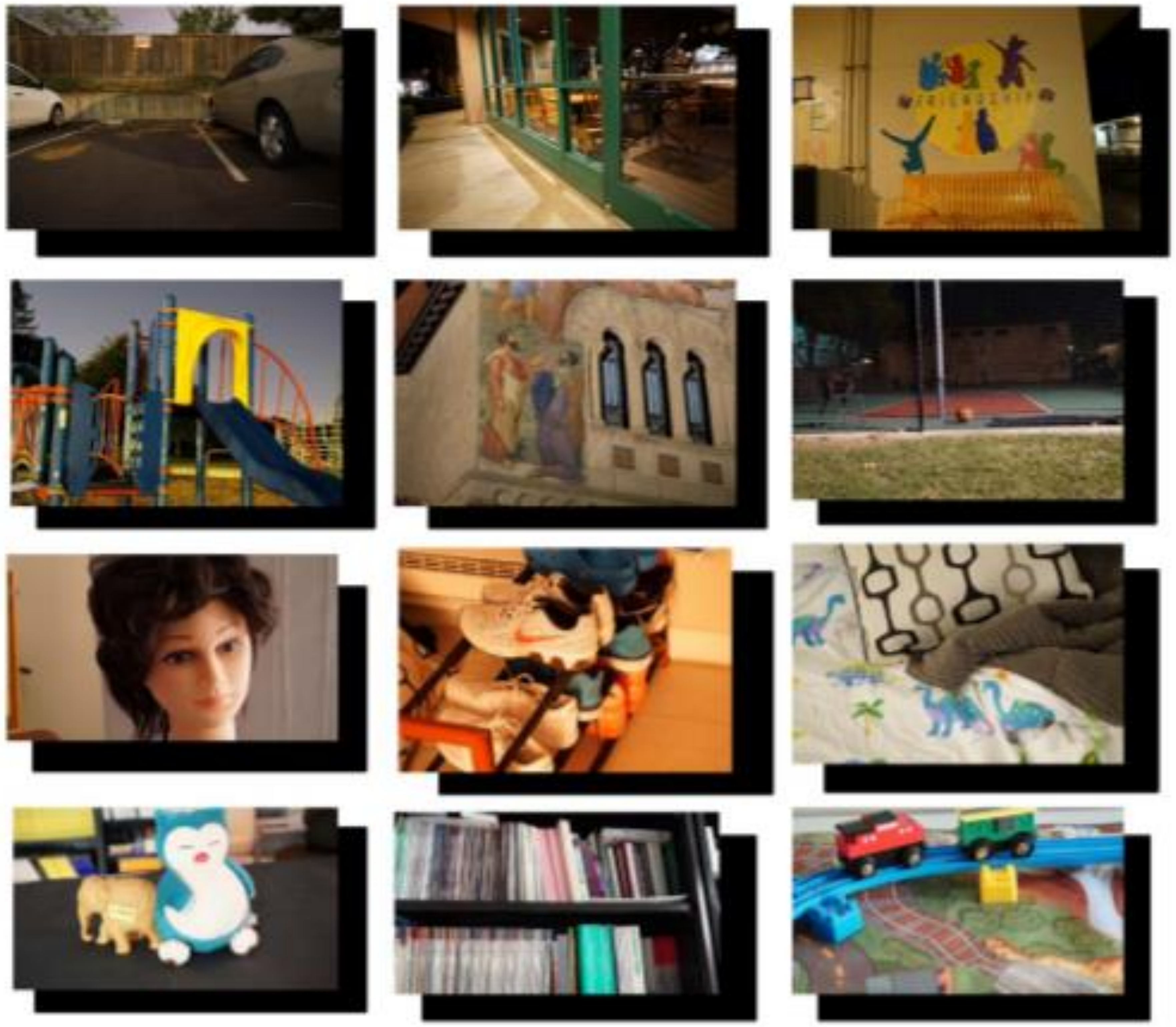


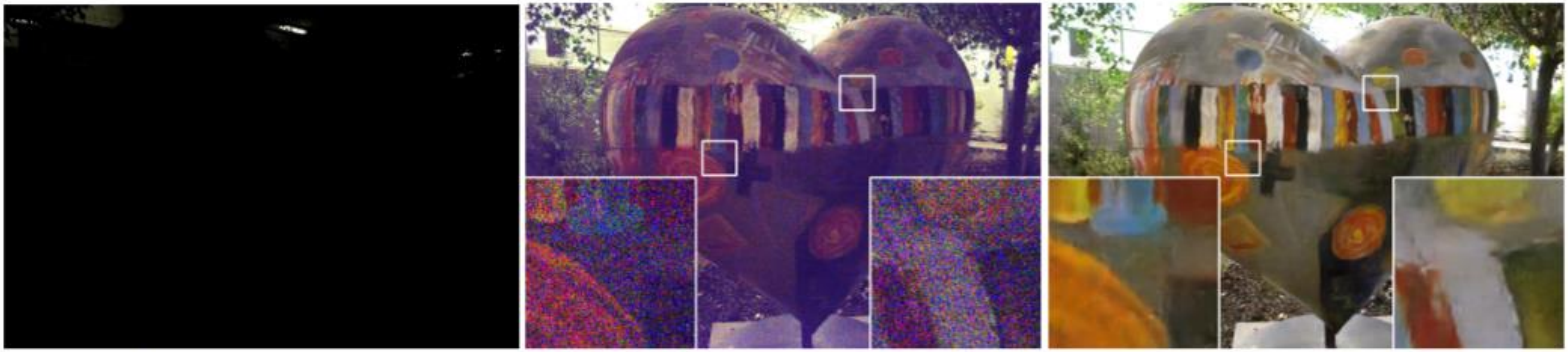
Figure 2. Example images in the SID dataset. Outdoor images in the top two rows, indoor images in the bottom rows. Long-exposure reference (ground truth) images are shown in front. Short-exposure input images (essentially black) are shown in the back. The illuminance at the camera is generally between 0.2 and 5 lux outdoors and between 0.03 and 0.3 lux indoors.

# Amplication Ratio

(a) x28 (b) x87 (c) x189 (d) x366

Figure 4. The effect of the amplification factor on a patch from an indoor image in the SID dataset (Sony x100 subset). The amplification factor is provided as an external input to our pipeline, akin to the ISO setting in cameras. Higher amplification factors yield brighter images. This figure shows the output of our pipeline with different amplification factors.

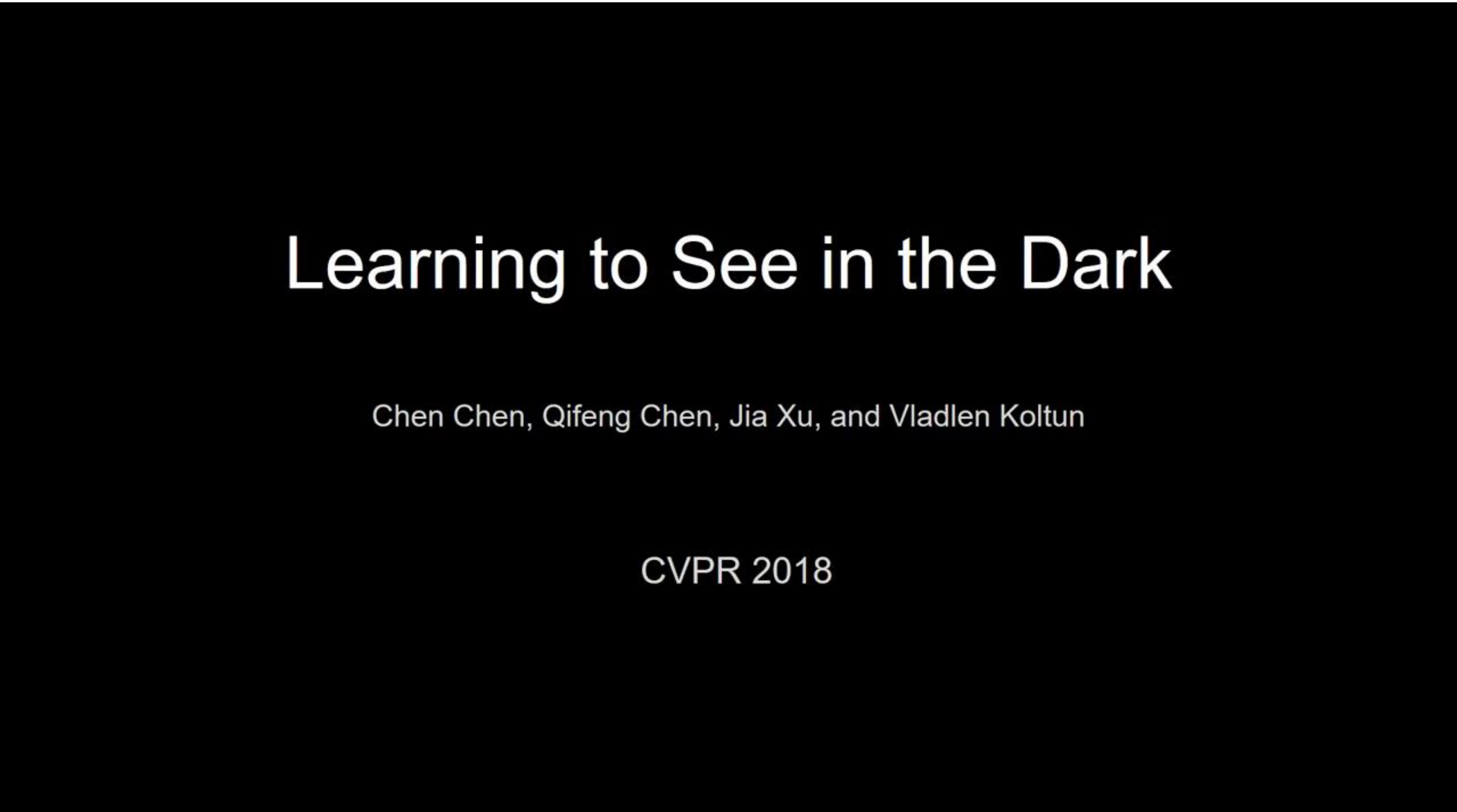
# Results



(a) JPEG image produced by camera (b) Raw data via traditional pipeline (c) Our result

Figure 5. (a) An image captured at night by the Fujifilm X-T2 camera with ISO 800, aperture f/7.1, and exposure of 1/30 second. The illuminance at the camera is approximately 1 lux. (b) Processing the raw data by a traditional pipeline does not effectively handle the noise and color bias in the data. (c) Our result obtained from the same raw data.

# Demo

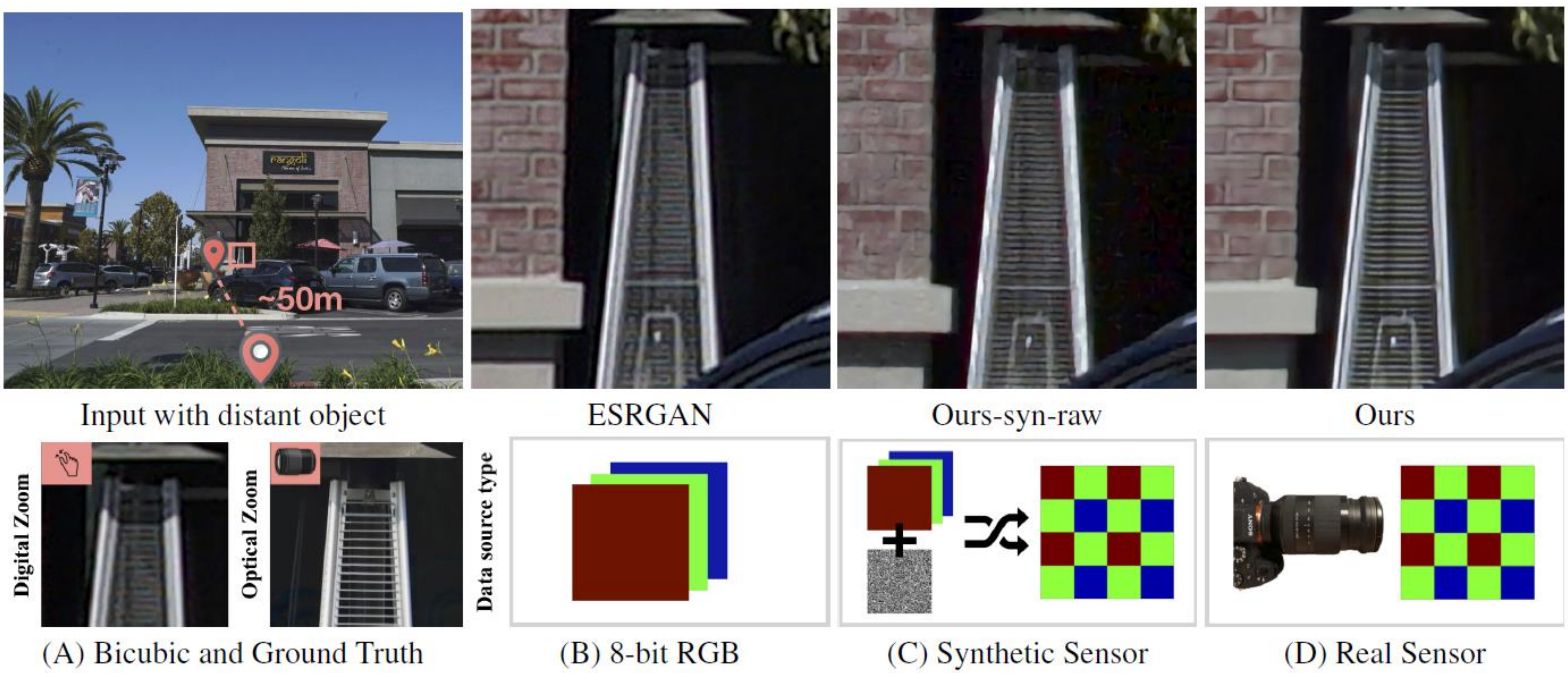

Learning to See in the Dark
Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun
CVPR 2018

# Results

| | Sony x300 set | Sony x100 set |
|---|---|---|
| Ours > BM3D | 92.4% | 59.3% |
| Ours > Burst | 85.2% | 47.3% |

Table 2. Perceptual experiments were used to compare the presented pipeline with BM3D and burst denoising. The experiment is skewed in favor of the baselines, as described in the text. The presented single-image pipeline still significantly outperforms the baselines on the challenging x300 set and is on par on the easier x100 set.

# Zoom to Learn, Learn to Zoom



# Data Collection

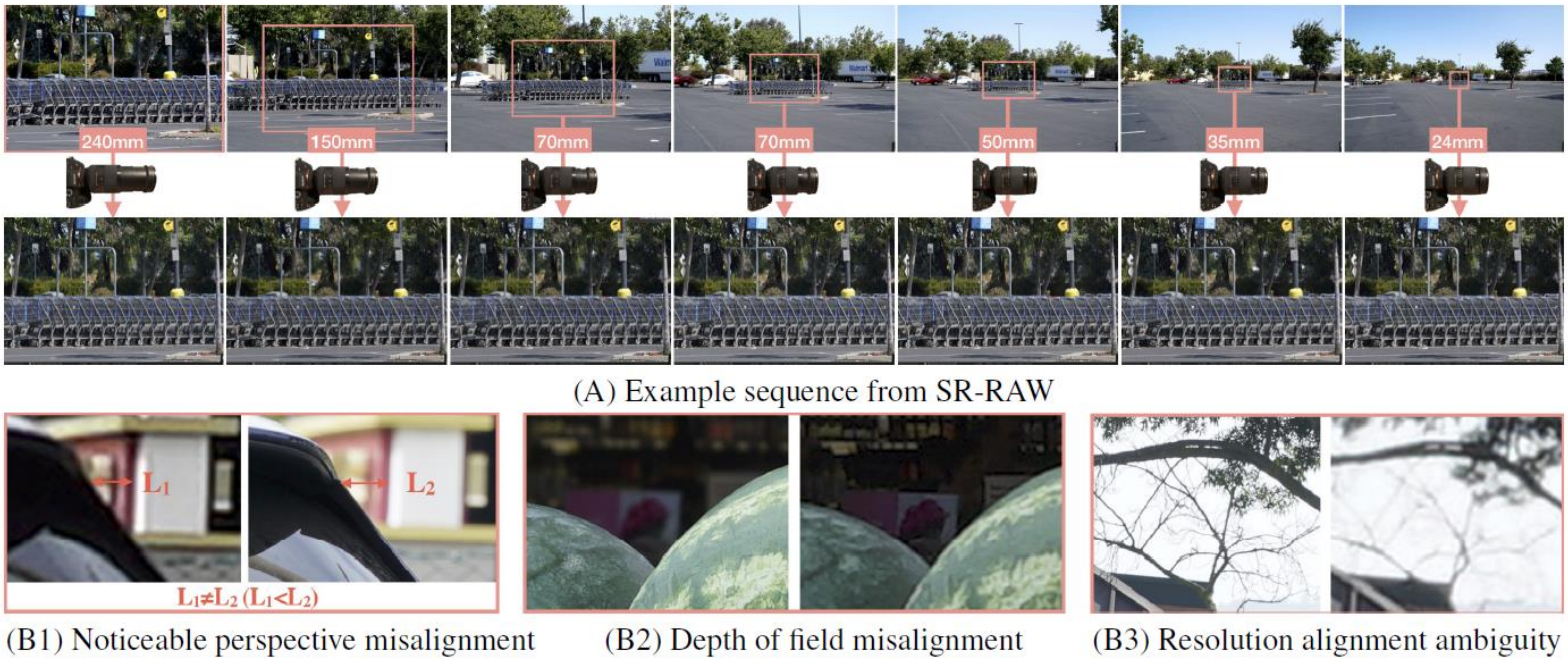
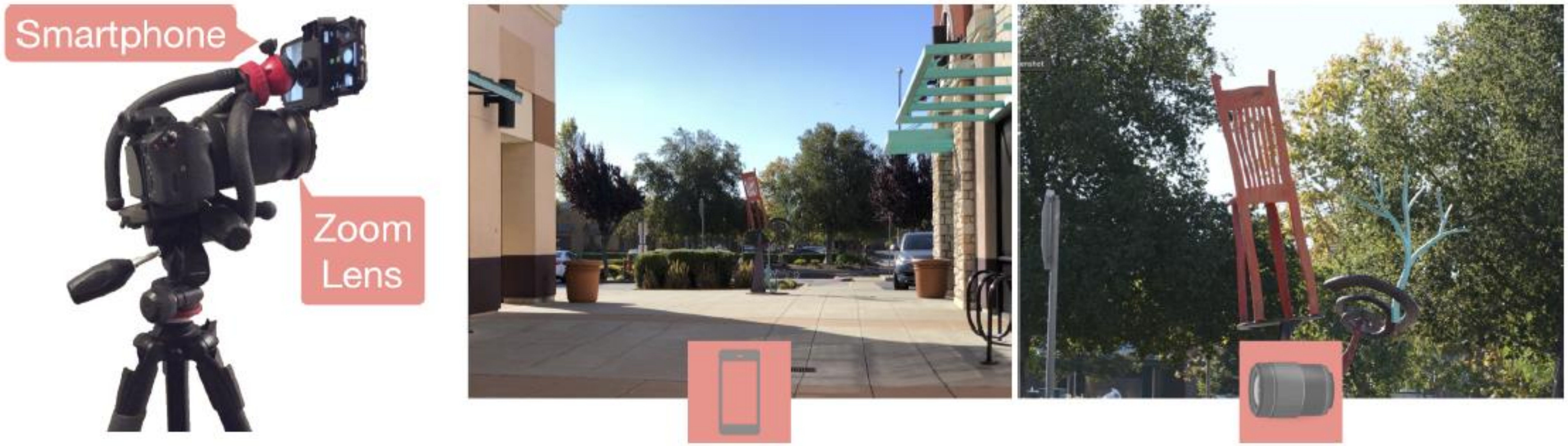



Figure 2: Example sequence from SR-RAW and three sources of misalignment in data capturing and pre-processing. The unavoidable misalignment drives us to propose a new similarity metric to correctly use SR-RAW for training.

# Data Collection

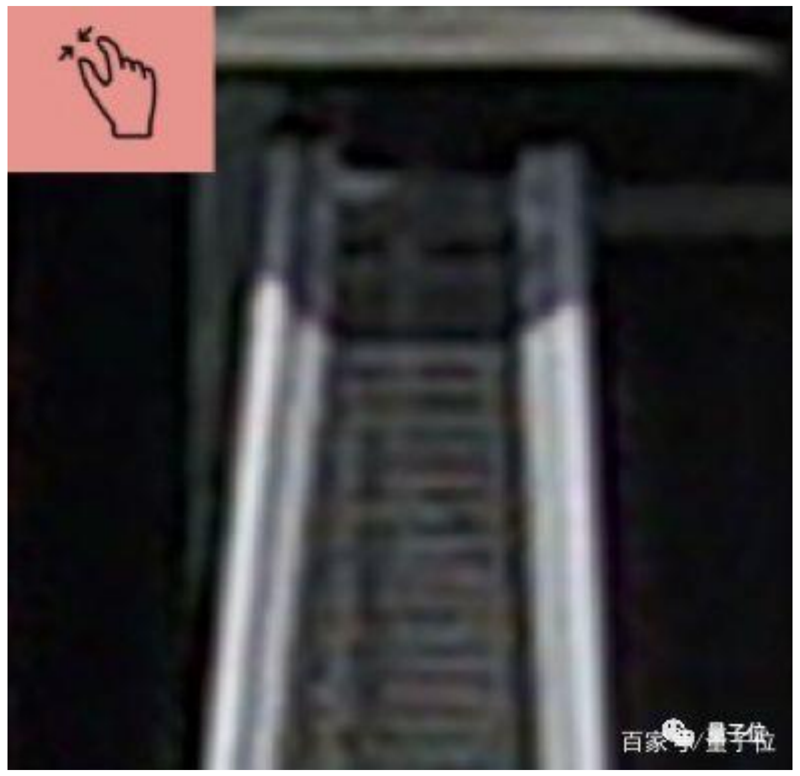



(A) Data Capture Setup (B) Example Smartphone Input (C) Example DSLR Target

Figure 2: Smartphone-DSLR data capture and an example data pair.

# What not just super-resolution with GANs?

- Existing super-resolution methods are trained on downsampled RGB images that contain little noise
- But in 8X digital zoom, noise is prominent
- RGB images are the output of ISP
  - High frequency is removed by denoising
- We train our model to recover underlying high-frequency details from noisy input



# Contextual Bilateral Loss

$$\mathrm{CX}(P,Q) = \frac{1}{N}\sum_{i}^{N} \min_{j=1,\dots,M}(\mathbb{D}_{p_i,q_j}).$$

Contextual Loss

$$\mathrm{CoBi}(P,Q) = \frac{1}{N}\sum_{i}^{N} \min_{j=1,\dots,M}(\mathbb{D}_{p_i,q_j} + w_s \mathbb{D}'_{p_i,q_j}), \quad (2)$$

where $\mathbb{D}'_{p_i,q_j} = \|(x_i,y_i),(x_j,y_j)\|_2$. $(x_i,y_i),(x_j,y_j)$ are spatial coordinates of features $p_i$ and $q_j$, respectively,

A novel loss (CoBi) for measuring similarity of slightly misaligned image pairs

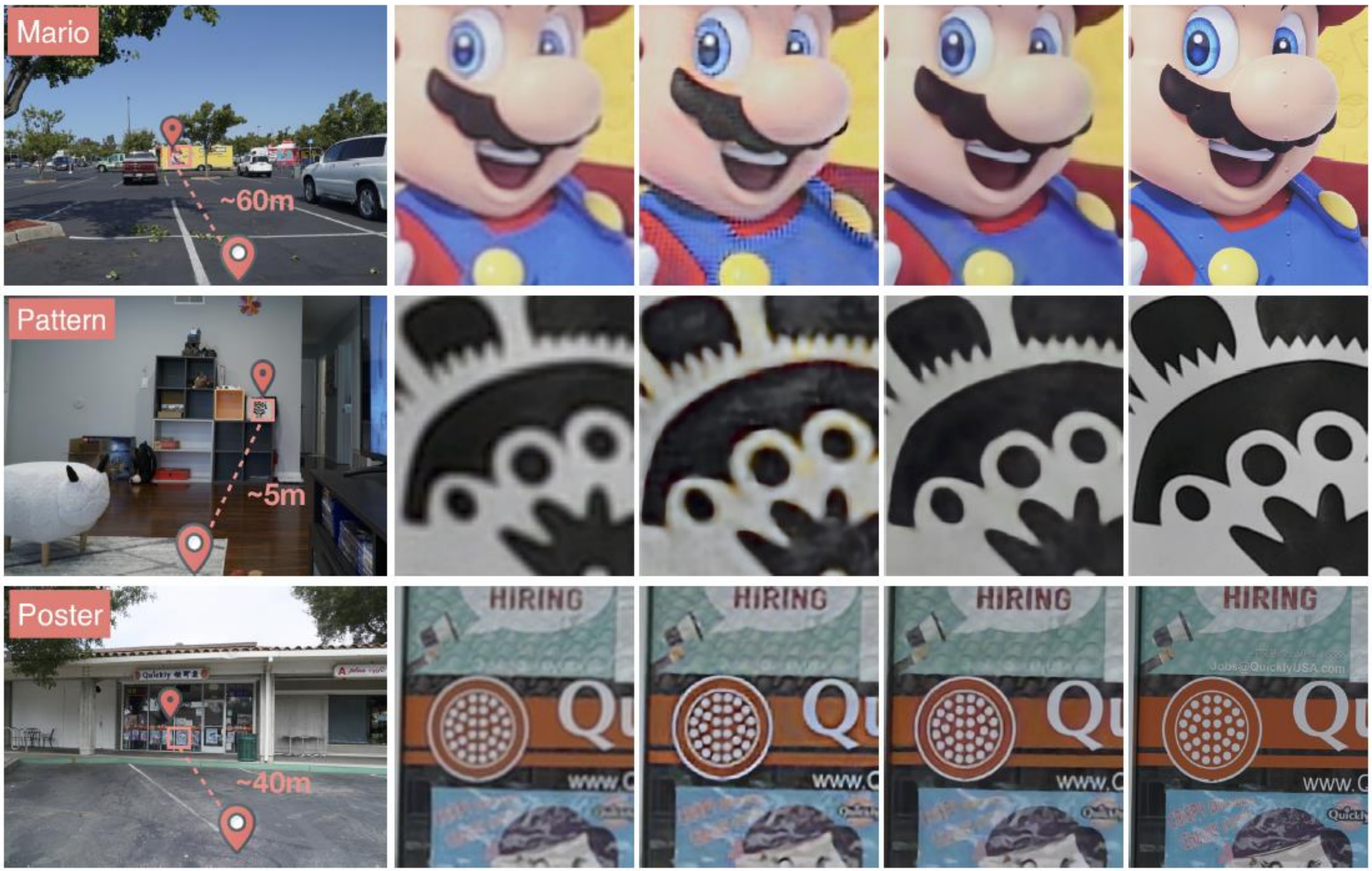
# Contextual Bilateral Loss


(A) Bicubic (B) Train with CX (C) Train with CoBi (D) Ground Truth

# Results


Figure 5: Our 4x zoom results show better perceptual performance in super-resolving distant objects against baseline methods that are trained under a synthetic setting and applied to processed RGB images.

# Results



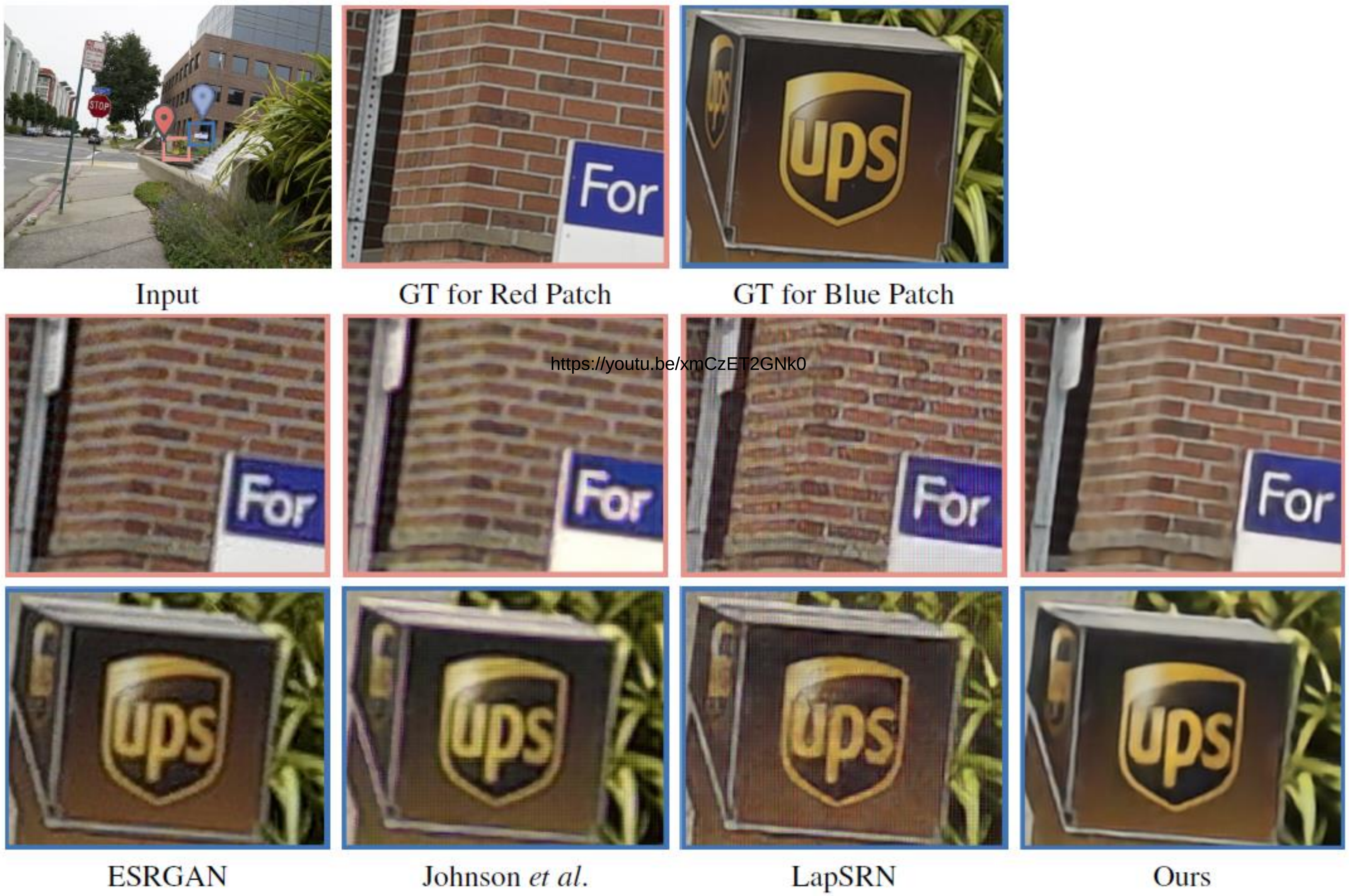

Input | Bicubic | Synthetic Sensor | Ours | GT

Figure 6: The model trained on synthetic sensor data produces artifacts such as jagged edges in "Mario" and "Poster" and color aberrations in "Pattern", while our model trained on real sensor data produces clean and high quality zoomed images.

# Results


Input
GT for Red Patch
GT for Blue Patch
32115
ESRGAN
Johnson *et al.*
LapSRN
Ours

# Results




https://youtu.be/xmCzET2GNk0

Figure 4 (Cont.): Our 4X zoom results show better perceptual performance in super-resolving distant objects against baseline methods that are trained under a synthetic setting and applied to processed RGB images.

# Going well

# A hazy day

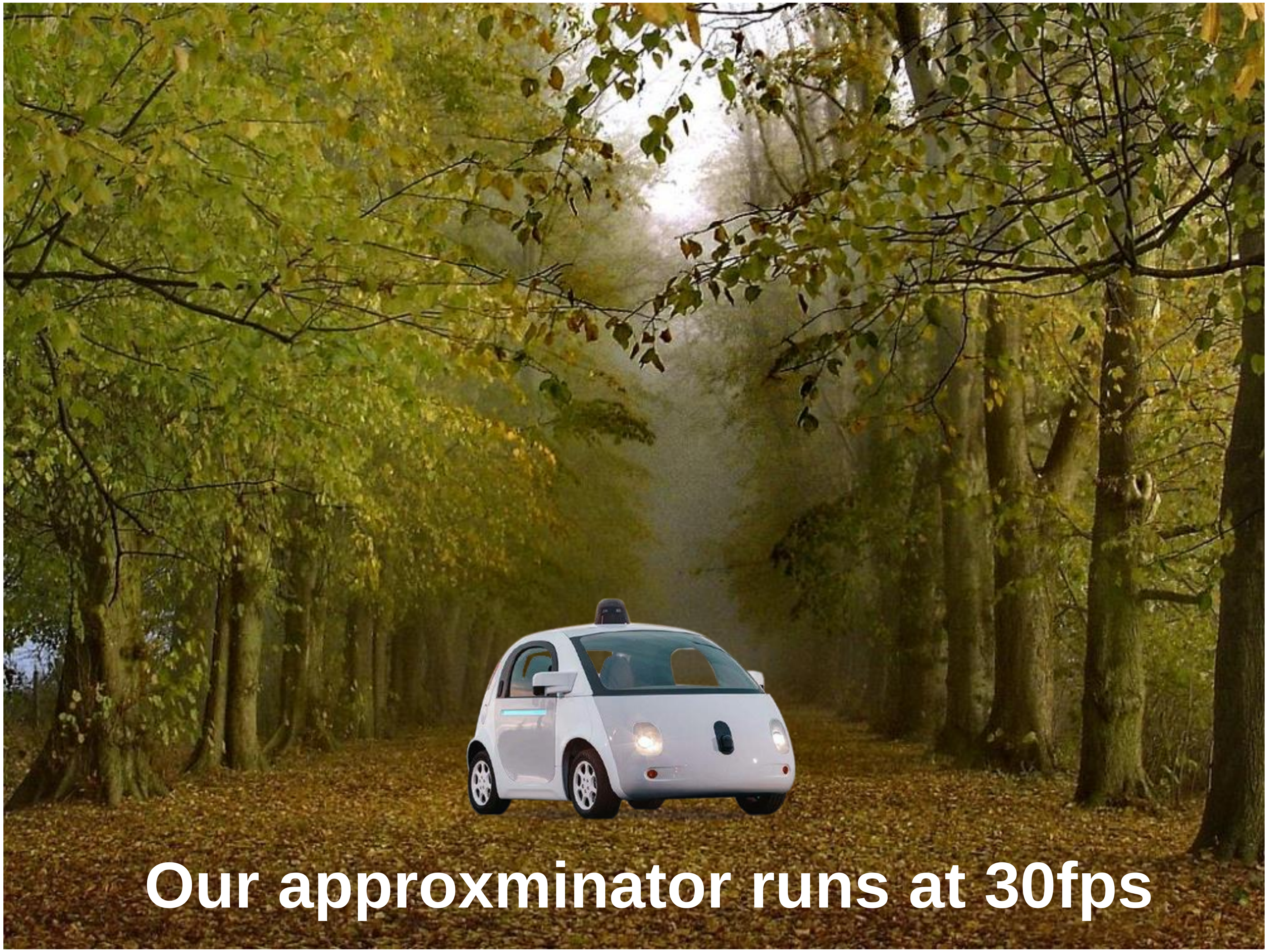
# Dehazed image

# But not practical

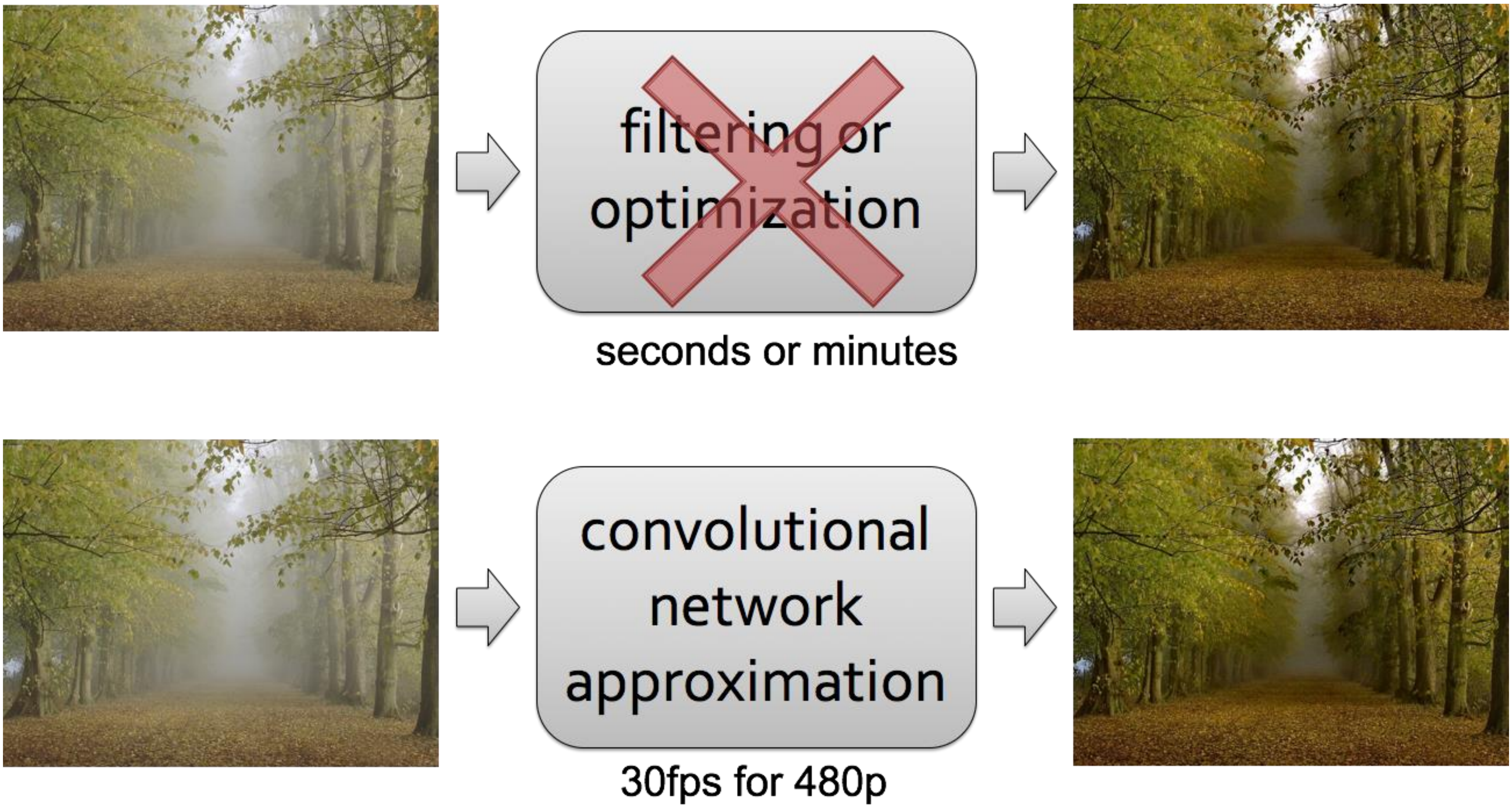
# Alternative solutions?

- Use another method
  - No state-of-the-art accuracy
- Accelerate implementation
  - Time consuming
- **Nonlinear Function Approximator**
  - Simple, general, accurate and fast

# Real-time performance

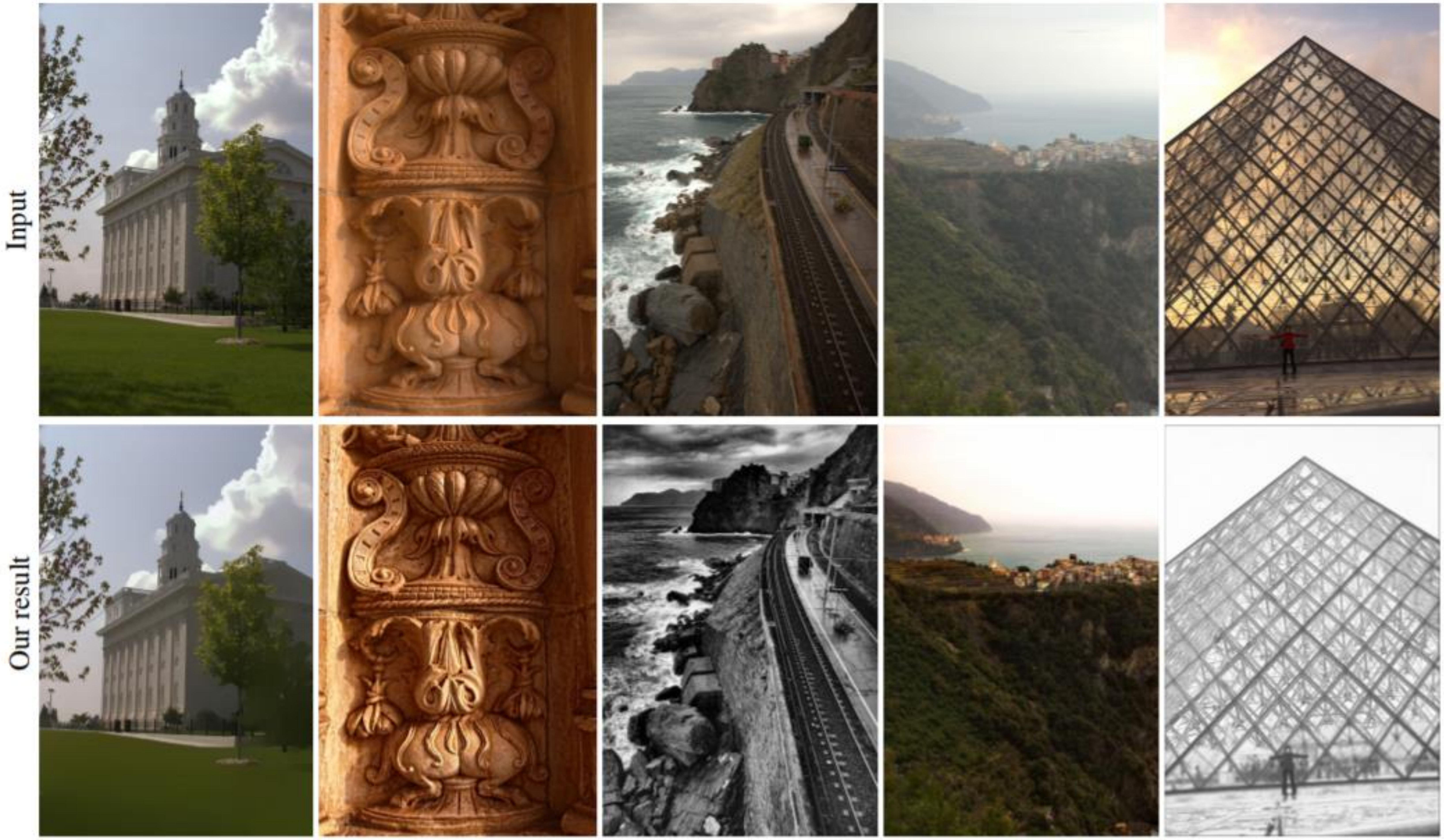


# Fast Image Processing




Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast Image Processing with Fully-Convolutional Networks, ICCV 2017

# Results



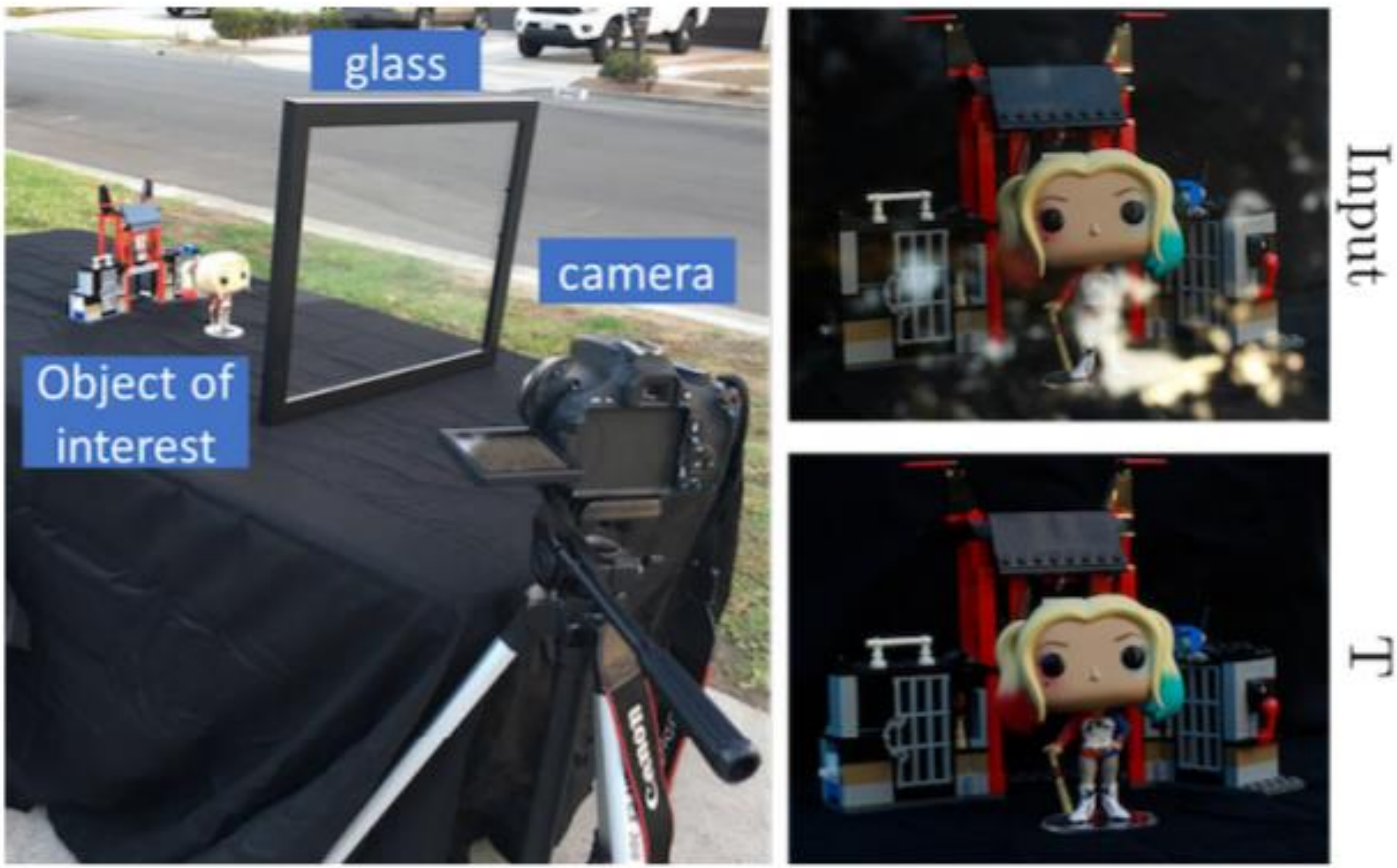

$L_0$ smoothing | Multiscale tone | Photographic style | Nonlocal dehazing | Pencil drawing

# Demo


Fast Image Processing with
Fully-Convolutional Networks
Qifeng Chen* Jia Xu* Vladlen Koltun
Intel Labs
* Joint first authors

# Single Image Reflection Removal

# Data Collection

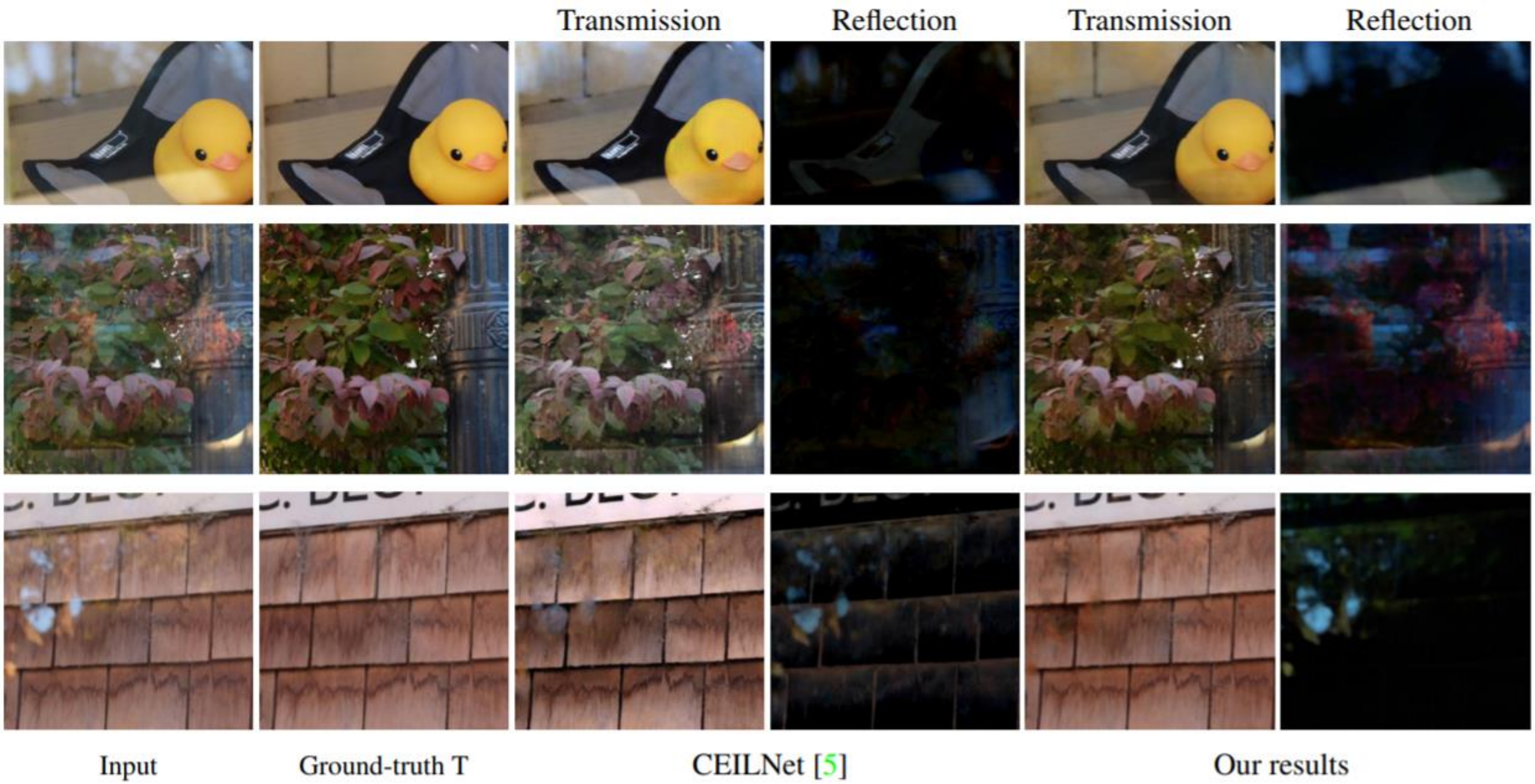


# Method


(a) Input | (b) Without $L_{\text{feat}}$ | (c) Without $L_{\text{adv}}$ | (d) Without $L_{\text{excl}}$ | (e) Complete model

Figure 2: Visual comparisons on the three perceptual loss functions, evaluated on a real-world image. In (b), we replace $L_{\text{feat}}$ with image space $L^1$ loss and observed overly-smooth output. (c) shows artifacts of color degradation and noticeable residuals without $L_{\text{adv}}$. In (d), the lack of $L_{\text{excl}}$ makes the predicted transmission have undesired reflection residuals. Our complete model in (e) is able to produce better and cleaner prediction.

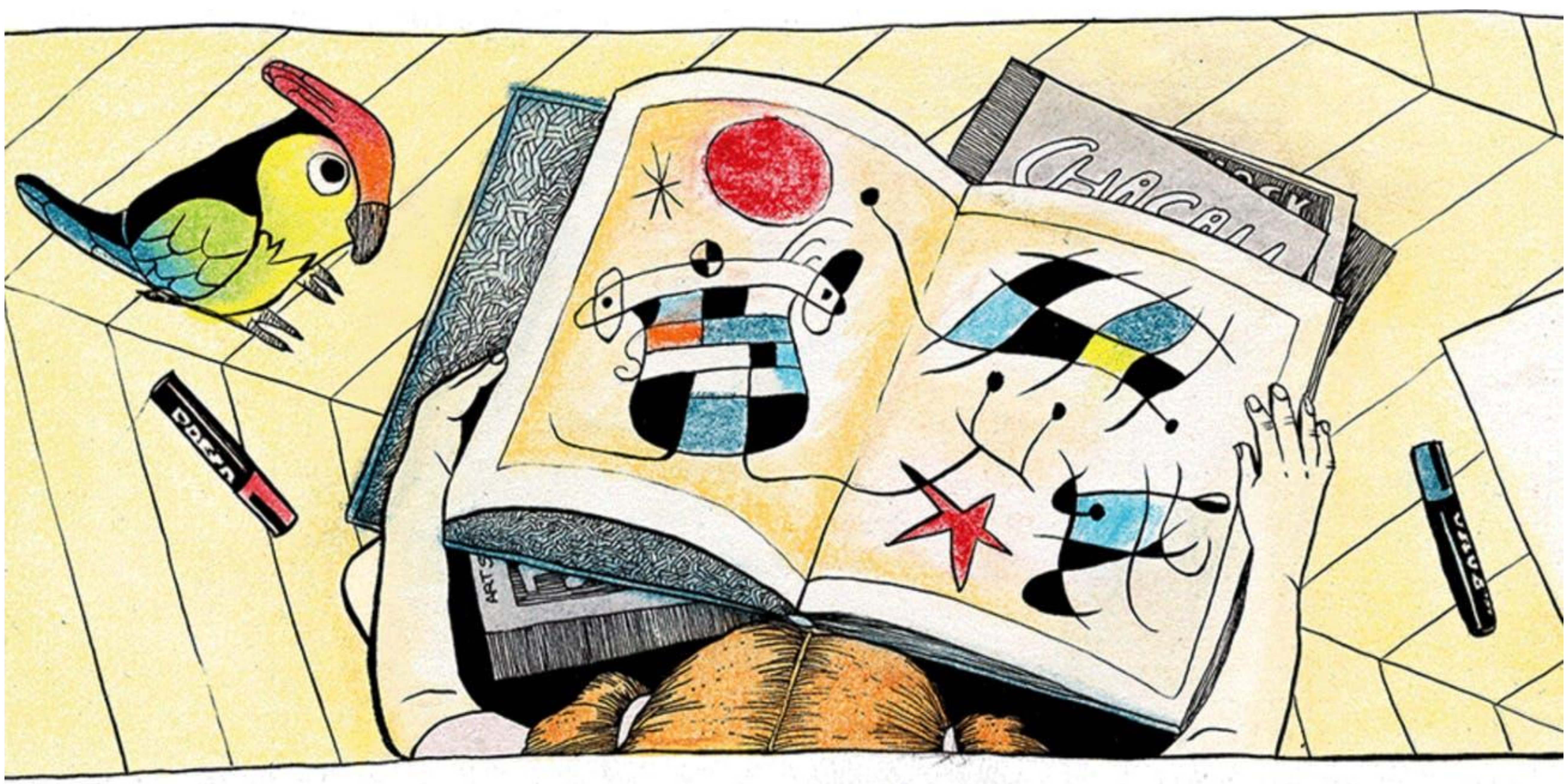
# Results



Transmission
Reflection
Transmission
Reflection
Input
Ground-truth T
CEILNet [5]
Our results

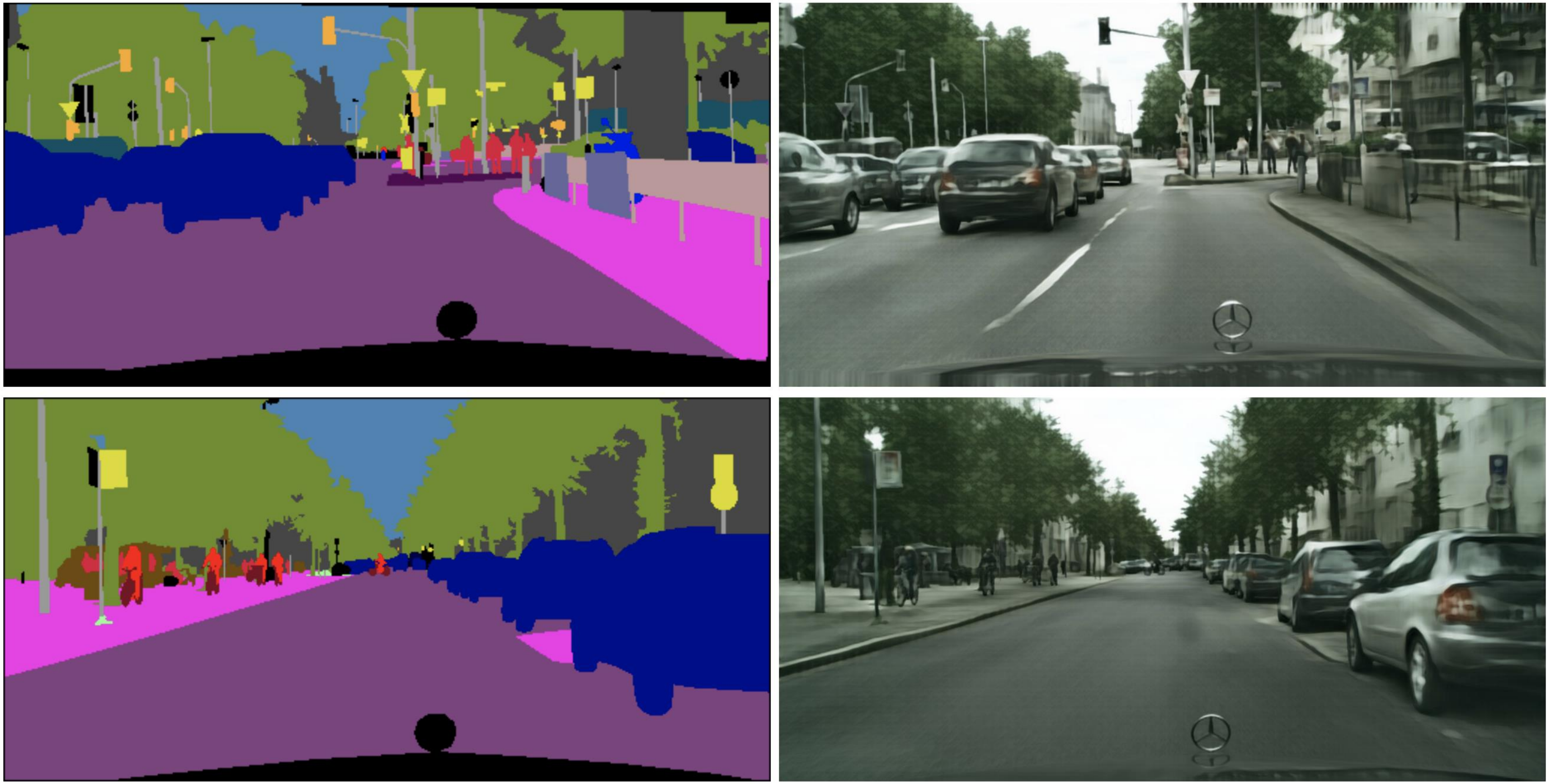
# Deep Image and Video Synthesis

# Art by Human Creation



# Art by Human Creation & AI

# Photographic image synthesis



Input semantic layouts

Synthesized images

Qifeng Chen and Vladlen Koltun. Photographic Image Synthesis with Cascaded Refinement Networks. ICCV 2017

# Motivation

- Computer graphics
  - Alternative route to photorealism
  - Capture photographic appearance
  - Fast image synthesis

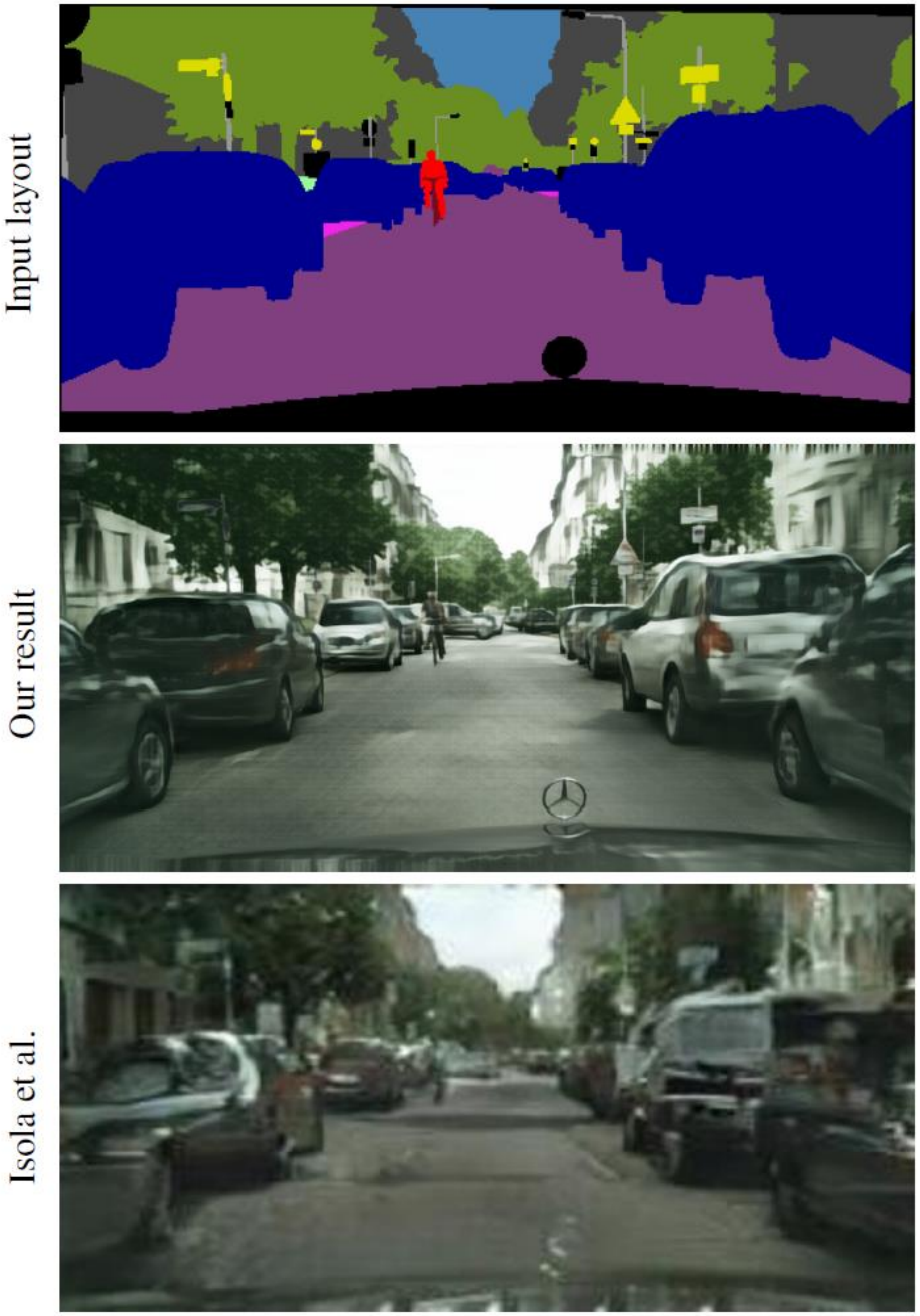
CARLA
Dosovitskiy et al., CoRL 2017

# Motivation

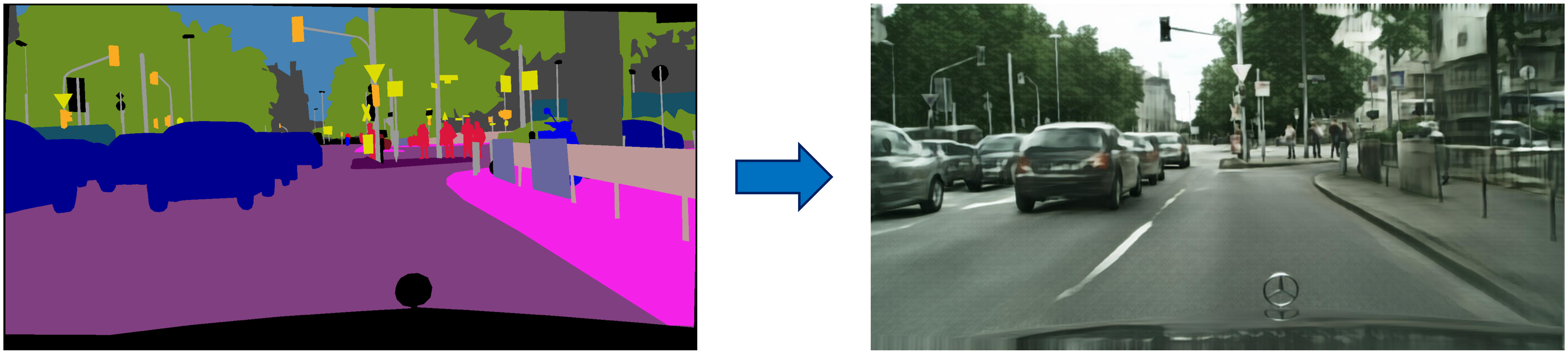
- Artificial Intelligence
  - Visual Imagination

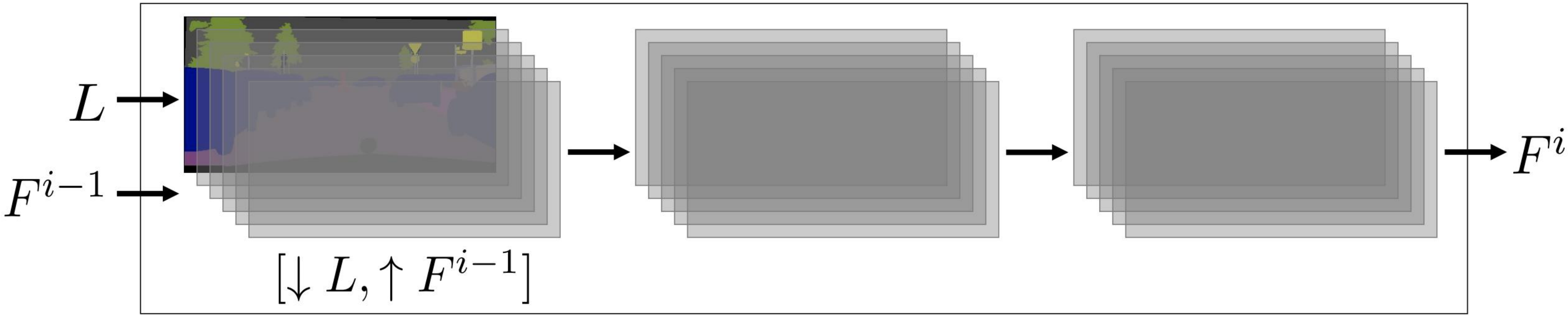
# Our approach

- Cascaded refinement networks
- Perceptual Loss
- Diversity



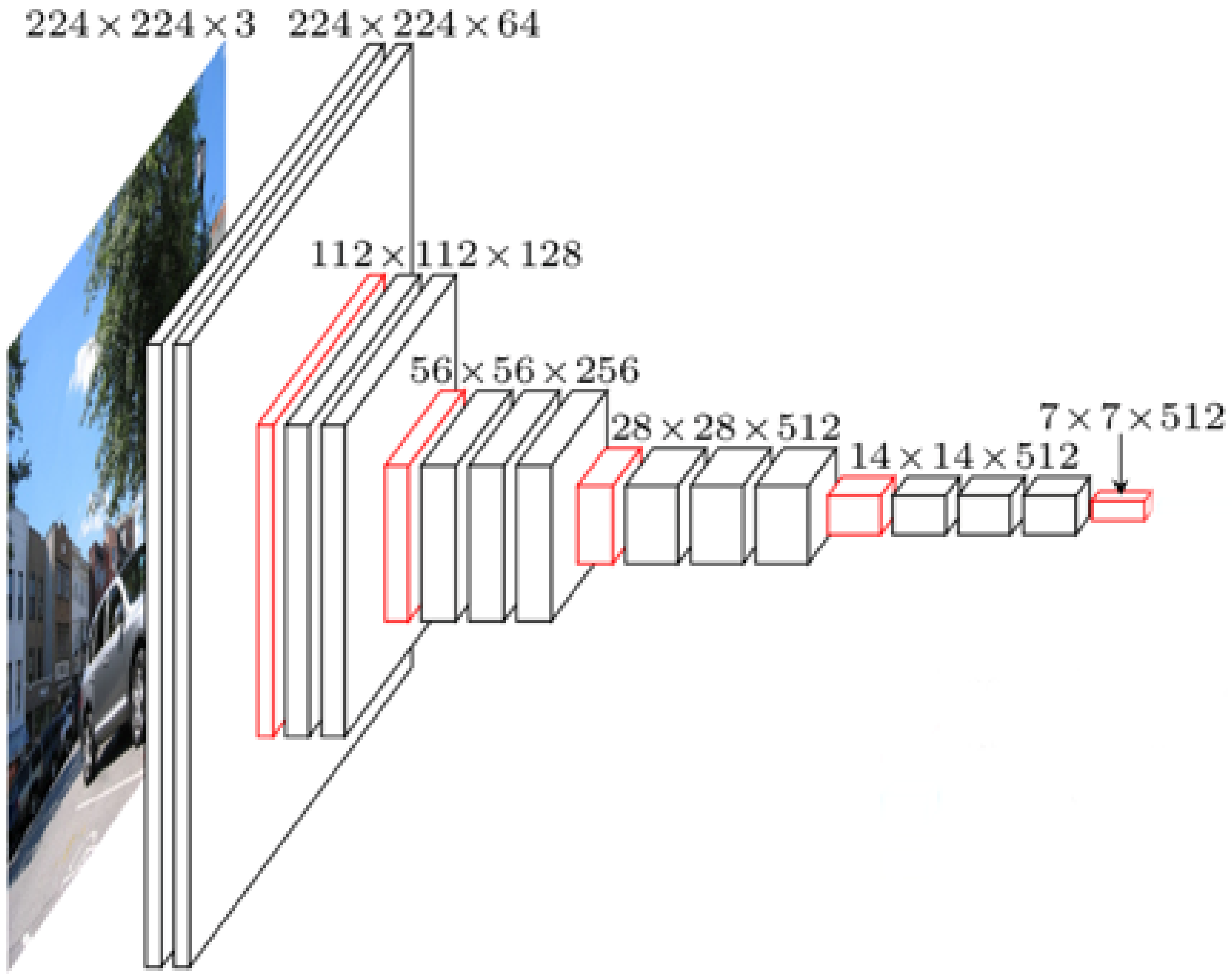
# Cascaded refinement networks






High Resolution

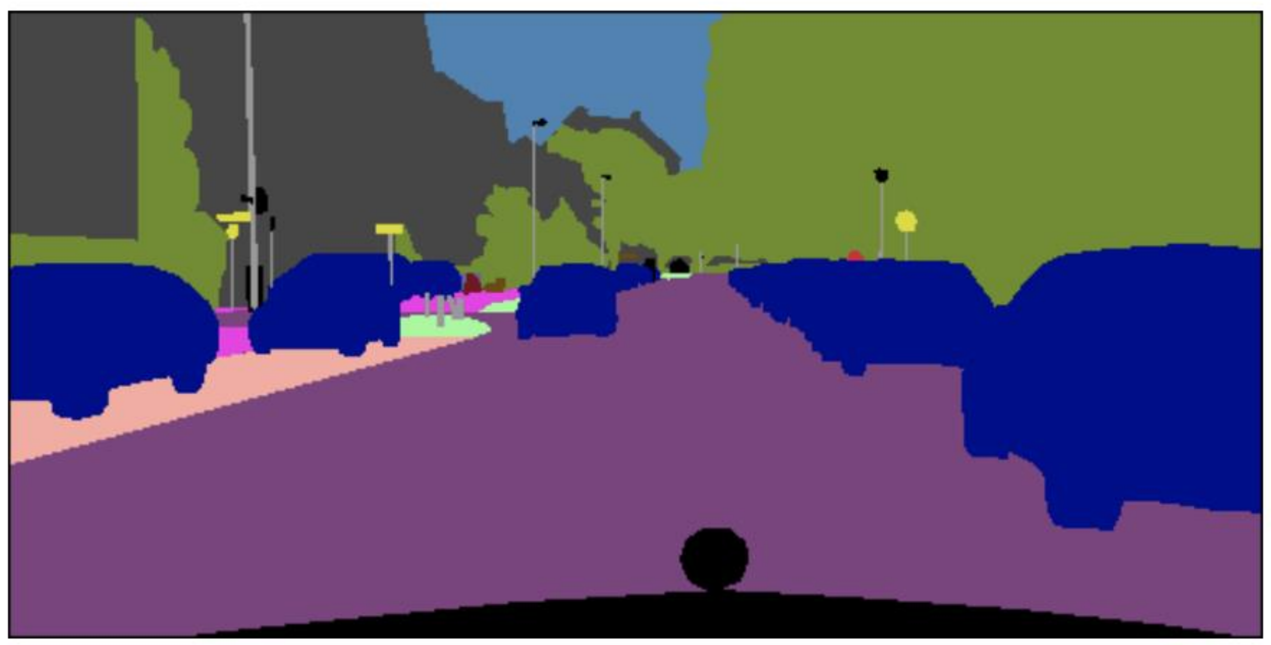
# Perceptual Loss




$$\mathcal{L}_{I,L}(\theta) = \sum_l \lambda_l \|\Phi_l(I) - \Phi_l(g(L;\theta))\|_1 .$$

# Diversity

# Comparisons on Cityscapes



Semantic layout

GAN+semantic segmenation

Full-resolution network

Our result

Isola et al. [16]

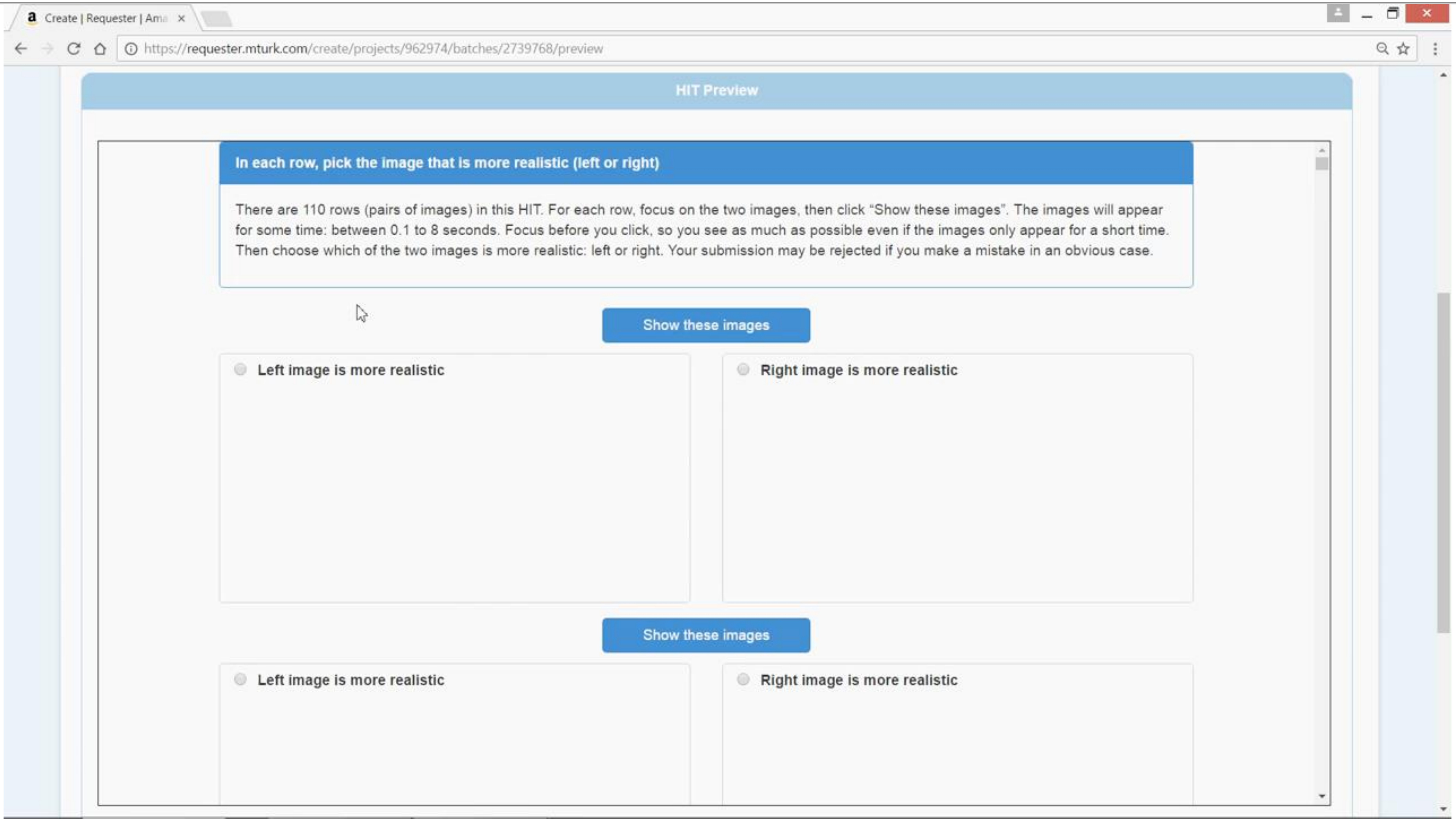
Encoder-decoder

# Results on NYU dataset

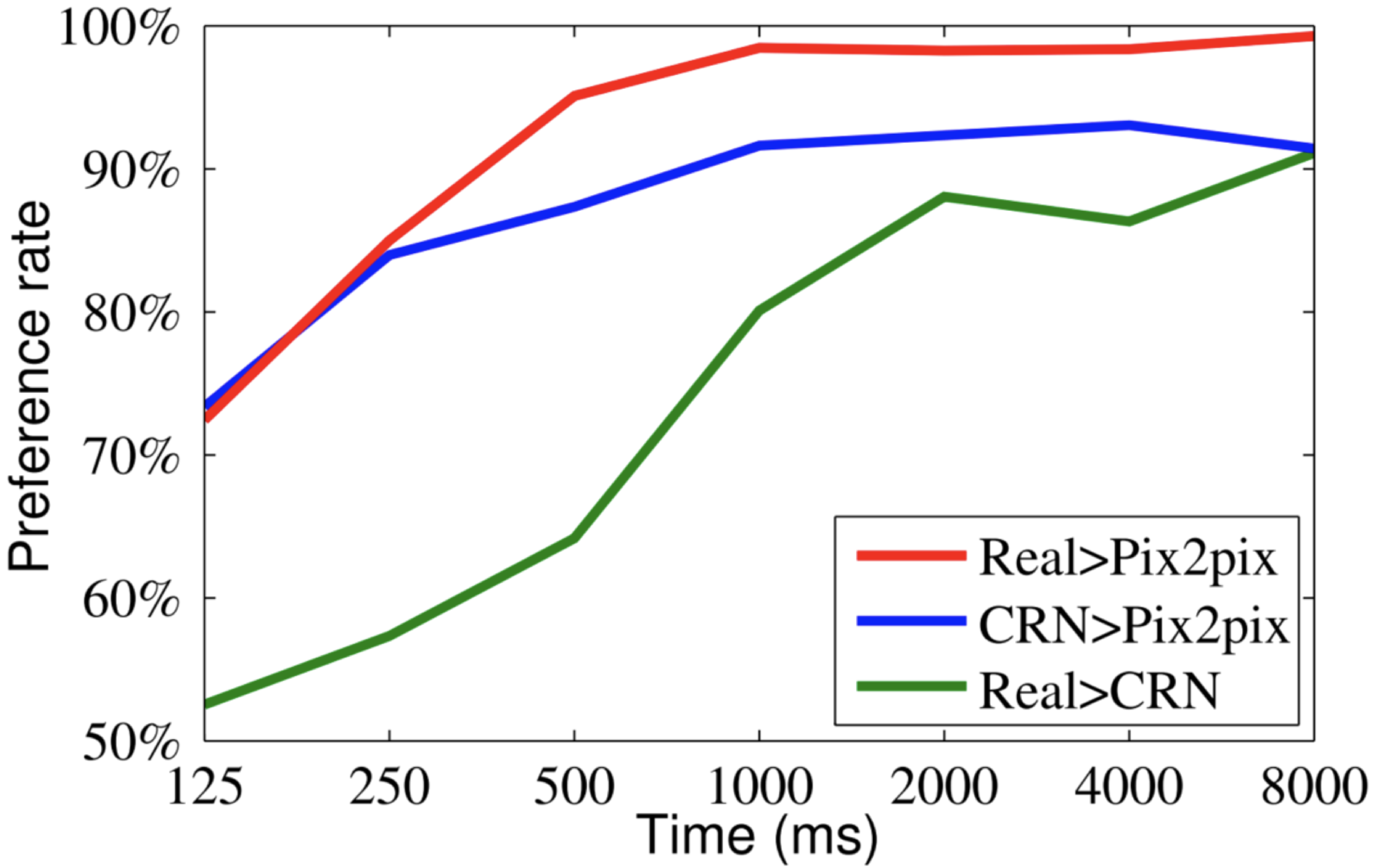

Semantic layout | Our result | Isola et al. [16] | Full-resolution network | Encoder-decoder

Figure 6. Qualitative comparison on the NYU dataset.

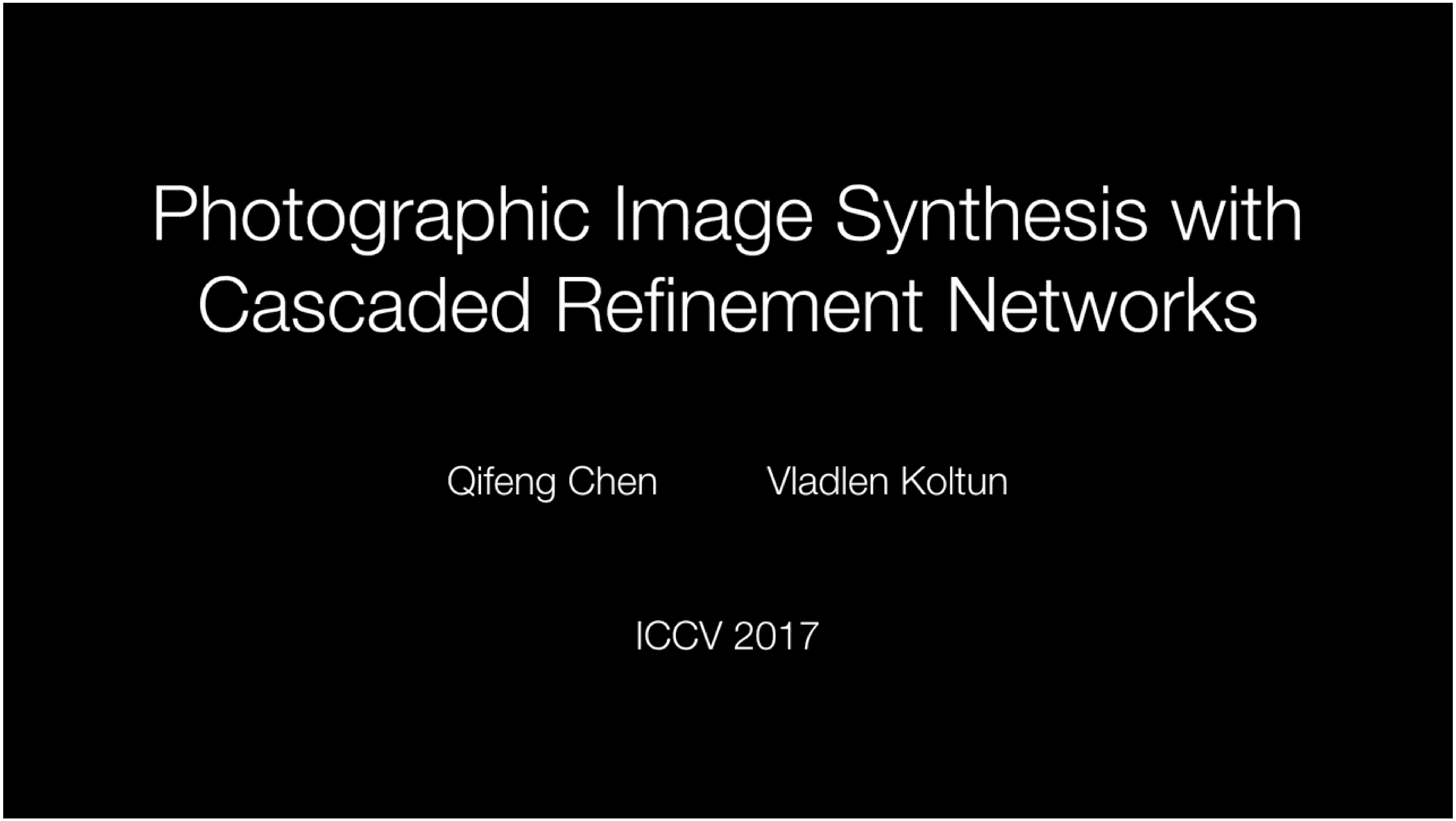
# User Study


Create | Requester | Ama
https://requester.mturk.com/create/projects/962974/batches/2739768/preview
HIT Preview
In each row, pick the image that is more realistic (left or right)
There are 110 rows (pairs of images) in this HIT. For each row, focus on the two images, then click "Show these images". The images will appear for some time: between 0.1 to 8 seconds. Focus before you click, so you see as much as possible even if the images only appear for a short time. Then choose which of the two images is more realistic: left or right. Your submission may be rejected if you make a mistake in an obvious case.
Show these images
Left image is more realistic
Right image is more realistic
Show these images
Left image is more realistic
Right image is more realistic

# User study


100%
90%
80%
70%
60%
50%
Preference rate
125
250
500
1000
2000
4000
8000
Time (ms)
Real>Pix2pix
CRN>Pix2pix
Real>CRN

# GTA5 and Demo Video



# Semi-parametric Image Synthesis


Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun Semi-parametric Image Synthesis. CVPR 2018

# Image Synthesis

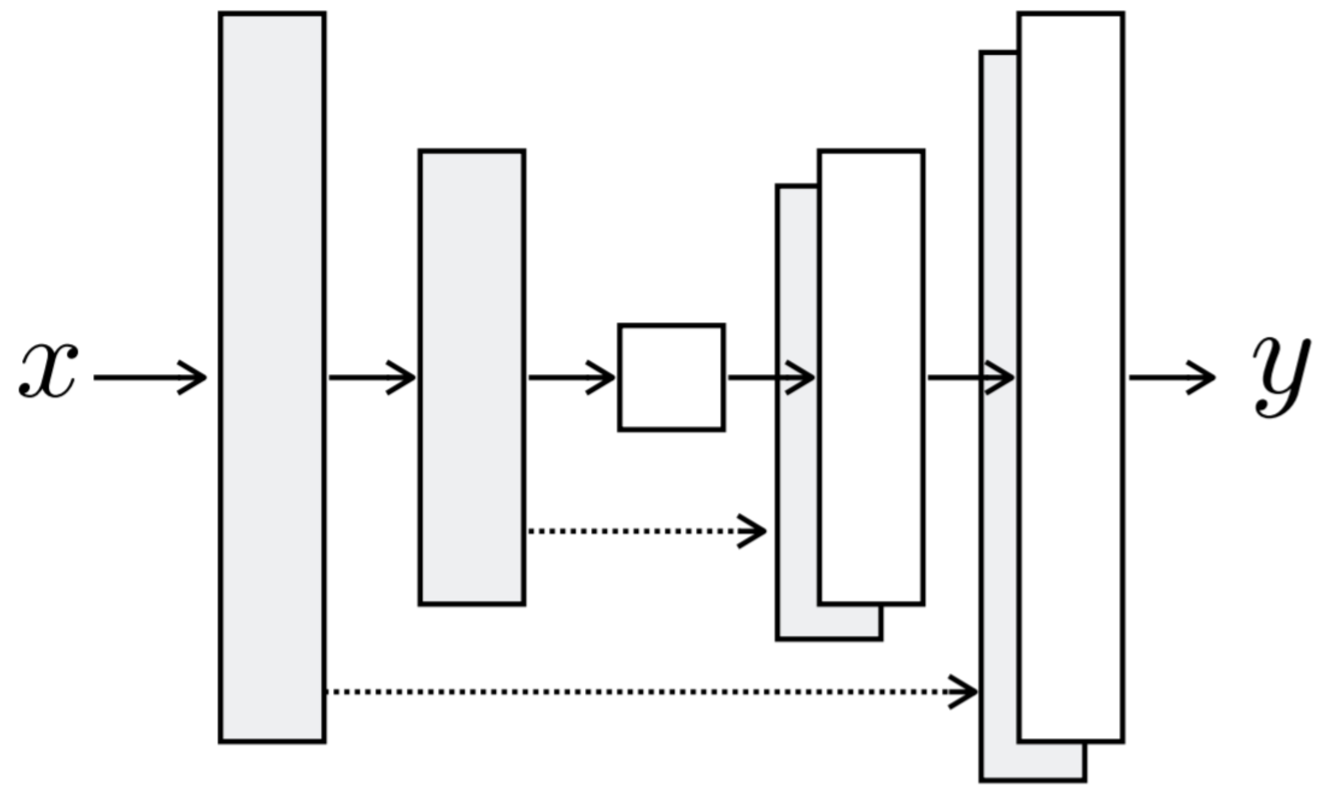
Semantic layouts

Our result

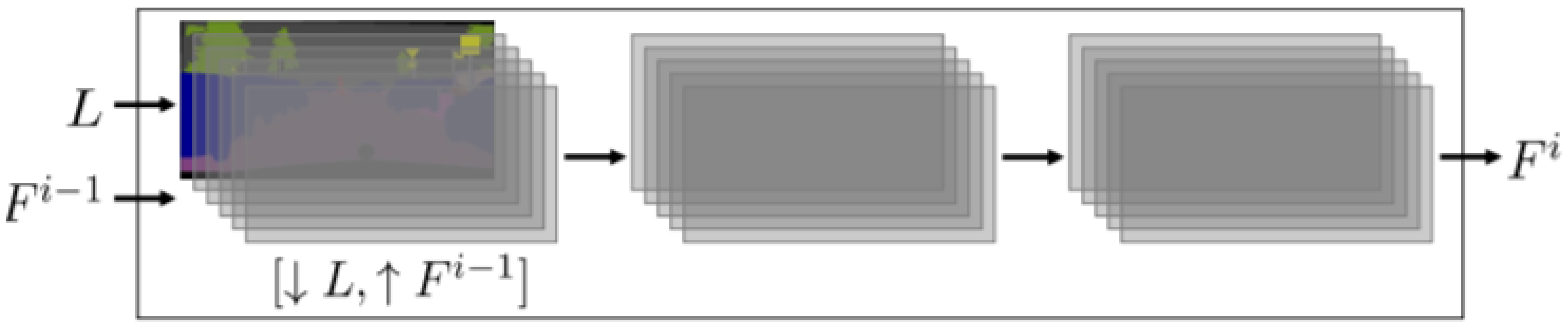
NYU dataset [Silberman et al. ECCV 2012]

ADE20K dataset [Zhou et al. 2017]

# Prior Work: Parametric Models

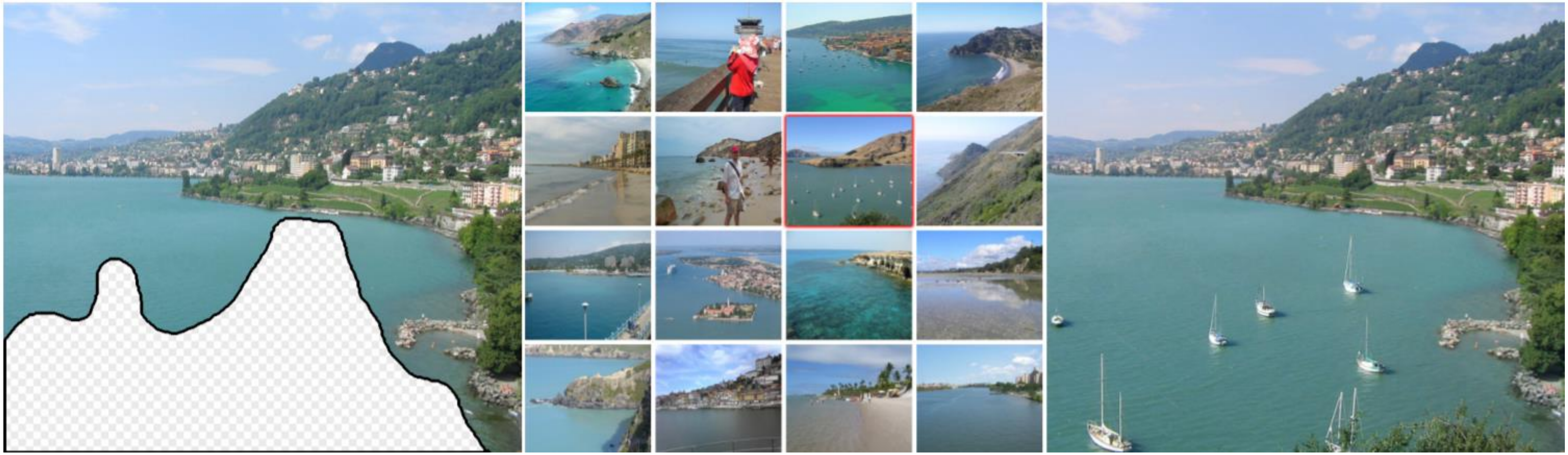



Pix2pix [Isola et al. 2017]



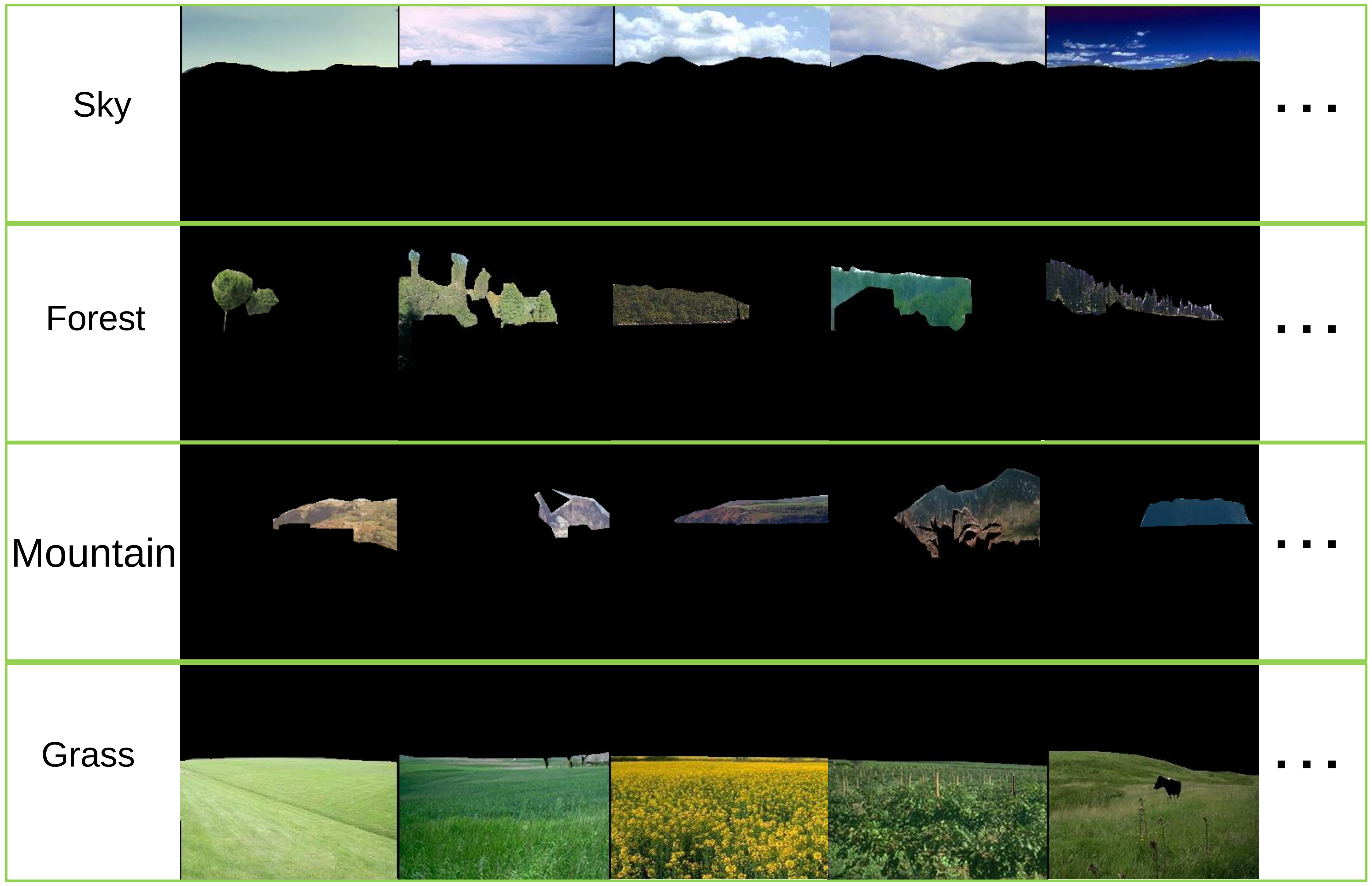

CRN [Chen and Koltun 2017]

# Prior Work: Non-parametric Models



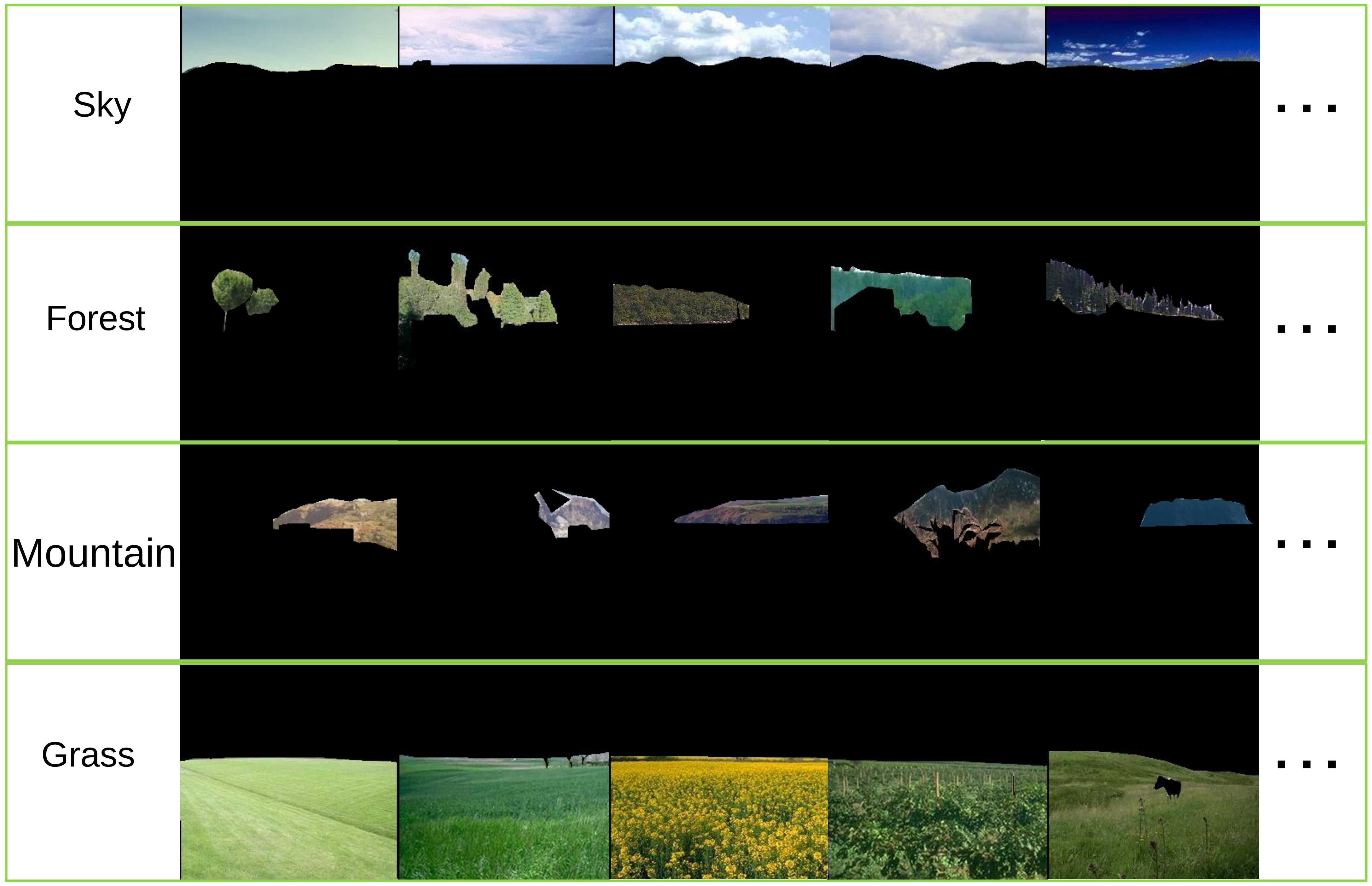
Scene Completion using Millions of Photographs [Hays and Efros 2007]

# Our Approach

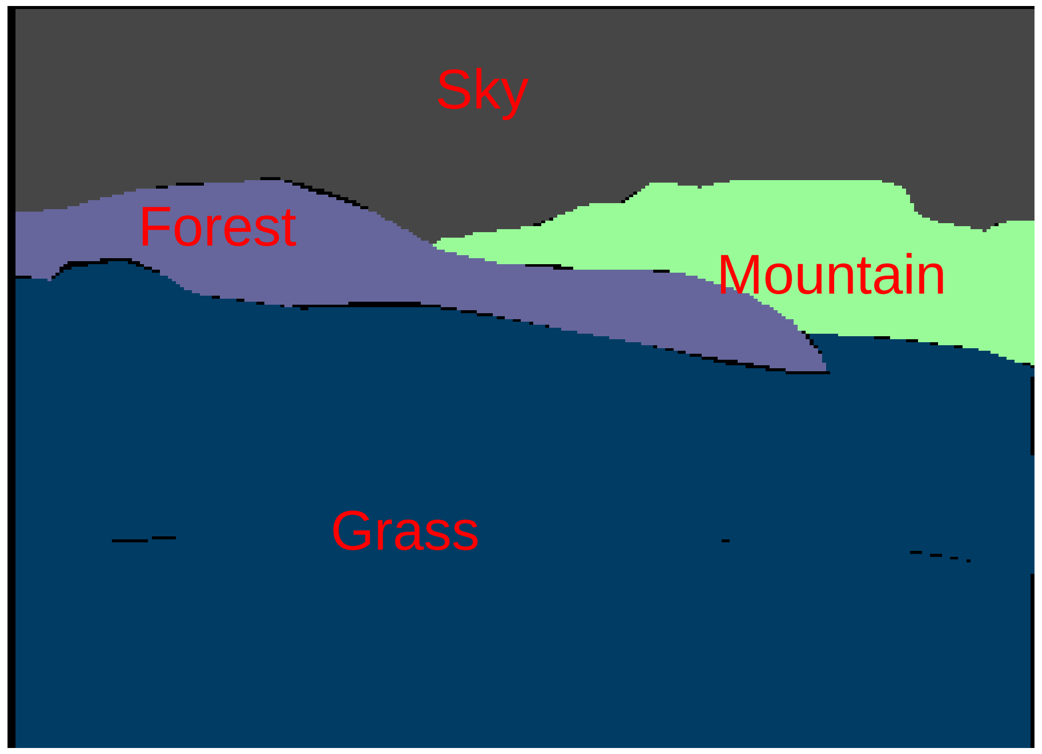



External memory

# Our Approach

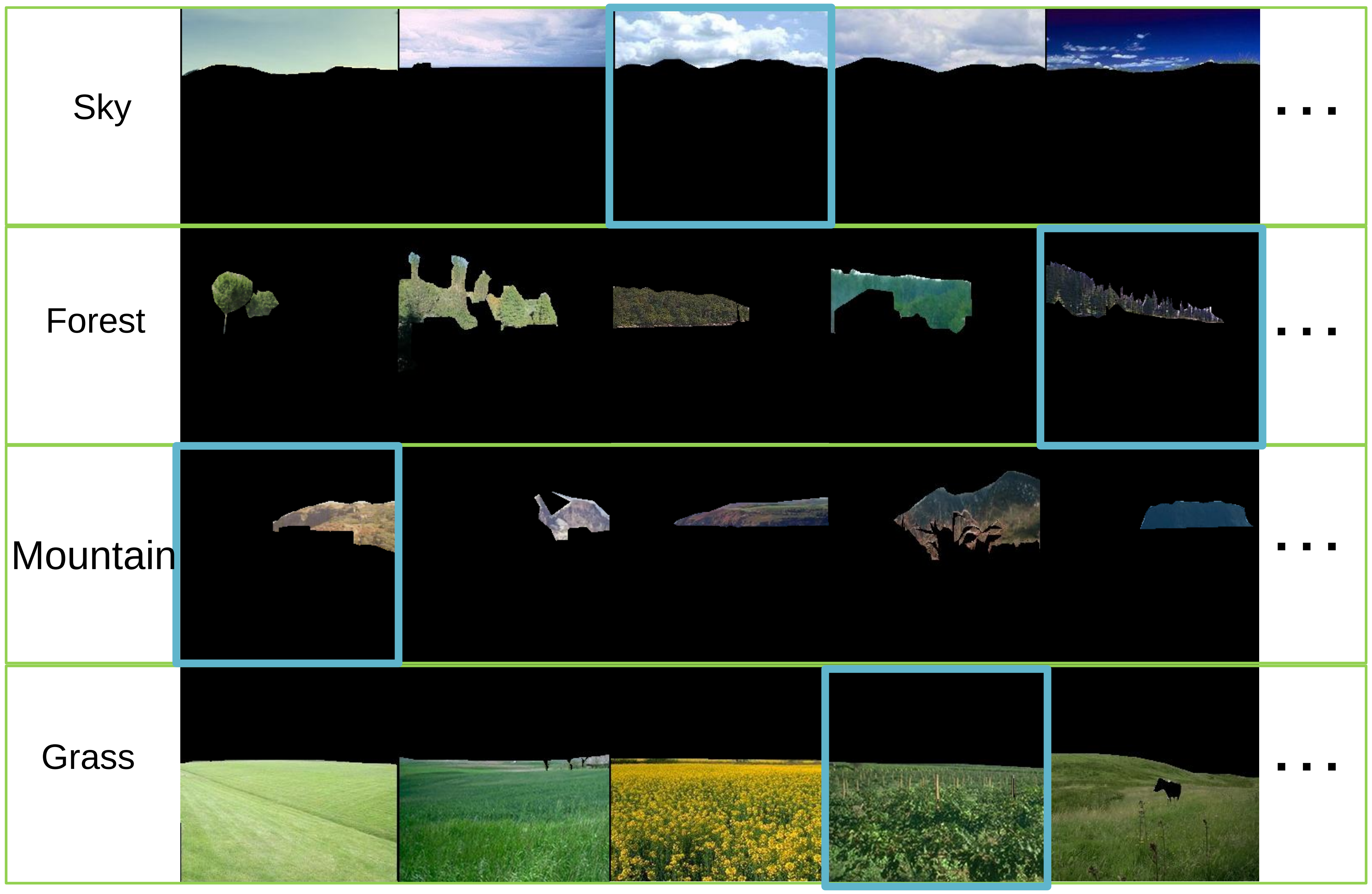



External memory

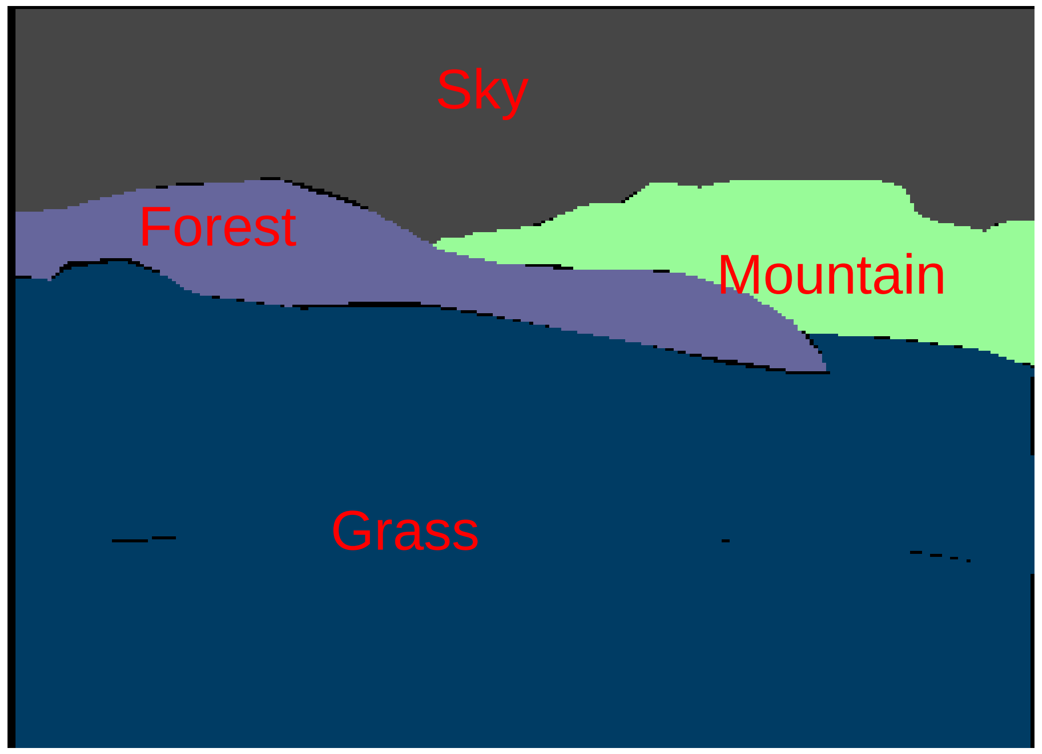



Semantic layout

# Our Approach

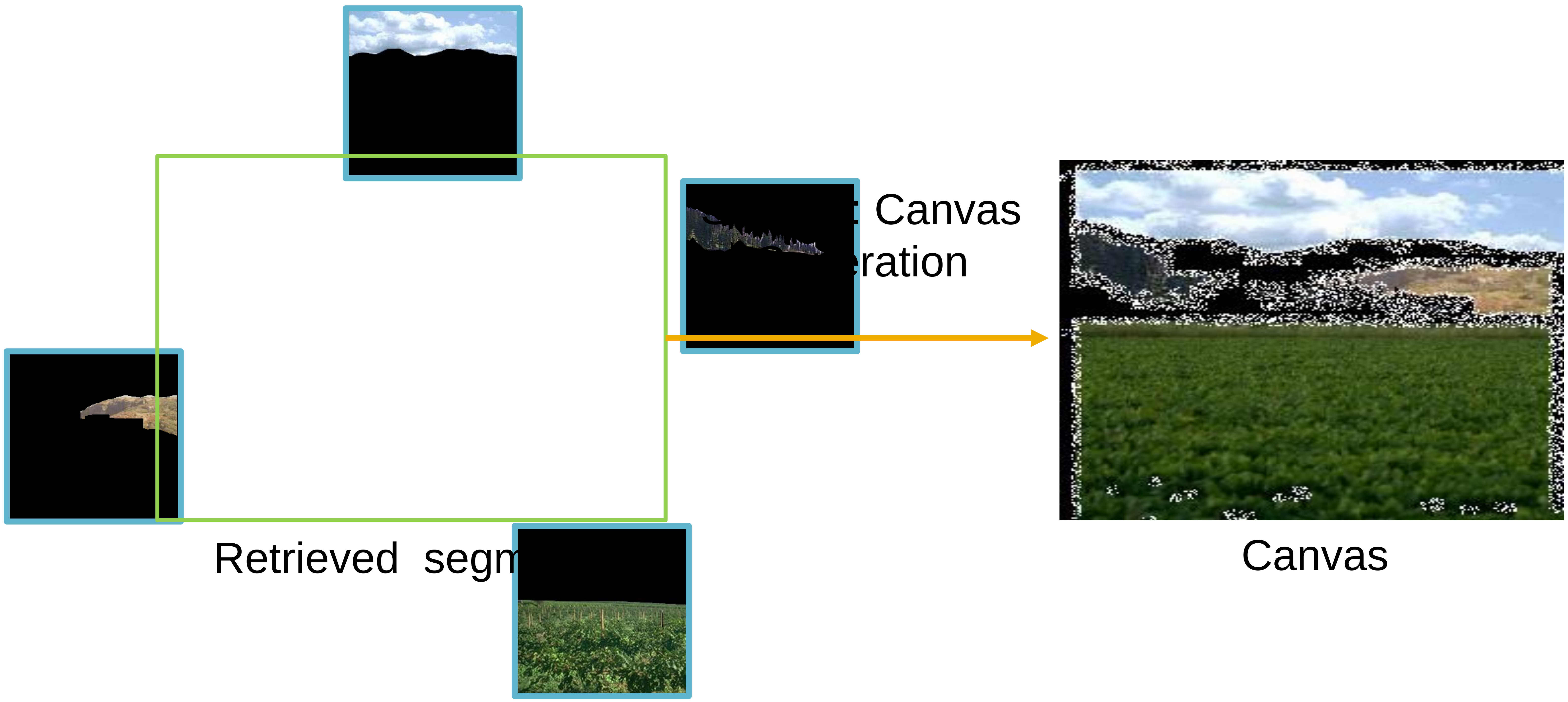



External memory



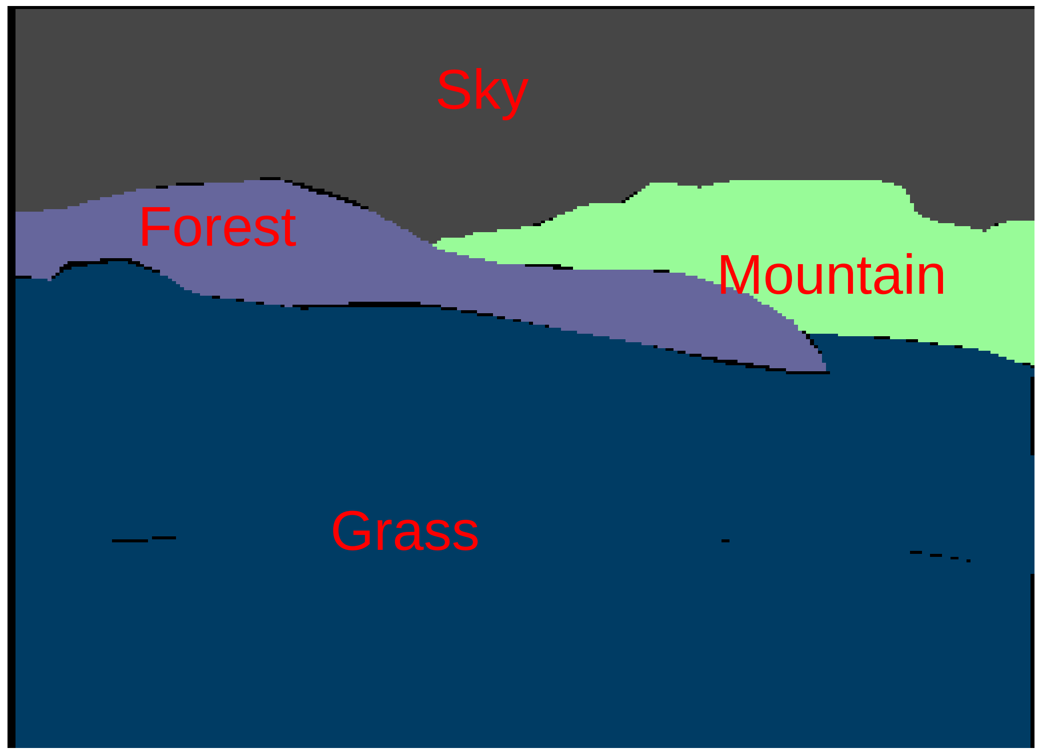

Semantic layout

# Our Approach



Canvas

# Our Approach

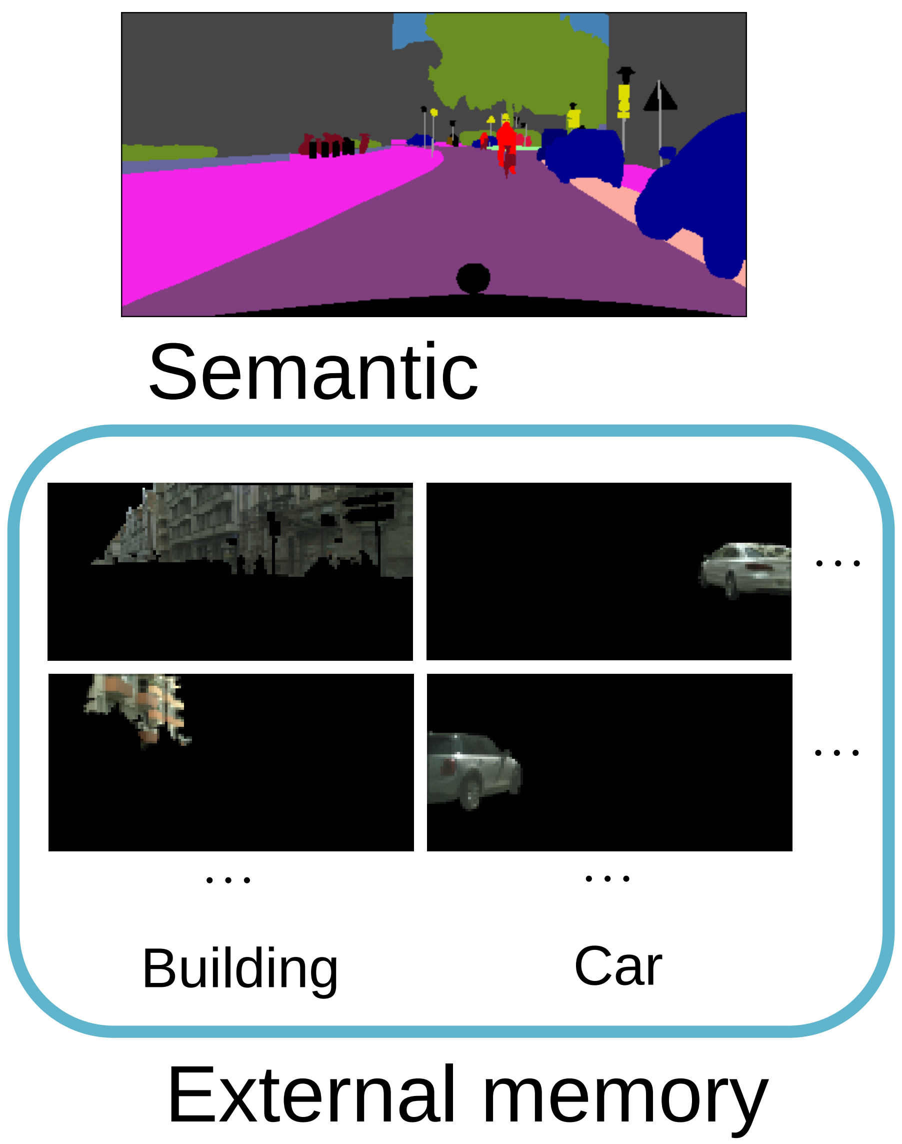


Semantic layout

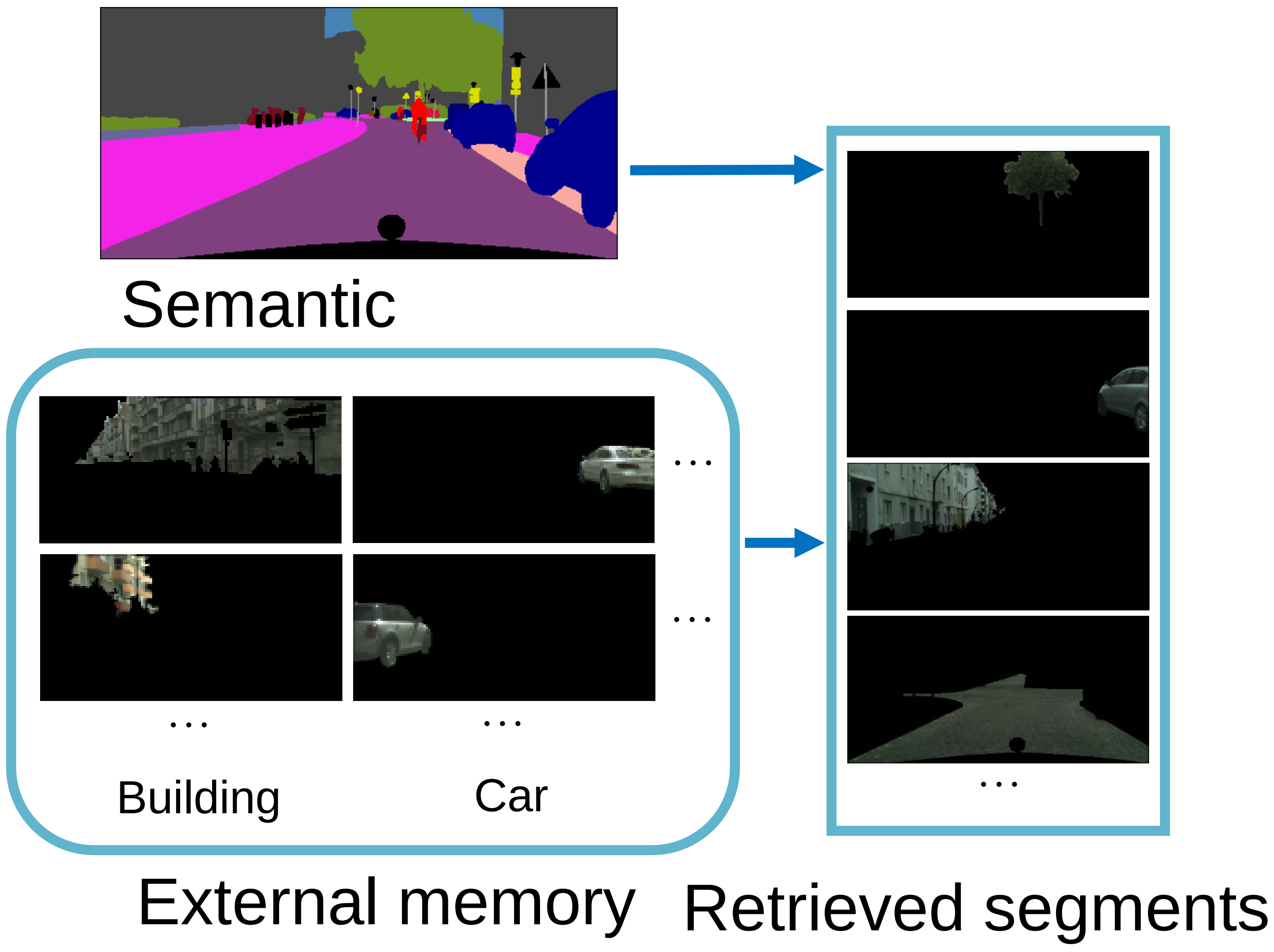
Stage 2: Image Synthesis


Canvas
Final result

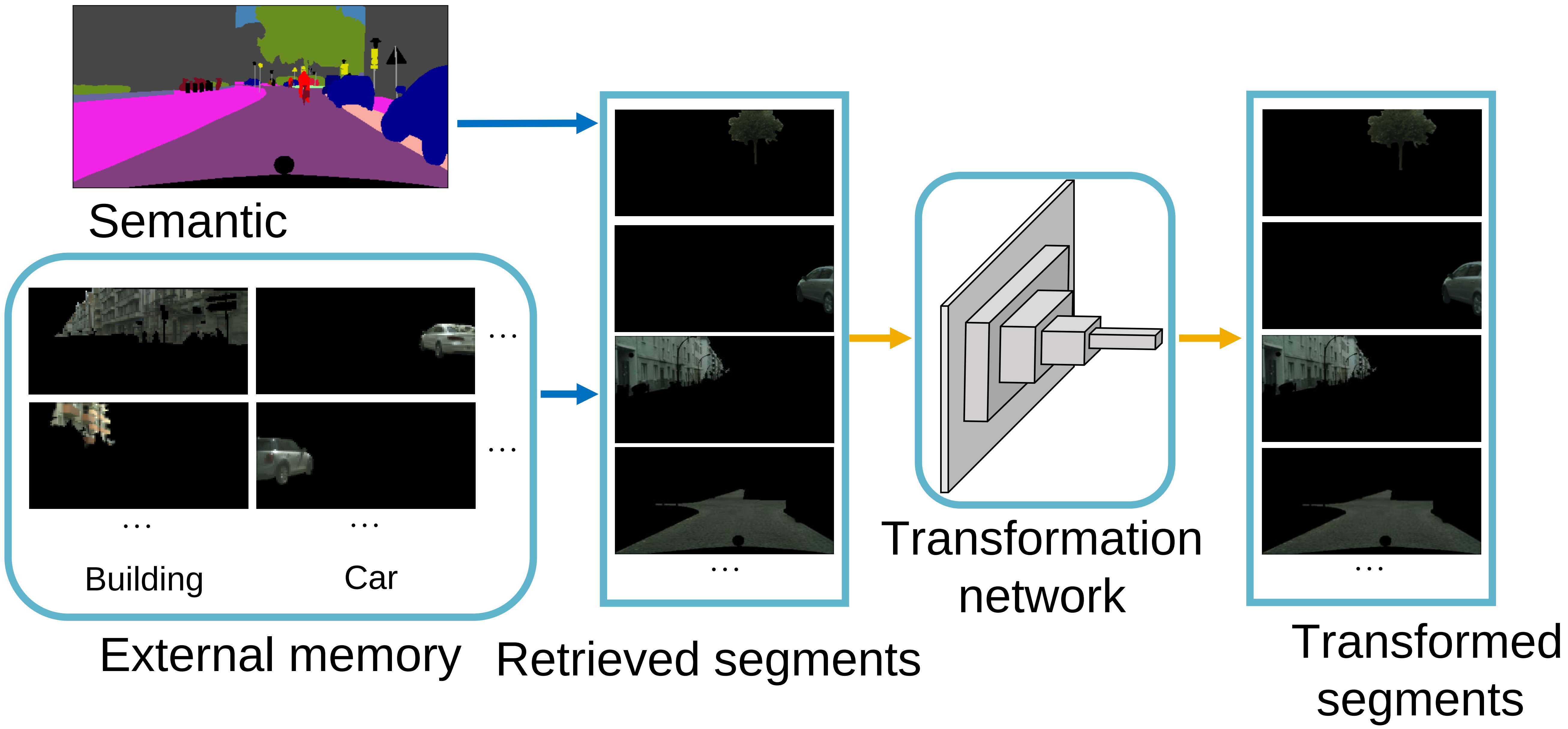
# SIMS: Canvas Generation



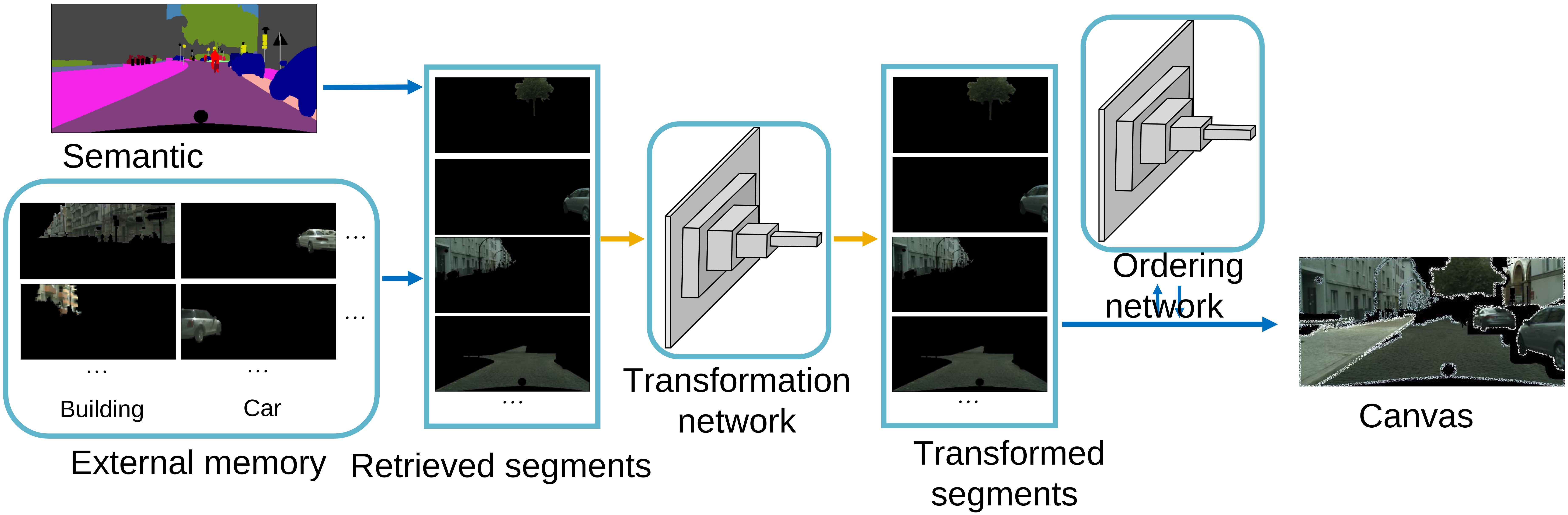
# SIMS: Canvas Generation



# SIMS: Canvas Generation

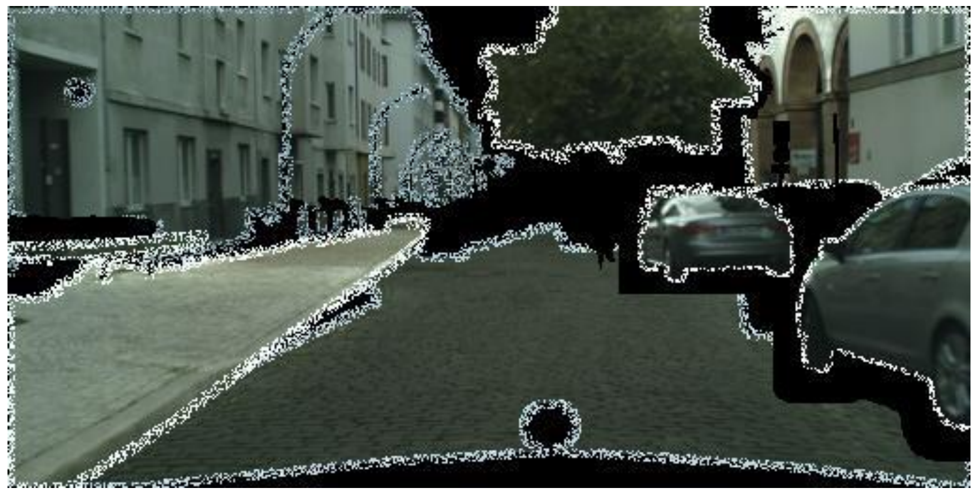


# SIMS: Canvas Generation



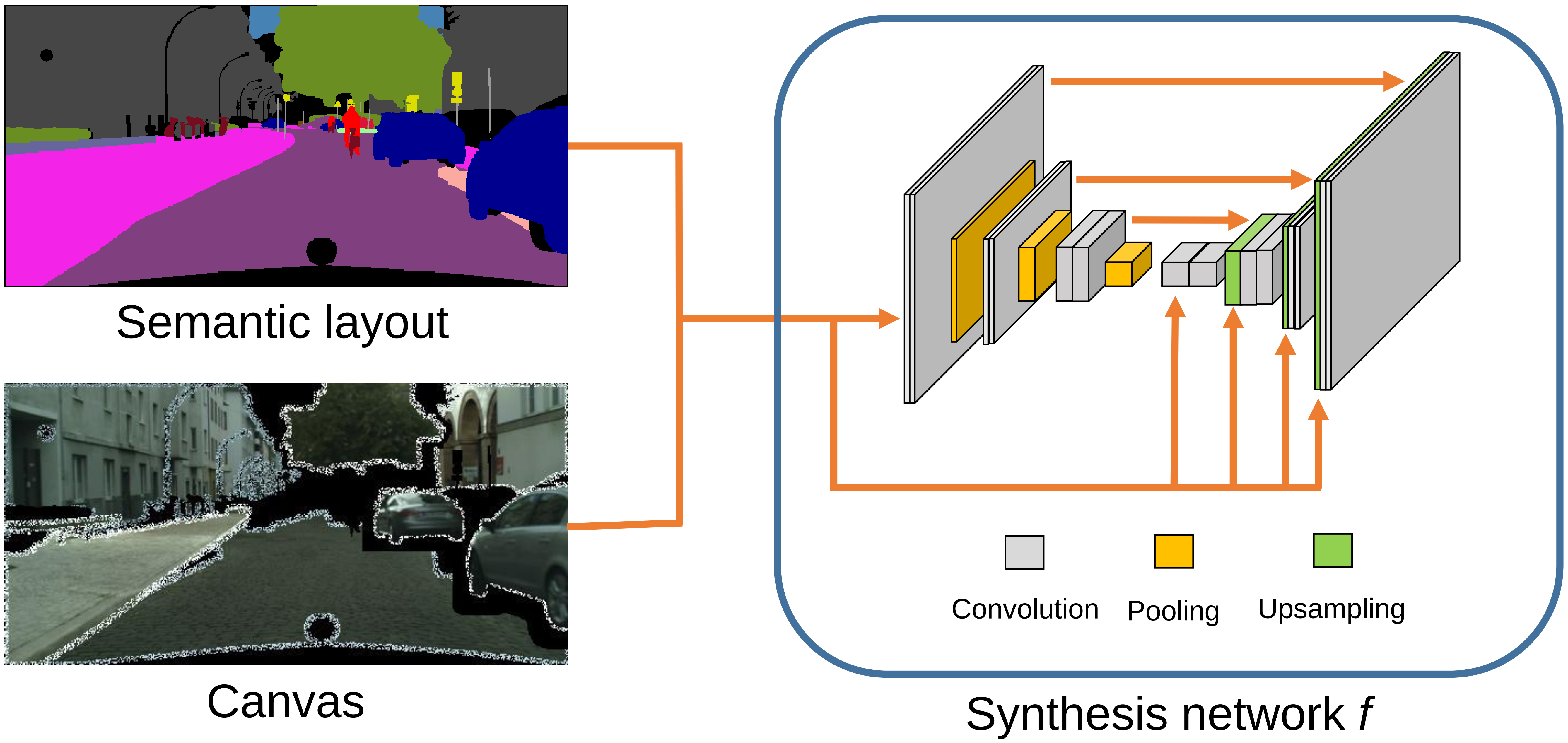
# SIMS: Image Synthesis

Semantic layout

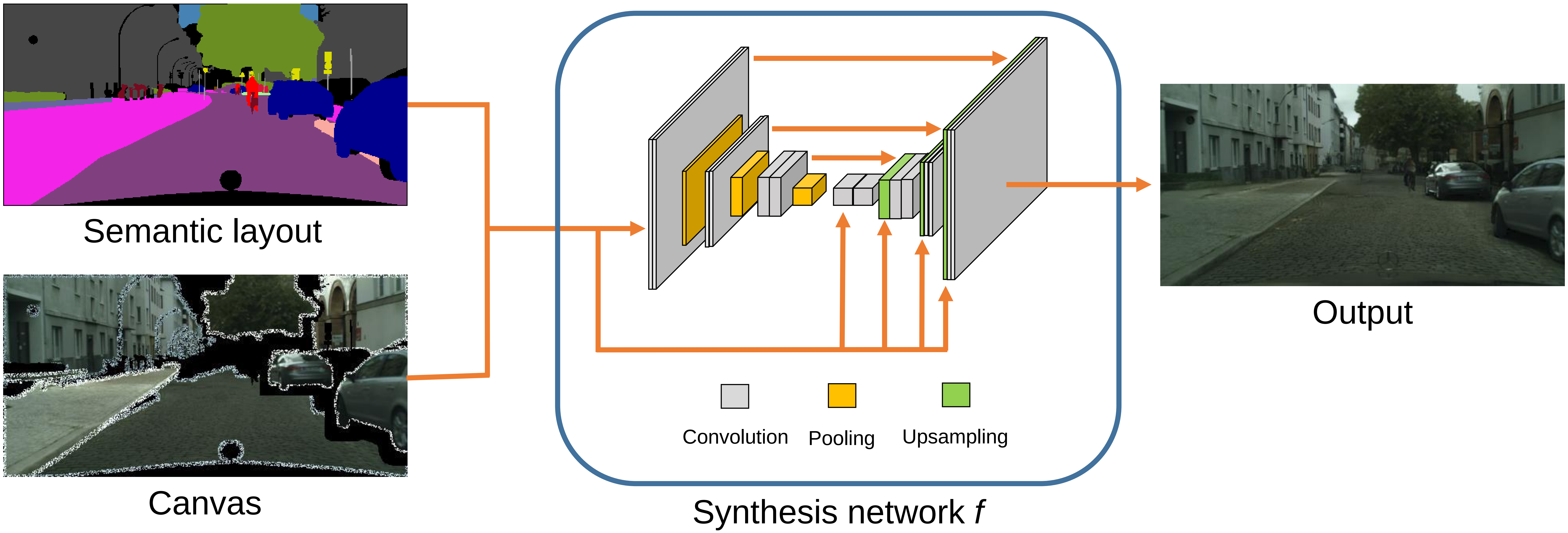


Canvas

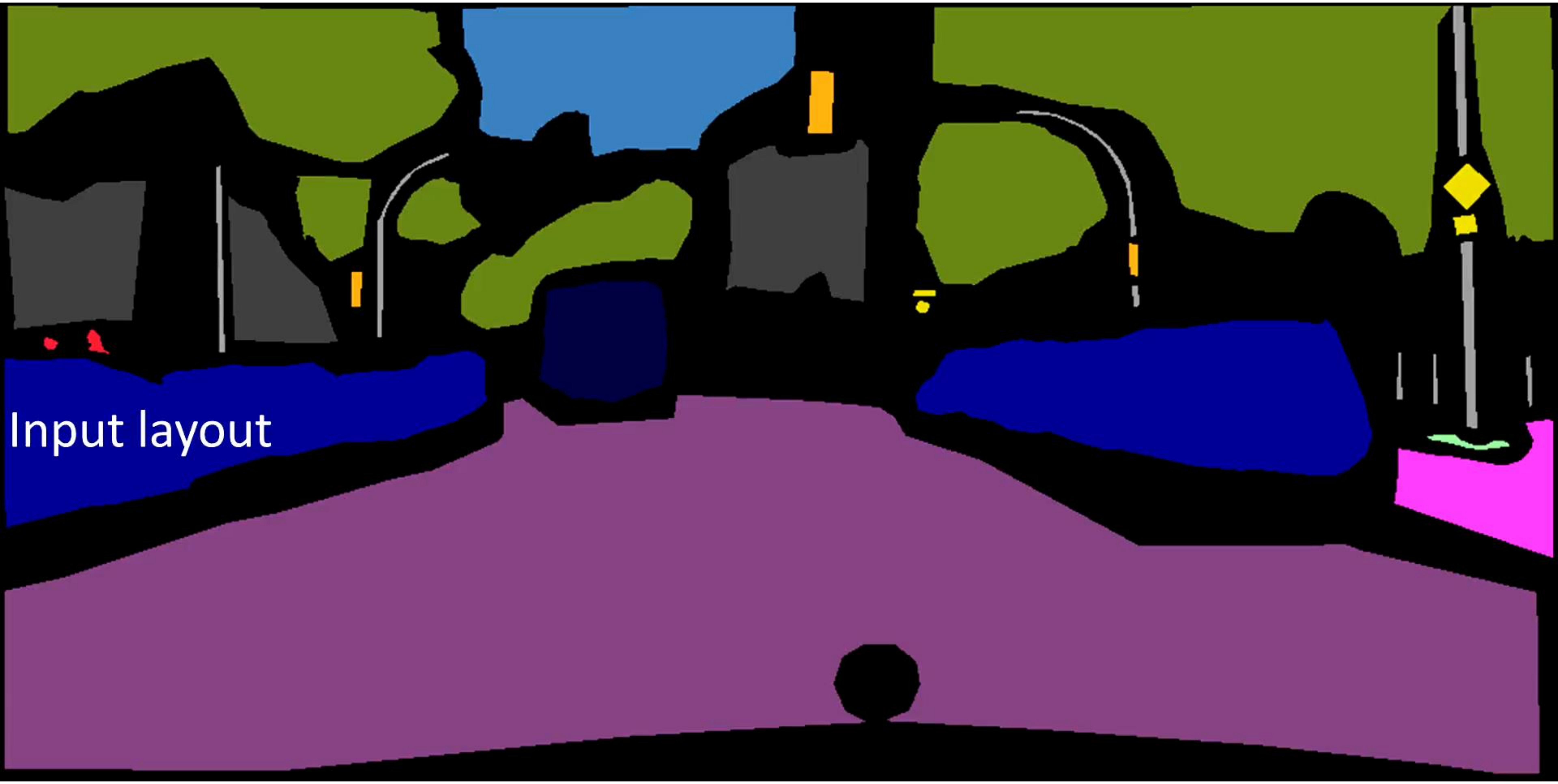
# SIMS: Image Synthesis

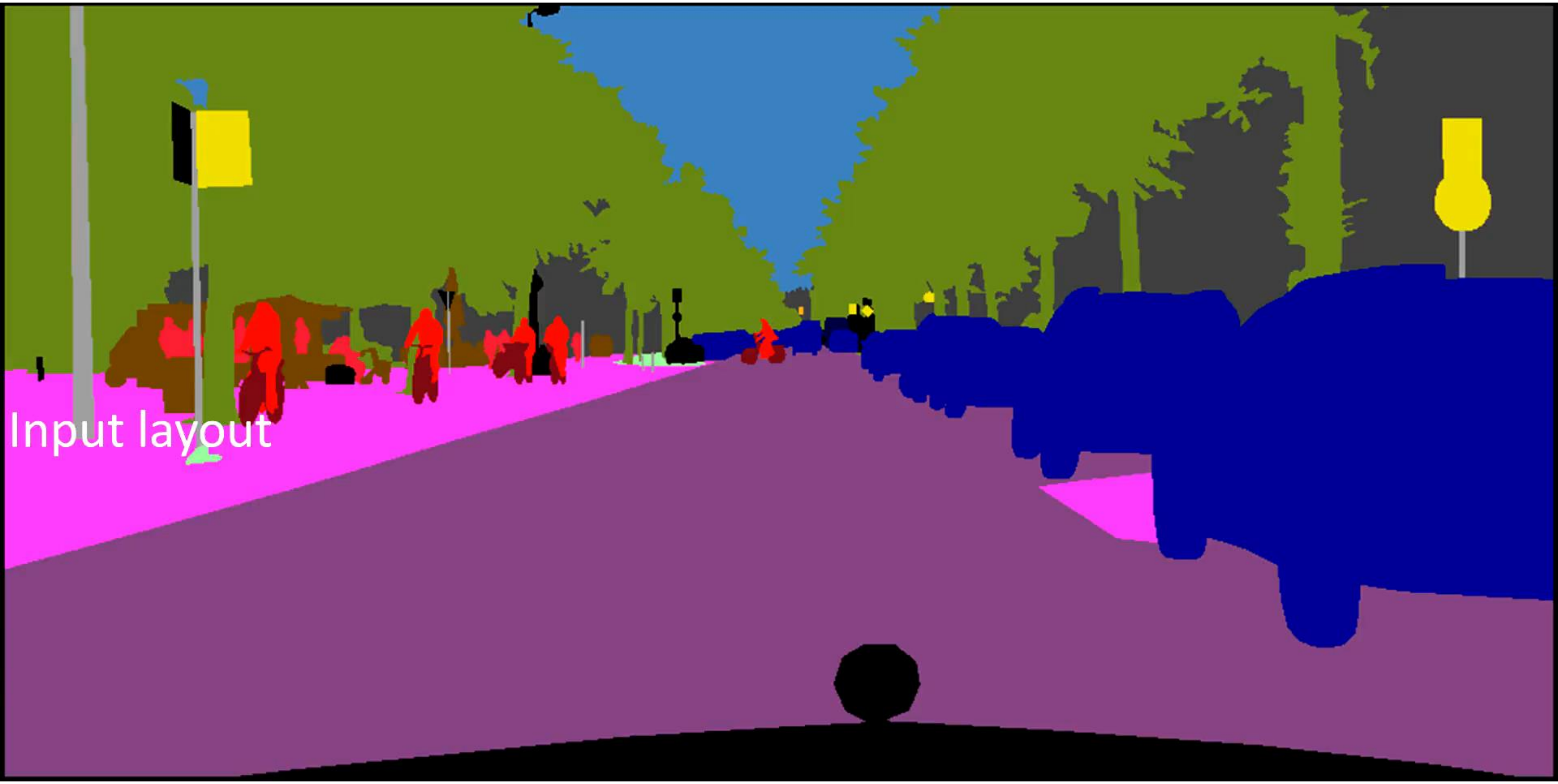


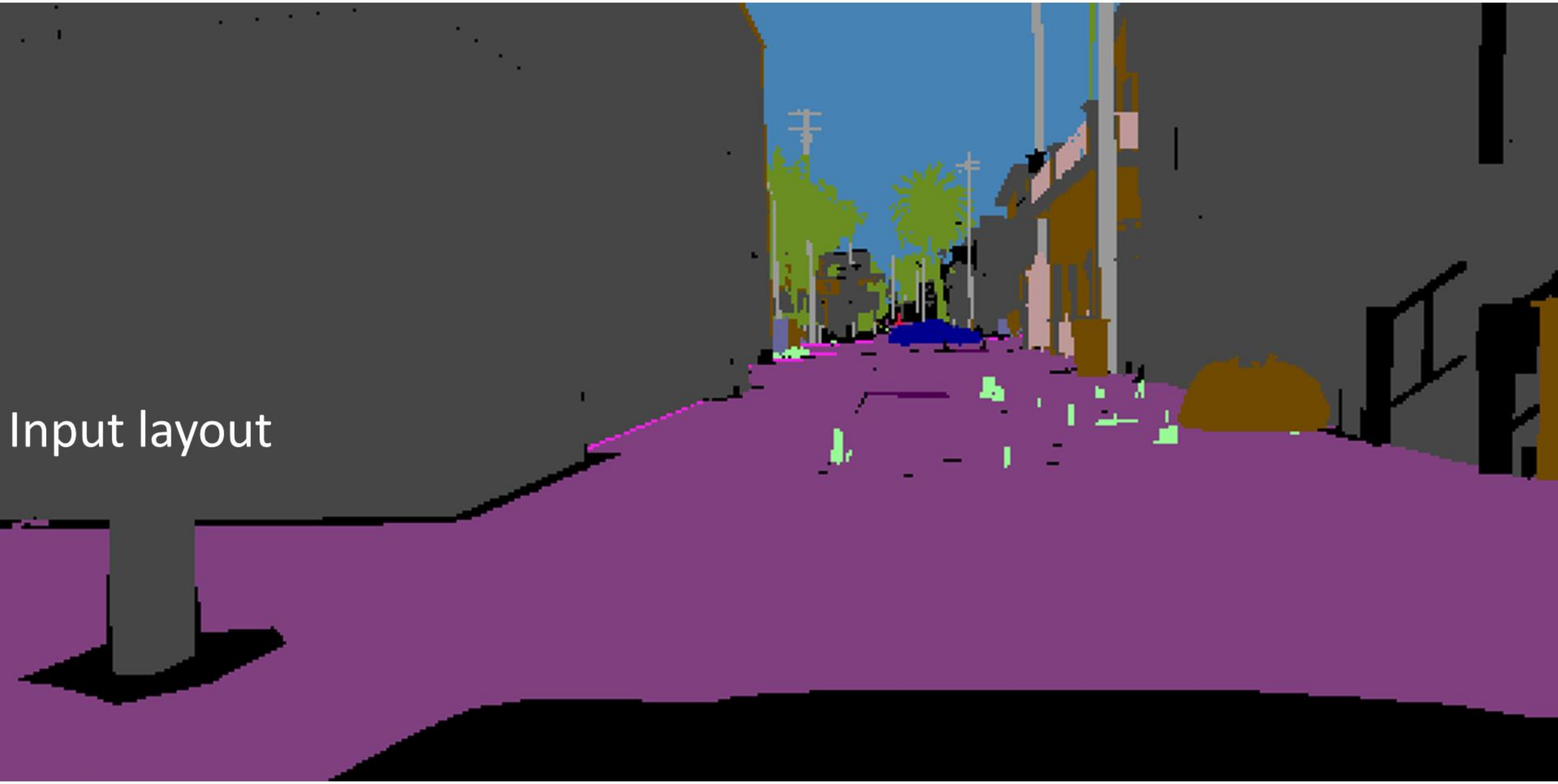
# SIMS: Image Synthesis

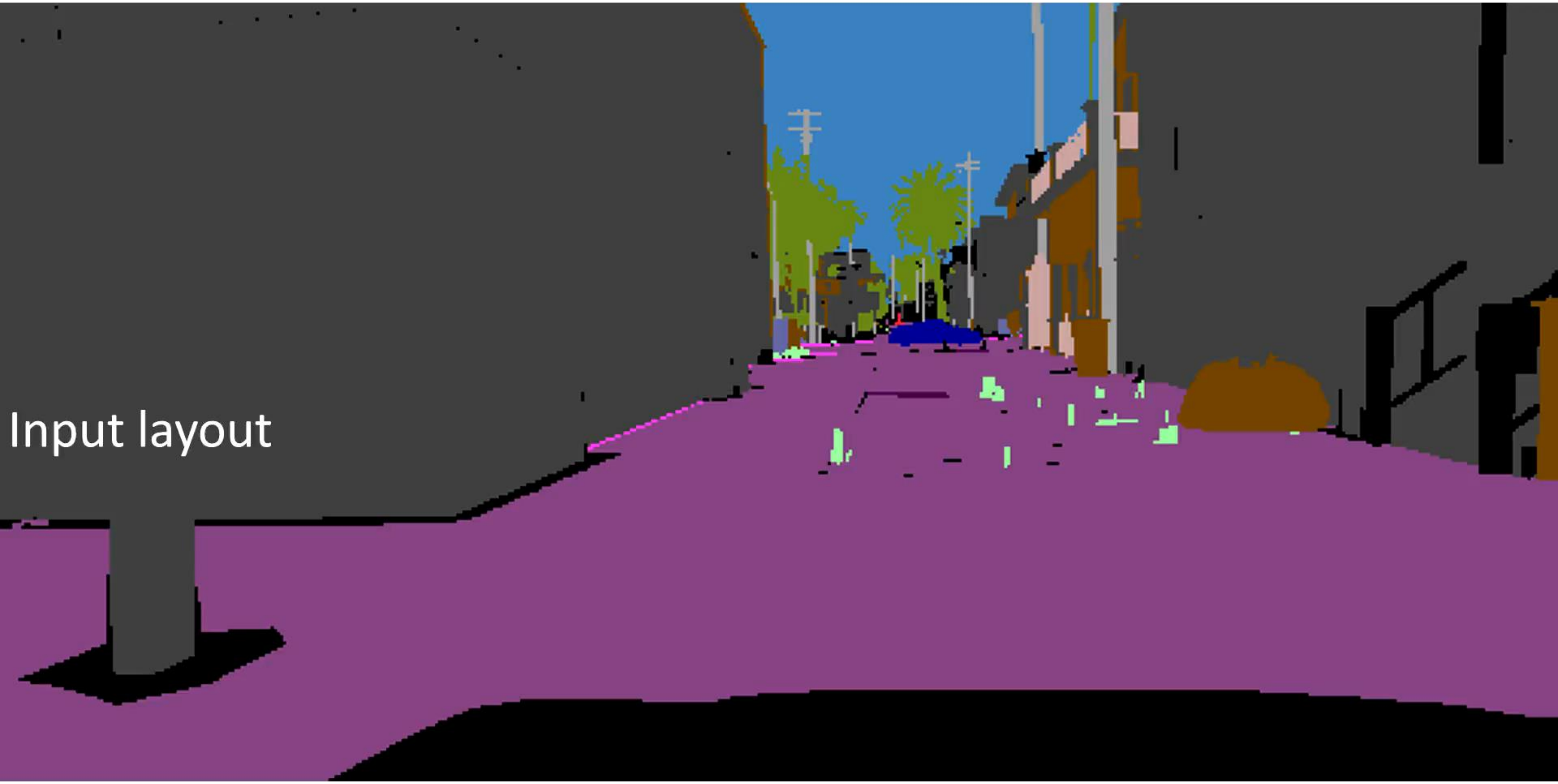


# Results


Input layout


Input layout


Input layout


Input layout

Semantic layout

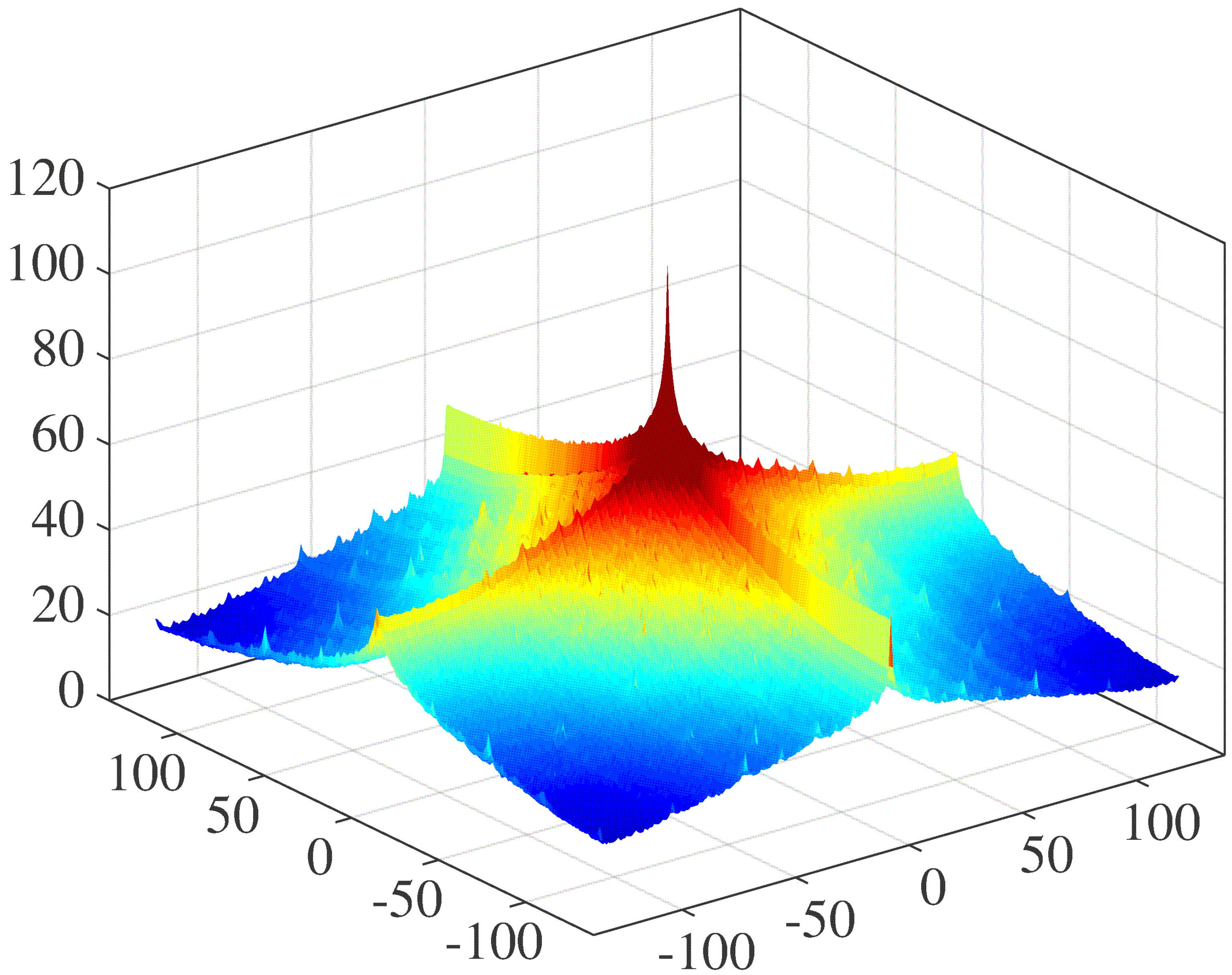
Pix2pix [Isola et al. 2017]

CRN [Chen and Koltun 2017]

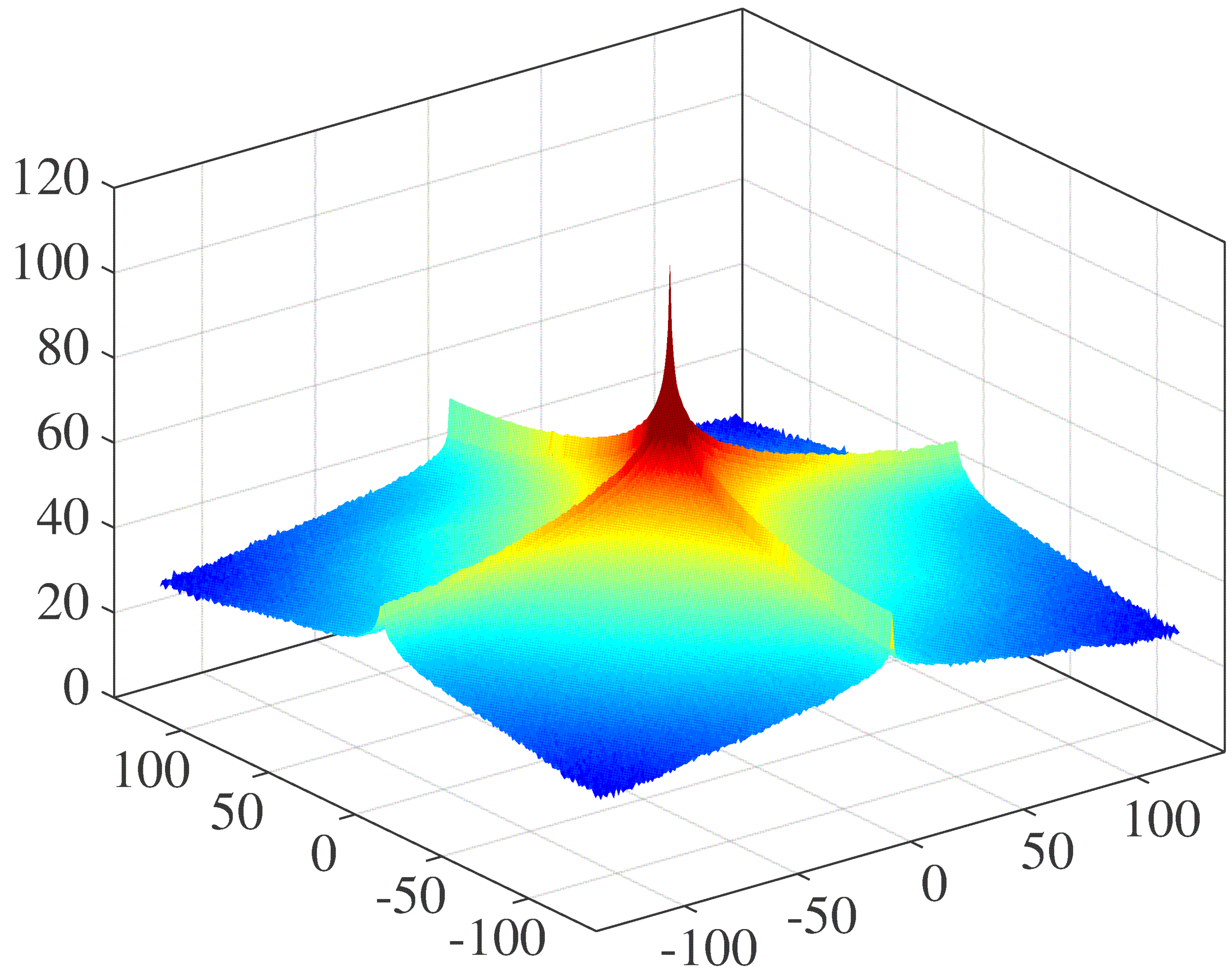
Our result

# Diversified Synthesis

# Image Statistics
# Mean Power Spectrum

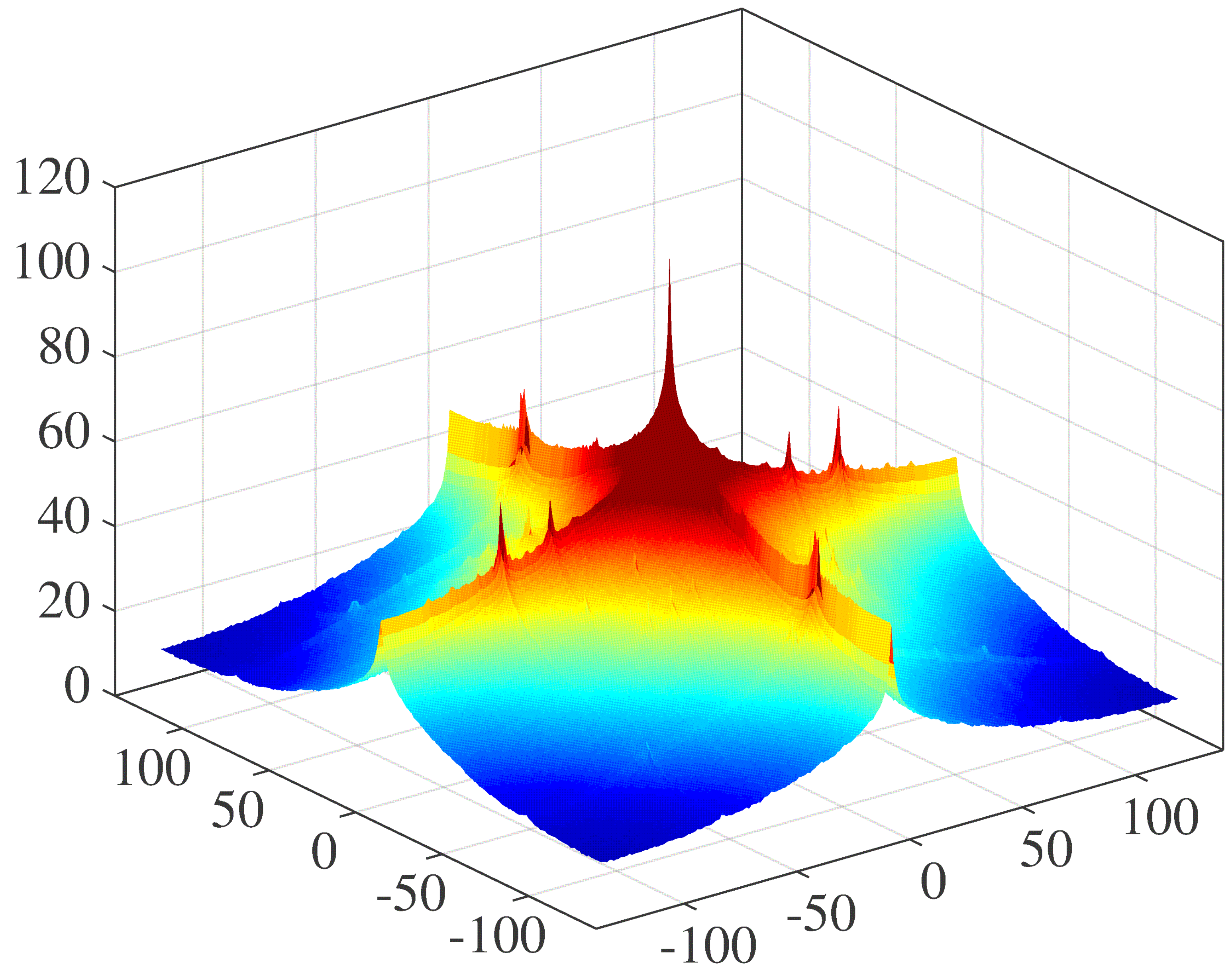



Pix2pix [Isola et al. 2017]

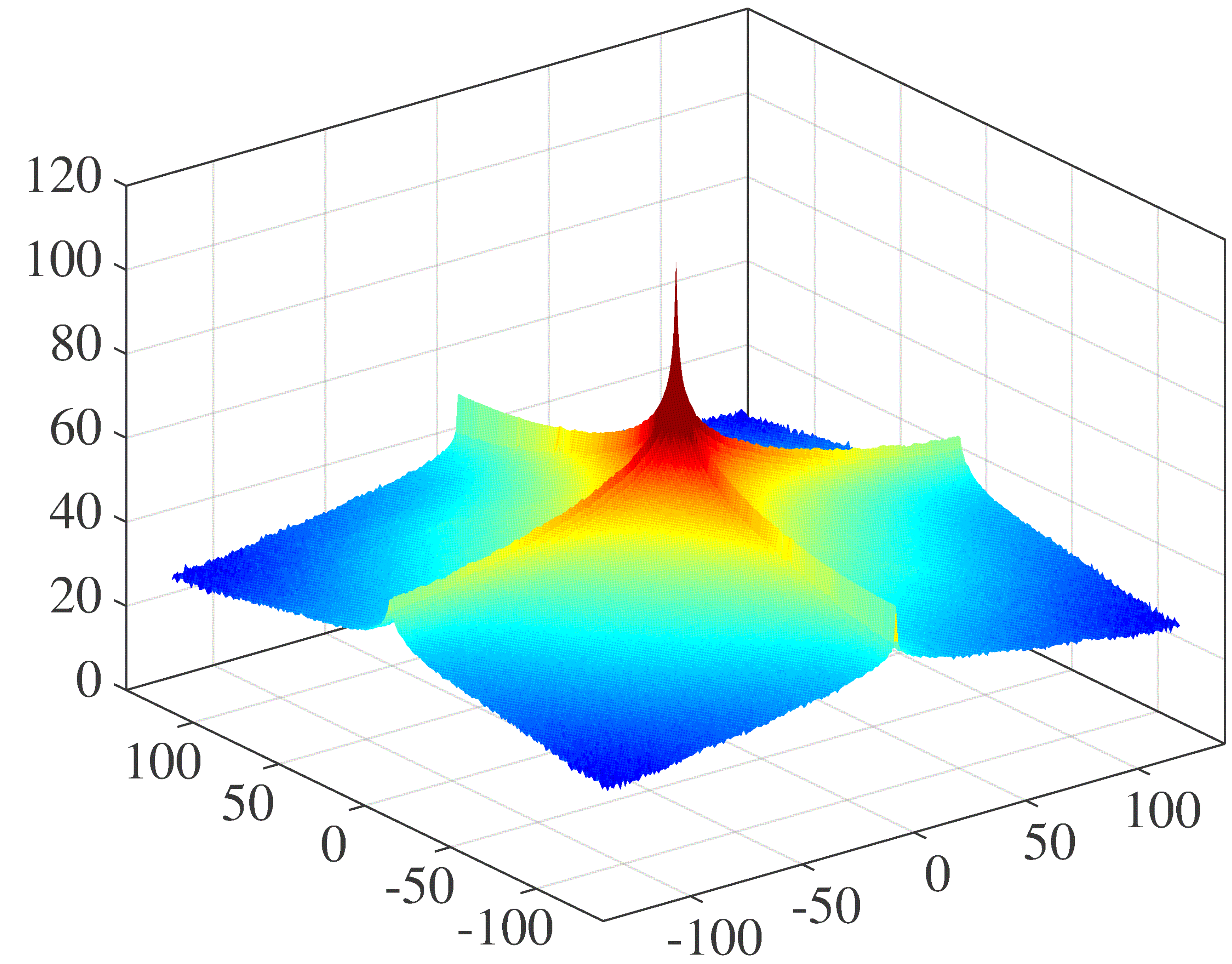



Real images

# Image Statistics

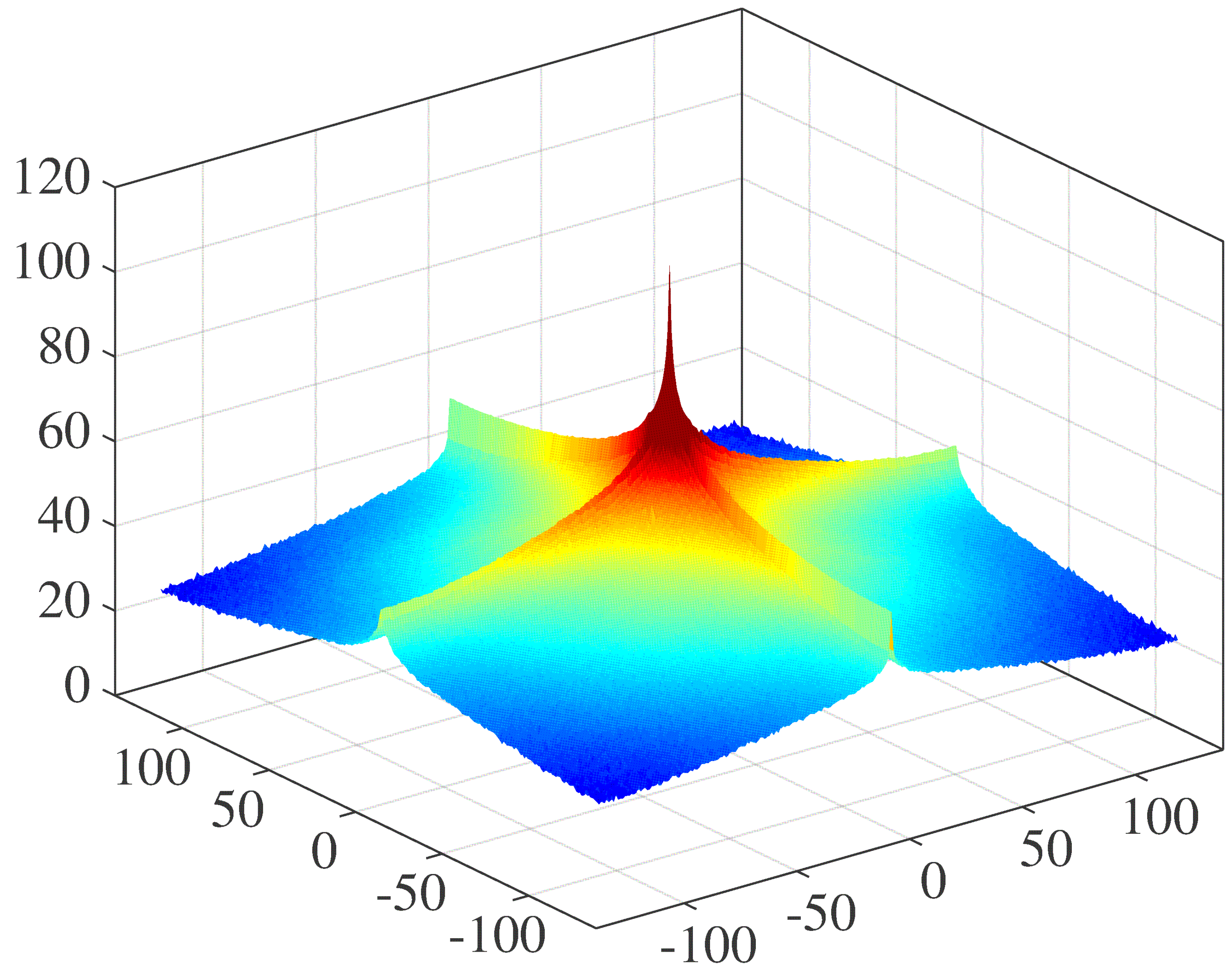



CRN [Chen and Koltun 2017]

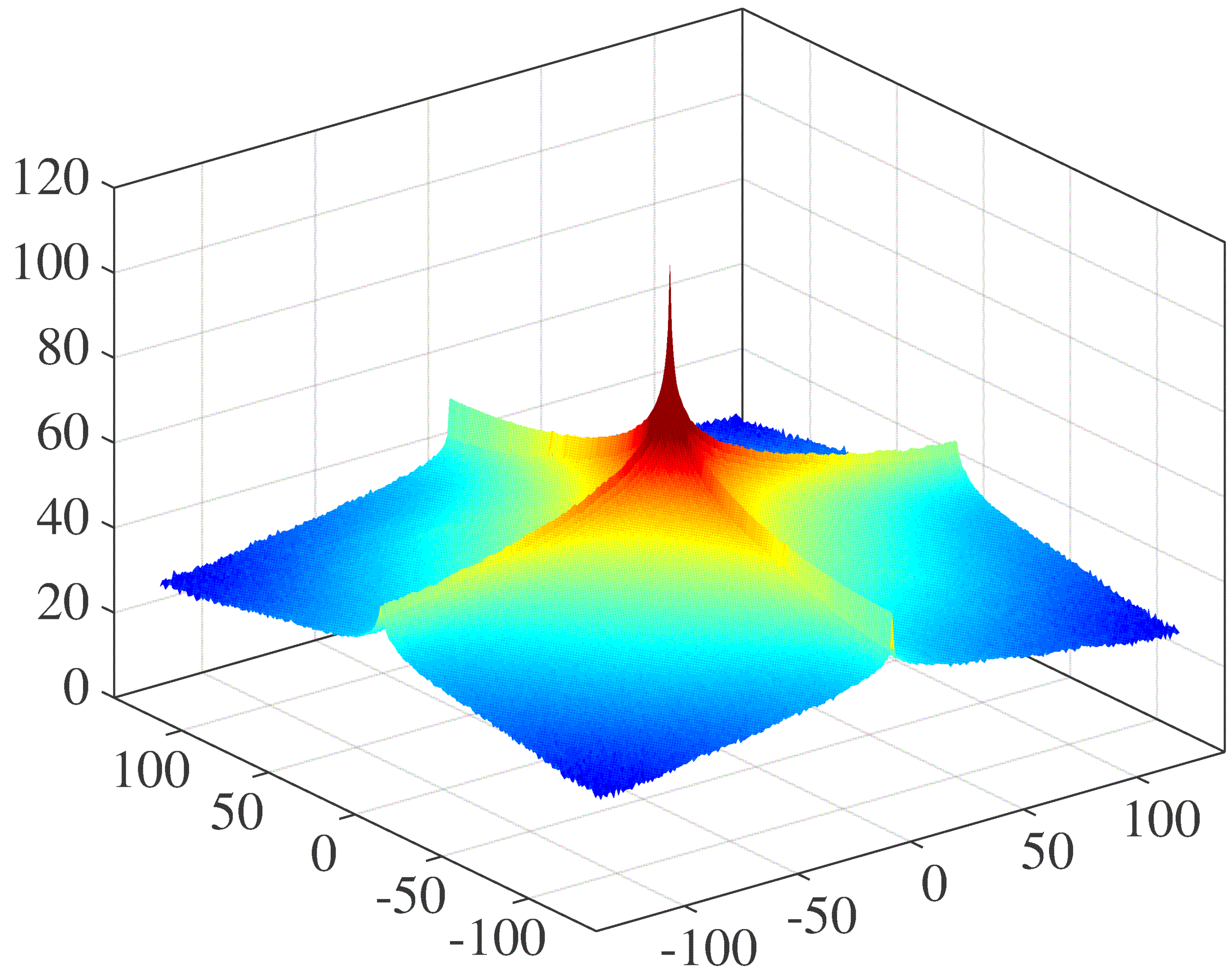



Real images

# Image Statistics




Our approach

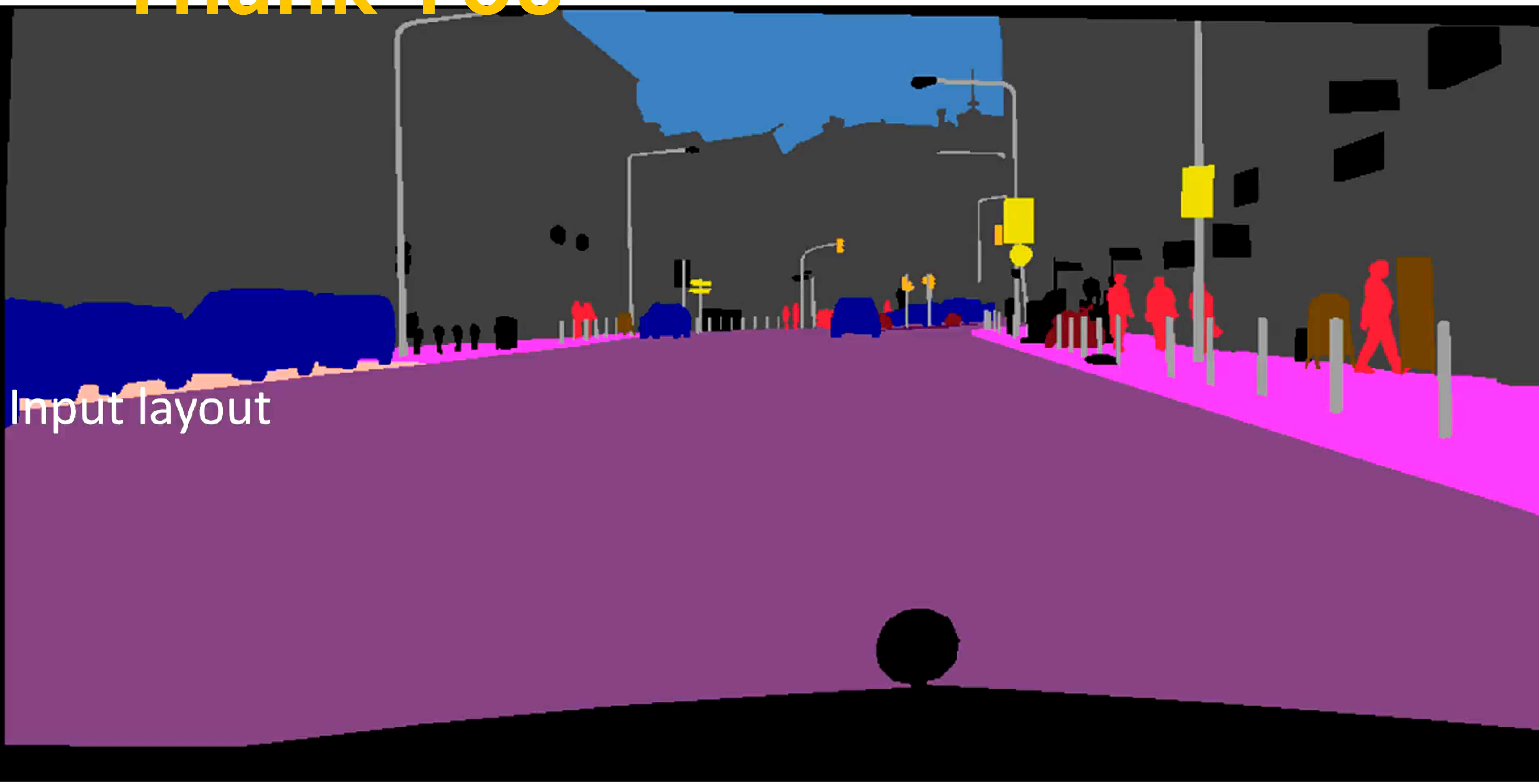



Real images

# Perceptual Experiments

| | Cityscapes (coarse） | Cityscapes (fine) | Cityscapes (GTA5) | NYU (fine) | ADE20K (coarse) | Mean |
|---|---|---|---|---|---|---|
| SIMS > Pix2pix | 94.2% | 98.1% | 95.7% | 94.9% | 87.6% | 94.1% |
| SIMS > CRN | 93.9% | 74.1% | 84.5% | 89.1% | 88.9% | 86.1% |

# Thank You

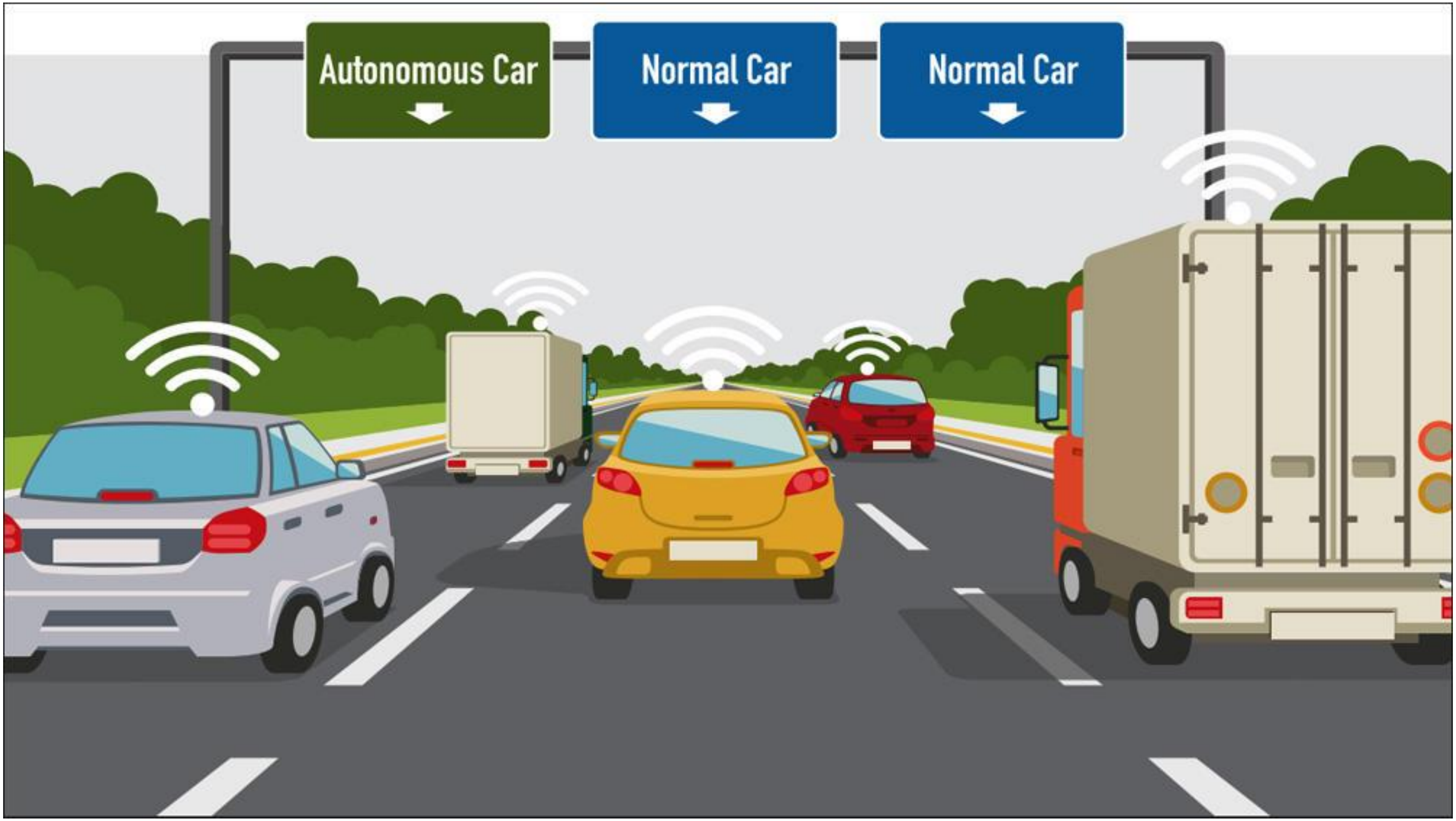


# Thank You

# Thank You

# Thank You

# Future Prediction



# Video Prediction

3D Motion Decomposition for RGBD Future Dynamic Scene Synthesis

Paper ID: 3727

# 3D Motion Decomposition for RGBD Future Dynamic Scene Synthesis

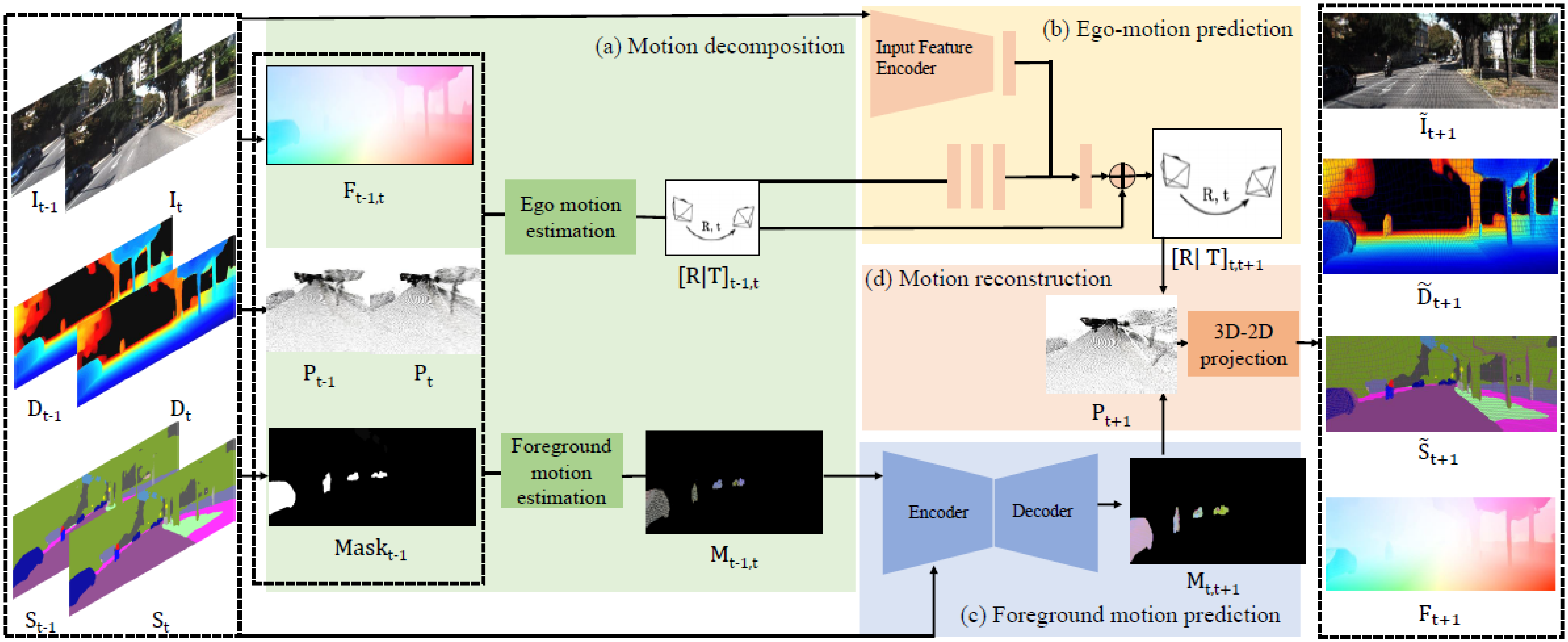



Figure 1: Motion forecasting with decomposition and composition. The input includes images ($I_{t-1}, I_t$), depth maps ($D_{t-1}, D_t$), and semantic maps ($S_{t-1}, S_t$). (a) The motion decomposition module decomposes motion into ego motion $[R|T]_{t-1,t}$ and moving object motion $M_{t-1,t}$. (b) The ego-motion prediction network and (c) the foreground motion prediction network generate future ego-motion $[R|T]_{t,t+1}$ and foreground motion $M_{t,t+1}$ respectively. (d) The motion composition module composes a predicted motion field and a new 3D point cloud $P_{t+1}$. $P_{t+1}$ is then projected to a 2D image plane. $M_{t-1,t}$ and $M_{t,t+1}$ are color coded where $R, G, B$ channels represent movement along $x, y, z$ directions.

# 3D Motion Decomposition for RGBD Future Dynamic Scene Synthesis

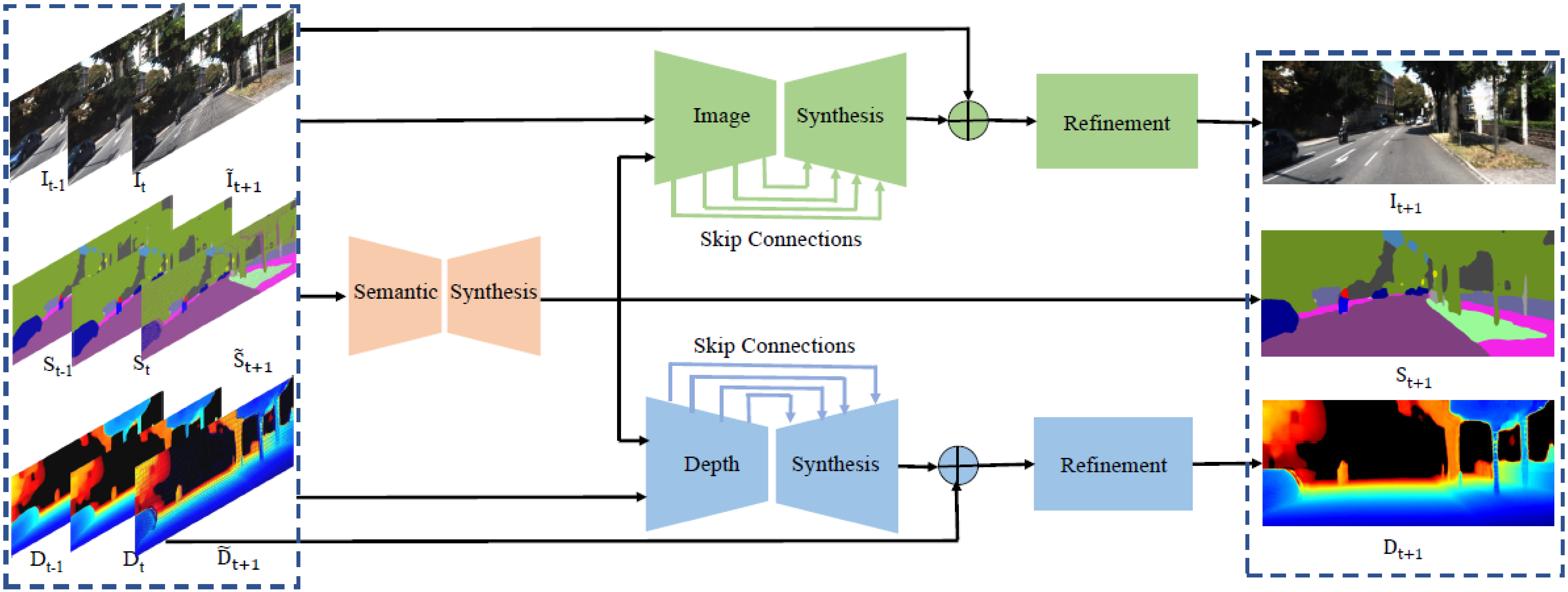



Figure 2: Refinement network. Taking as input the color images $(I_{t-1}, I_t, \tilde{I}_{t+1})$, depth maps $(D_{t-1}, D_t, \tilde{D}_{t+1})$, and semantic maps $(S_{t-1}, S_t, \tilde{S}_{t+1})$, the refinement network synthesizes image $I_{t+1}$, depth map $D_{t+1}$ and semantic map $S_{t+1}$ by refining the projected image $\tilde{I}_{t+1}$, depth $\tilde{D}_{t+1}$ and $\tilde{S}_{t+1}$.

# Results

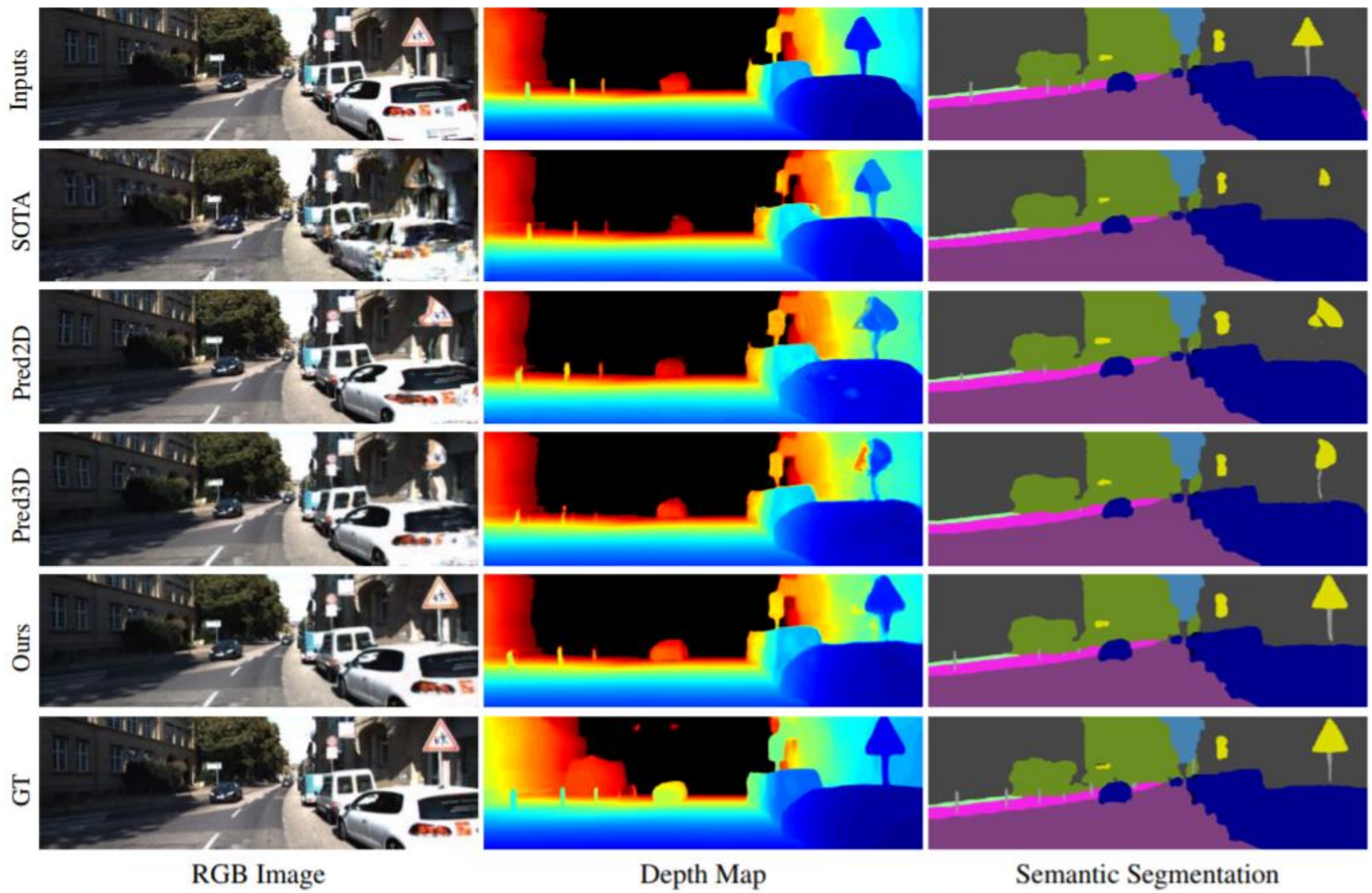



Figure 3: Visualization of different methods on next-frame prediction on the KITTI dataset. Input images are at time $t$. In the second row, the image is produced by MCNet [29] and depth map is produced by PredNet [13] while the segmentation map is from S2S [15].

# Results

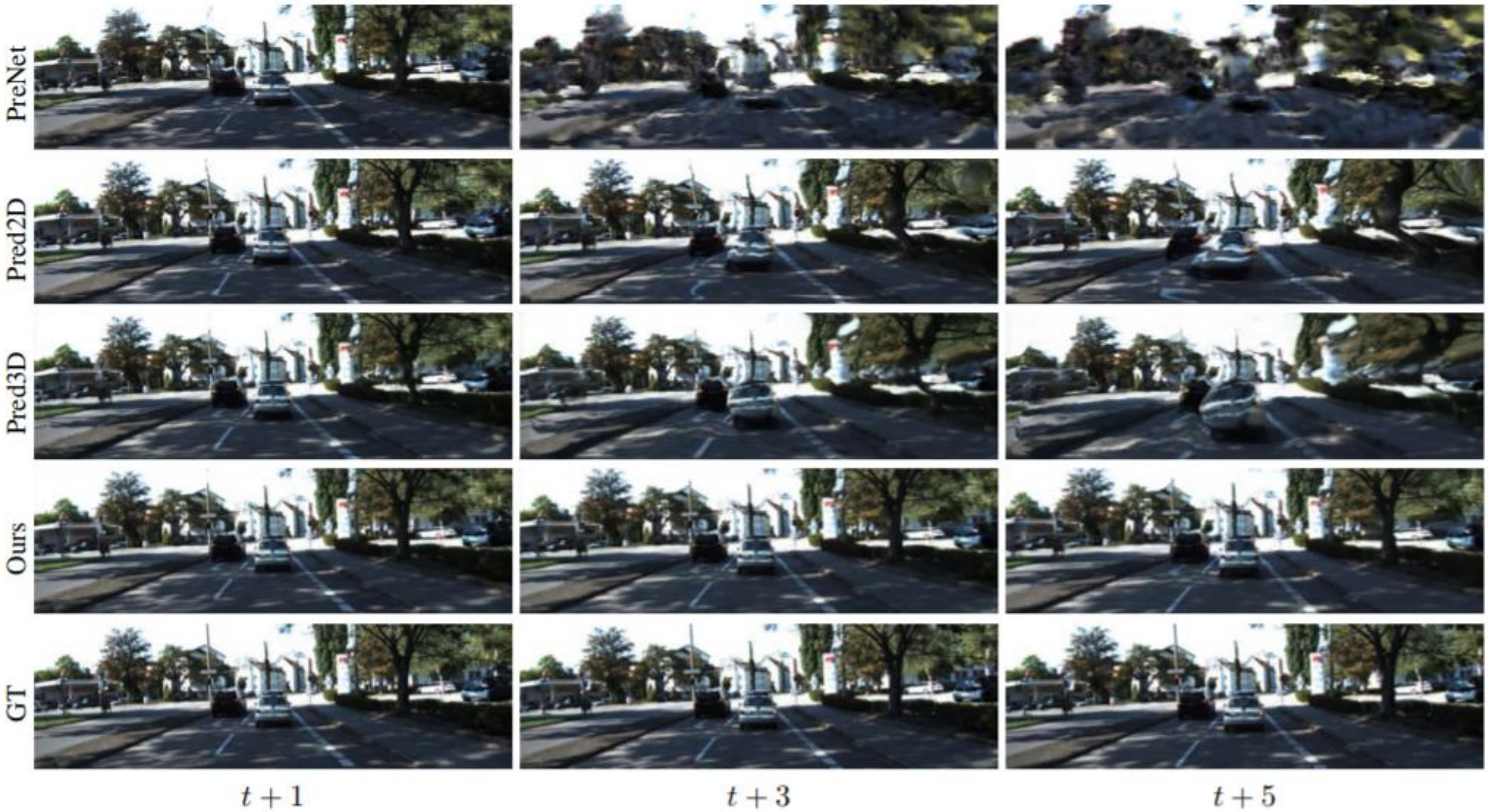



Figure 5: Results of predicting multiple frames. Depth and segmentation are provided in the supplement.

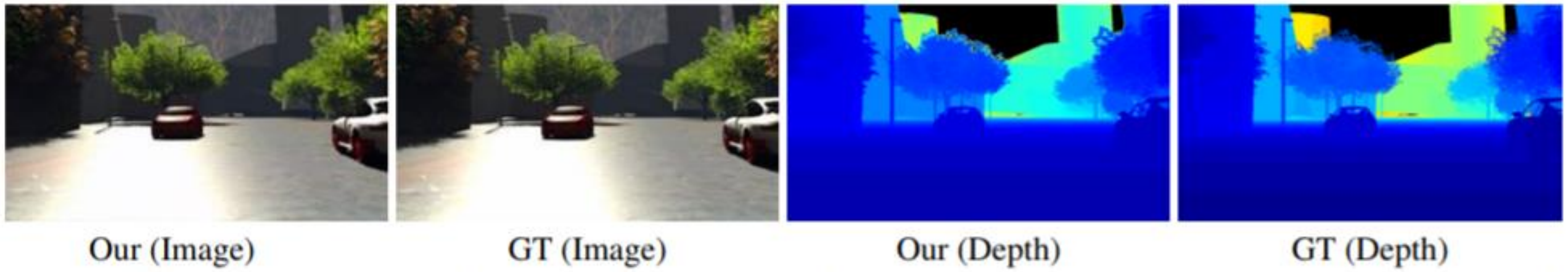



Figure 6: Visualization of our results on the Driving dataset for next frame prediction. "GT" stands for ground truth.

# Results

| | Flow | Depth | | Image | | Seg |
|---|---|---|---|---|---|---|
| | EPE ↓ | MAE ↓ | iMAE ↓ | PSNR ↑ | SSIM ↑ | IoU ↑ |
| S2S [15] | - | - | - | - | - | 37.31 |
| PredNet [13] | - | 3.71 | 5.72 | 12.37 | 0.35 | - |
| Copy | 11.88 | 3.25 | 5.38 | 12.36 | 0.36 | 31.85 |
| Warp | 11.51 | 3.32 | 5.67 | 12.48 | 0.35 | 32.67 |
| Pred2D | 8.63 | 3.92 | 7.77 | 12.41 | 0.37 | 37.33 |
| Pred3D | 10.56 | 3.09 | 5.38 | 11.99 | 0.38 | 31.87 |
| Ours | **5.57** | **2.63** | **4.17** | **13.05** | **0.41** | **41.70** |

Table 2: Qualitative results of predicting five future frames. ↑ means the higher the better. ↓ means the lower the better. "-" means invalid field.

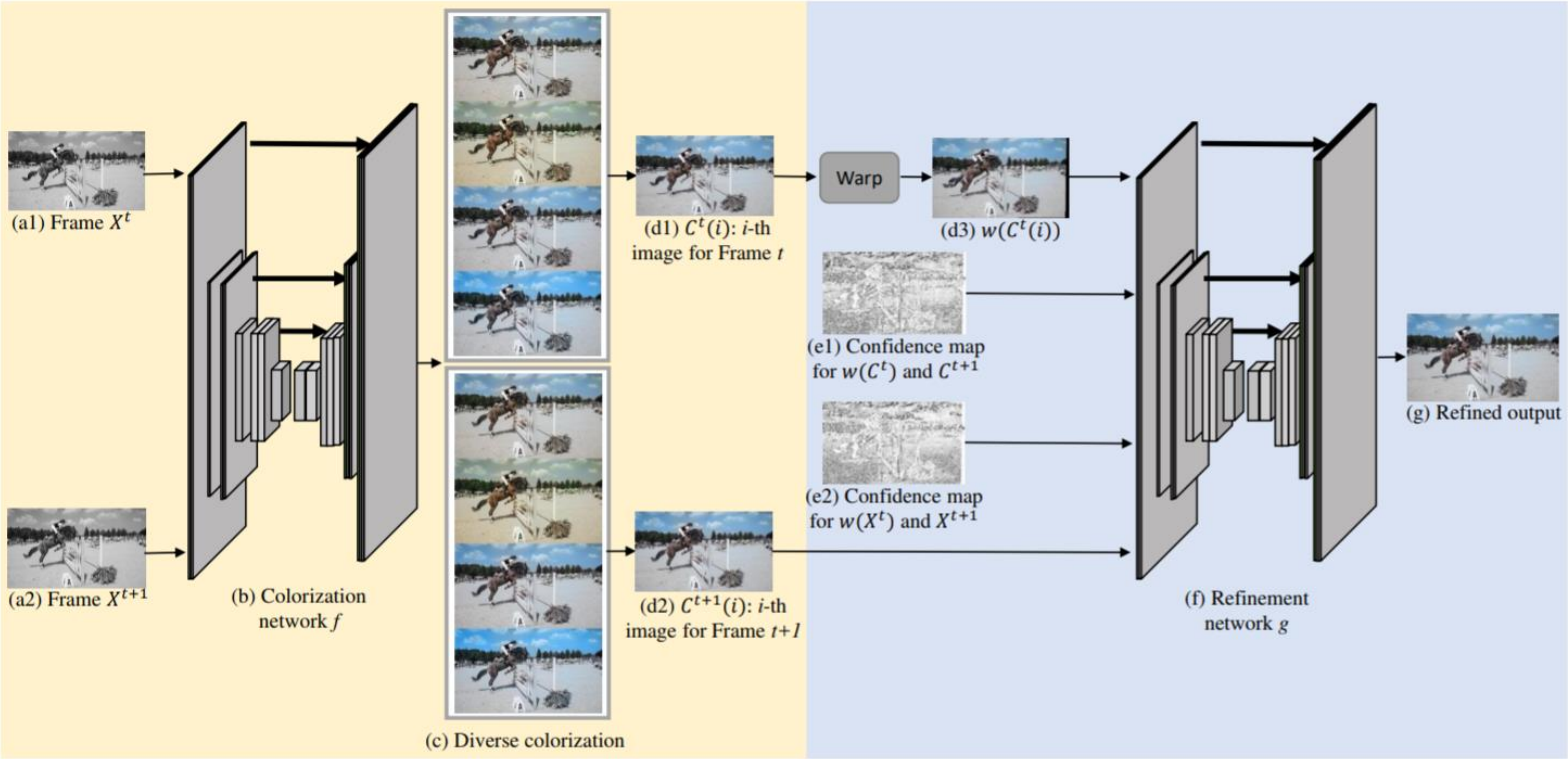
# Video Colorization

Fully Automatic Video Colorization with Self-Regularization and Diversity

Chenyang Lei　　Qifeng Chen

CVPR 2019

# Fully Automatic Video Colorization with Self Regularization and Diversity



# Diversity

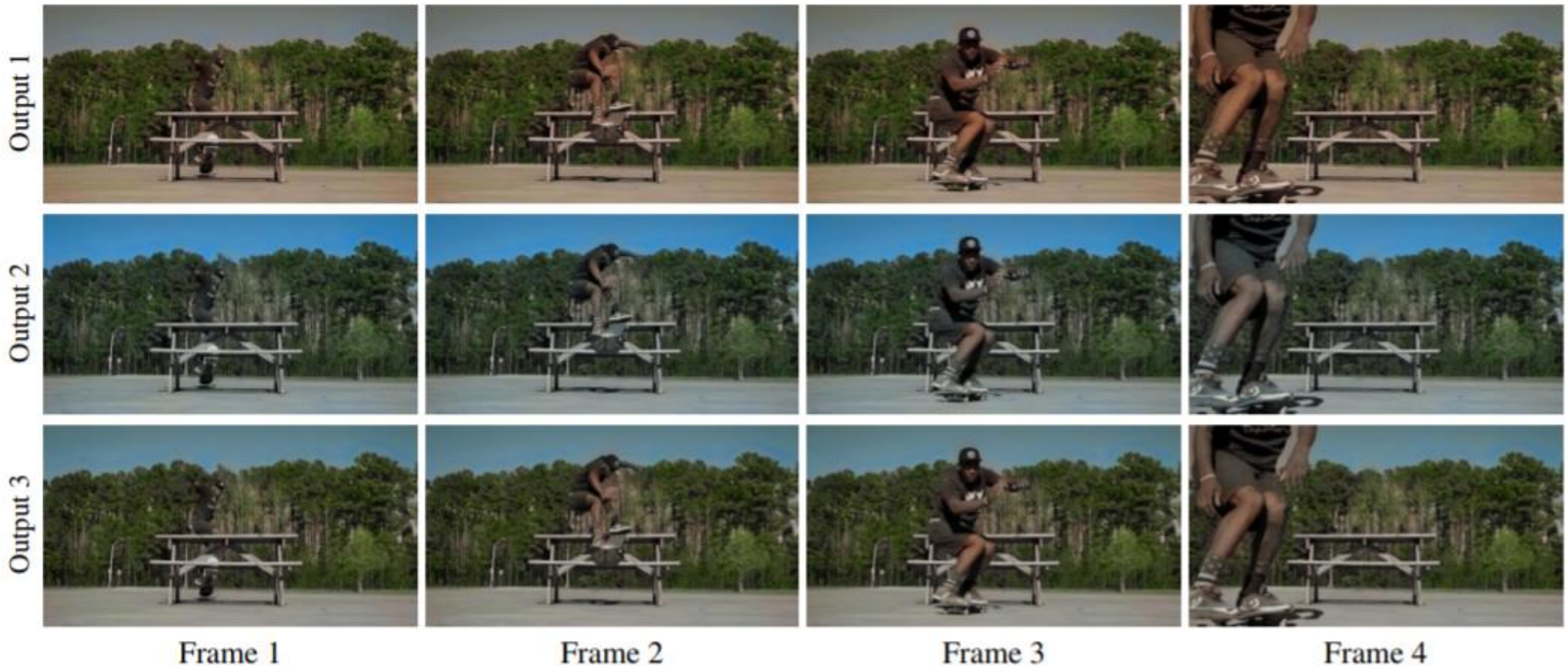



Figure 3. Four frames of three different videos colorized by our approach with diversity. Our approach is able to colorize videos in different ways. In general, different videos exhibit different global styles.

# Results

| Comparison | Preference rate DAVIS | Videvo |
|---|---|---|
| Ours > Zhang et al.[32] + BTC [15] | 80.0% | 88.8% |
| Ours > Iizuka et al. [12]+ BTC [15] | 72.8% | 63.3% |

Table 1. The results of perceptual user study. Both baselines are enhanced with temporal consistency by BTC [15]. Our model consistently outperforms both state-of-the-art colorization methods by Zhang et al. [32] and Iizuka et al. [12].

| Comparison | Preference rate DAVIS |
|---|---|
| Ours > Ours without self-reg. | 67.9% |
| Ours > Ours without diversity | 61.5% |

Table 2. The results of the ablation study of comparisons between our full model and ablated models. The evaluation is performed by perceptual user study with 15 participants. The results indicate that self-regularization and diversity are key components in our model to achieve state-of-the-art performance in fully automatic video colorization.

# Thank You

https://cqf.io