

# MeSS: City Mesh-Guided Outdoor Scene Generation with Cross-View Consistent Diffusion

Xuyang Chen<sup>1,2,\*</sup>, Zhijun Zhai<sup>3</sup>, Kaixuan Zhou<sup>1</sup>, Zengmao Wang<sup>3</sup>, Jianan He<sup>1</sup>, Dong Wang<sup>1</sup>, Yanfeng Zhang<sup>1</sup>, Mingwei Sun<sup>1,3</sup>, Xuqin Wang<sup>1,2</sup>, Rüdiger Westermann<sup>2</sup>, Tao Wu<sup>1</sup>, Konrad Schindler<sup>4</sup>, Liqiu Meng<sup>2</sup>



Project page



ETH zürich

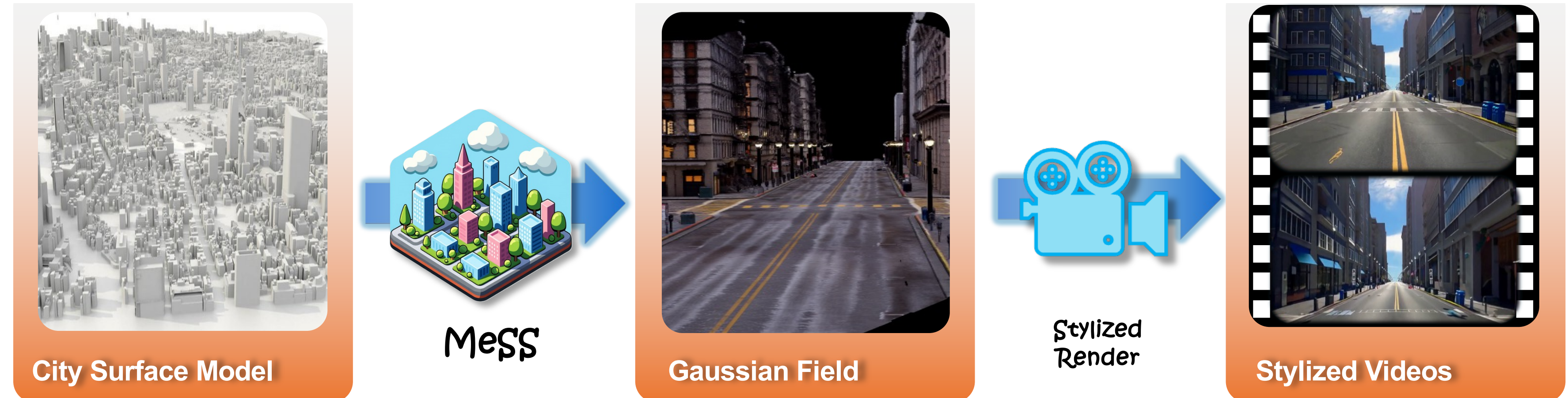
<sup>1</sup>Huawei Hilbert Research Center <sup>2</sup>Technical University of Munich, Munich, Germany  
<sup>3</sup>Wuhan University, Wuhan, China <sup>4</sup>ETH Zurich, Zurich, Switzerland

## 01 Introduction

City meshes have geometry, but lack realistic appearance.

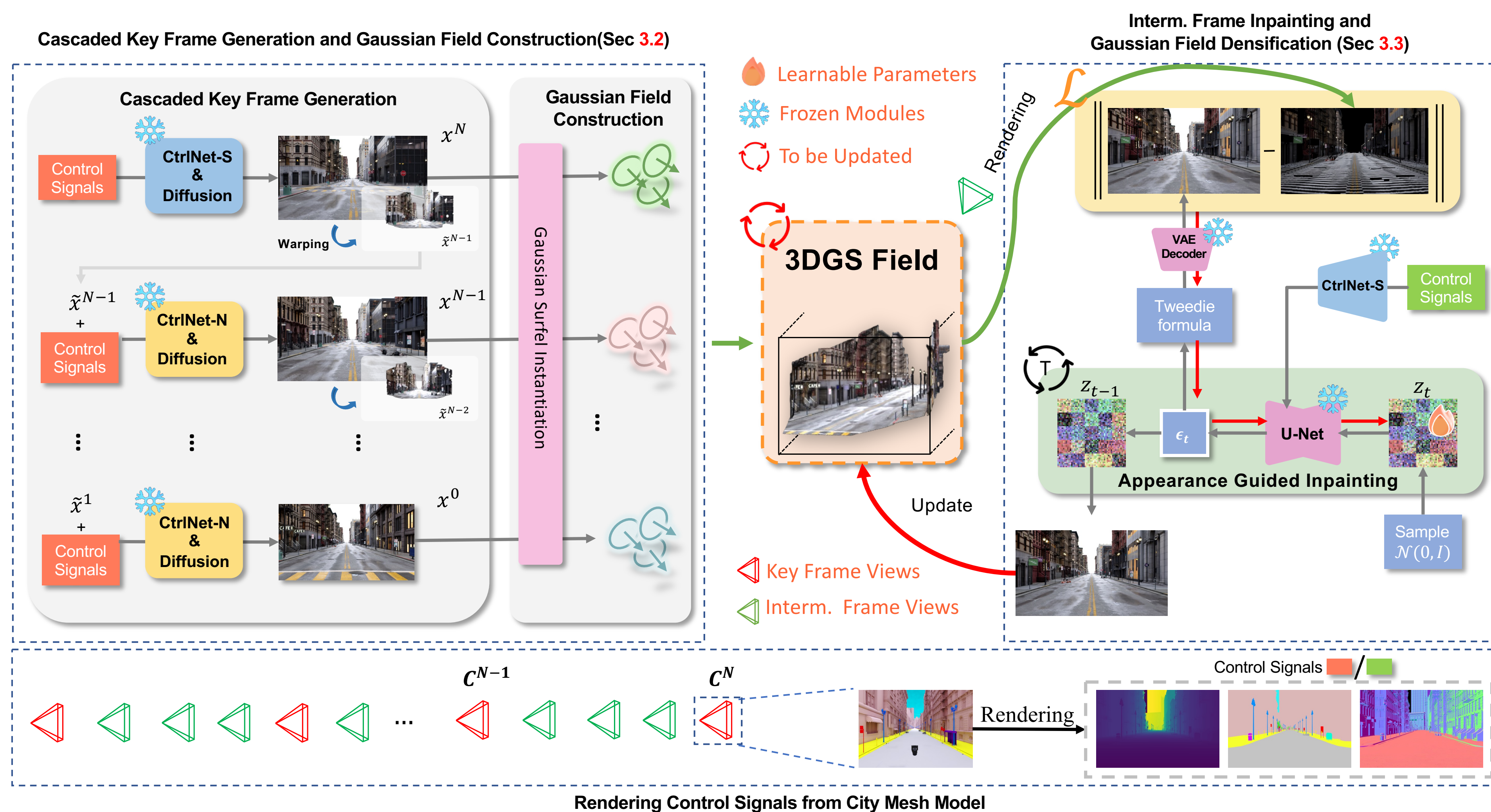
Generating street-level scenes must preserve the given geometry and avoid appearance drift along long camera paths.

**MeSS turns an untextured city mesh into a coherent 3D Gaussian Splatting scene.**



Geometry-aligned synthesis supports novel-view rendering and subsequent stylization.

## 02 Methods



### Match the control distribution

Train ControlNets on mesh-rendered depth, normals and semantics; use the same type of controls at inference.

#### 1 Generate sparse key views

Cascaded CtrlNets warp and outpaint backwards along the path. Initialize Gaussian surfels directly on the mesh.

#### 2 Fill intermediate views

AGInpaint guides diffusion with known rendered pixels to repair holes and thin silhouette regions.

#### 3 Align global appearance

GCAAlign harmonizes exposure across views through diffusion and 3DGS rendering.

Sparse-to-dense reconstruction uses metric mesh depth and surface normals, avoiding monocular depth estimation.

## 03 Experiments & Results

**31% lower LPIPS**

0.348 vs. VistaDream<sup>†</sup>: 0.508

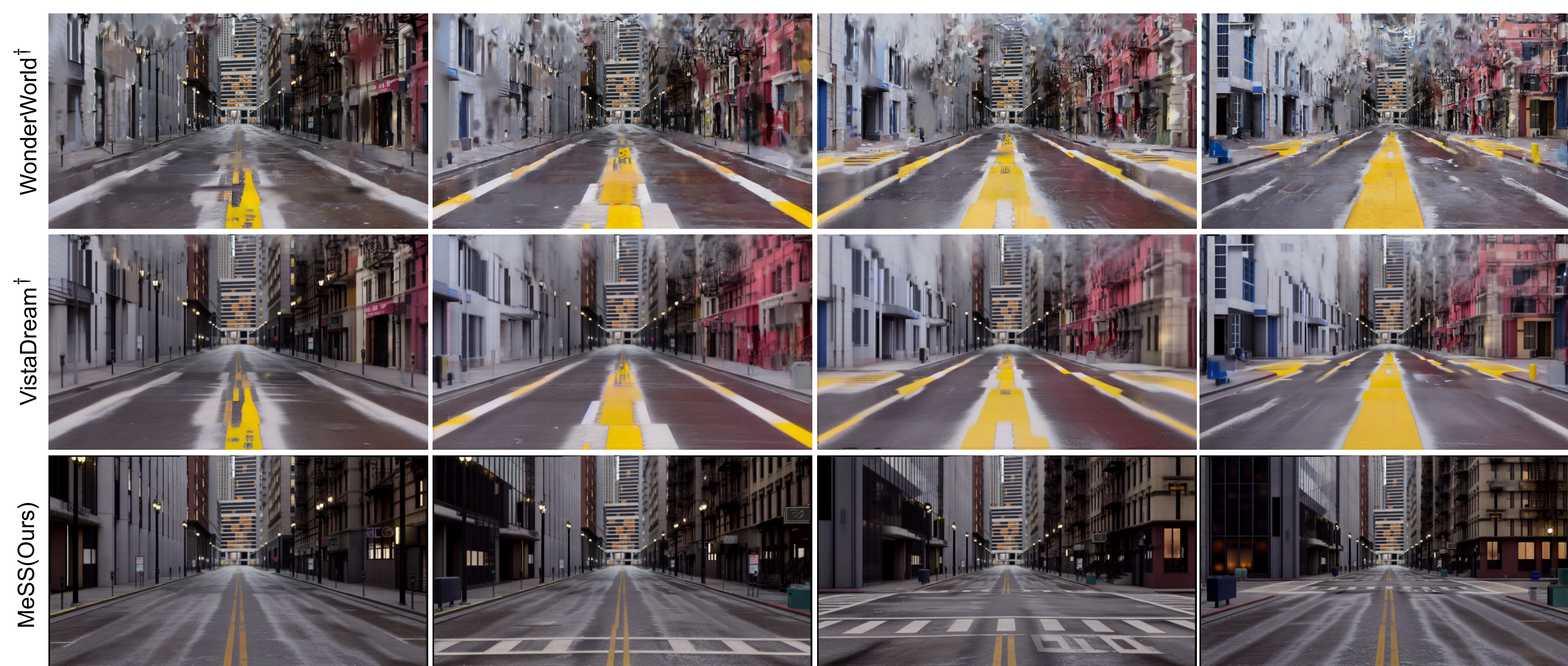
**36% lower KID**

0.016 vs. StreetScapes: 0.025

**FID 28.17**

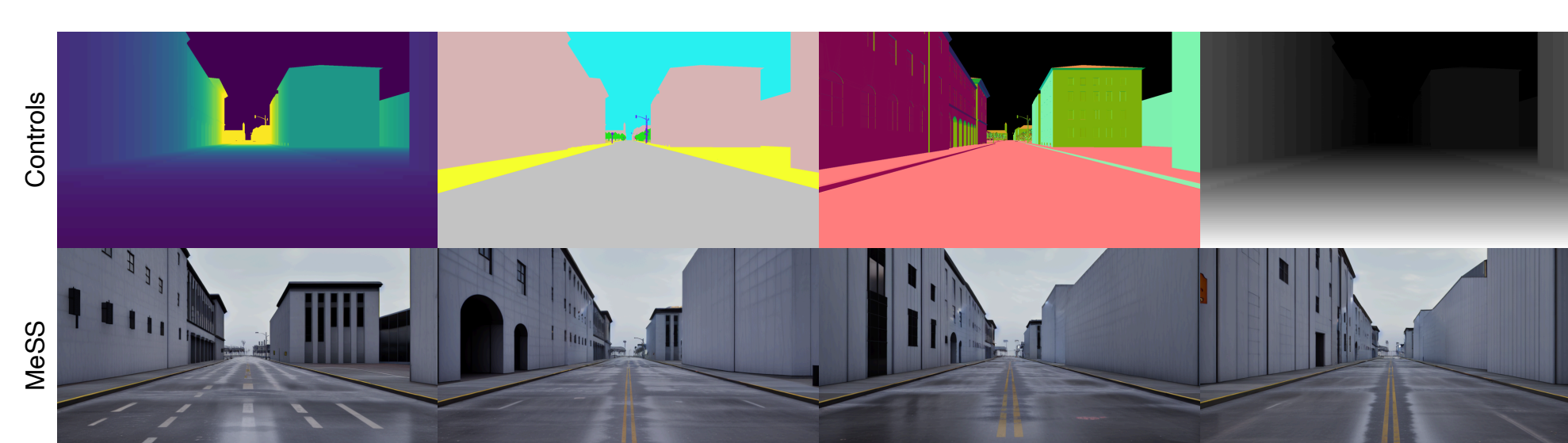
Lowest among the reported methods

Long trajectories: 200 frames / 200 m



Matched baselines use the same initial image, metric mesh depth and aligned inpainting model. MeSS reduces accumulated appearance drift.

**Zero-shot transfer to real LoD3 meshes**



TUM2Twin, Munich: no retraining. Facades follow the mesh geometry.

**Limitation:** real scenes retain a UE5-style texture prior.

### UE5 City Sample benchmark

16k training images; 3,200 evaluation images. All metrics below: lower is better.

| Method                   | LPIPS        | FID          | KID          |
|--------------------------|--------------|--------------|--------------|
| CityDreamer4D            | —            | 88.48        | 0.049        |
| WonderWorld <sup>†</sup> | 0.516        | 75.81        | 0.076        |
| VistaDream <sup>†</sup>  | 0.508        | 72.44        | 0.073        |
| StreetScapes             | 0.519        | 29.93        | 0.025        |
| <b>MeSS</b>              | <b>0.348</b> | <b>28.17</b> | <b>0.016</b> |

<sup>†</sup>Geometry-matched baselines. CityDreamer4D and StreetScapes are paradigm references; StreetScapes uses published results and a different evaluation protocol. Percentage gains compare with the best competitor on each metric.

### What drives the improvement?

| Ablation      | LPIPS | FID   | KID   |
|---------------|-------|-------|-------|
| w/o CtrlNet-N | 0.382 | 54.54 | 0.053 |
| w/o AGInpaint | 0.422 | 51.12 | 0.046 |
| w/o GCAAlign  | 0.346 | 26.25 | 0.013 |
| Full MeSS     | 0.348 | 28.17 | 0.016 |

GCAAlign removes exposure seams, with a small loss in fine detail; disabling it slightly improves the numerical metrics.