Neural scene representations for learning-based view synthesis and its applications

Phong Nguyen-Ha

Center for Machine Vision and Signal Analysis

University of Oulu, Finland

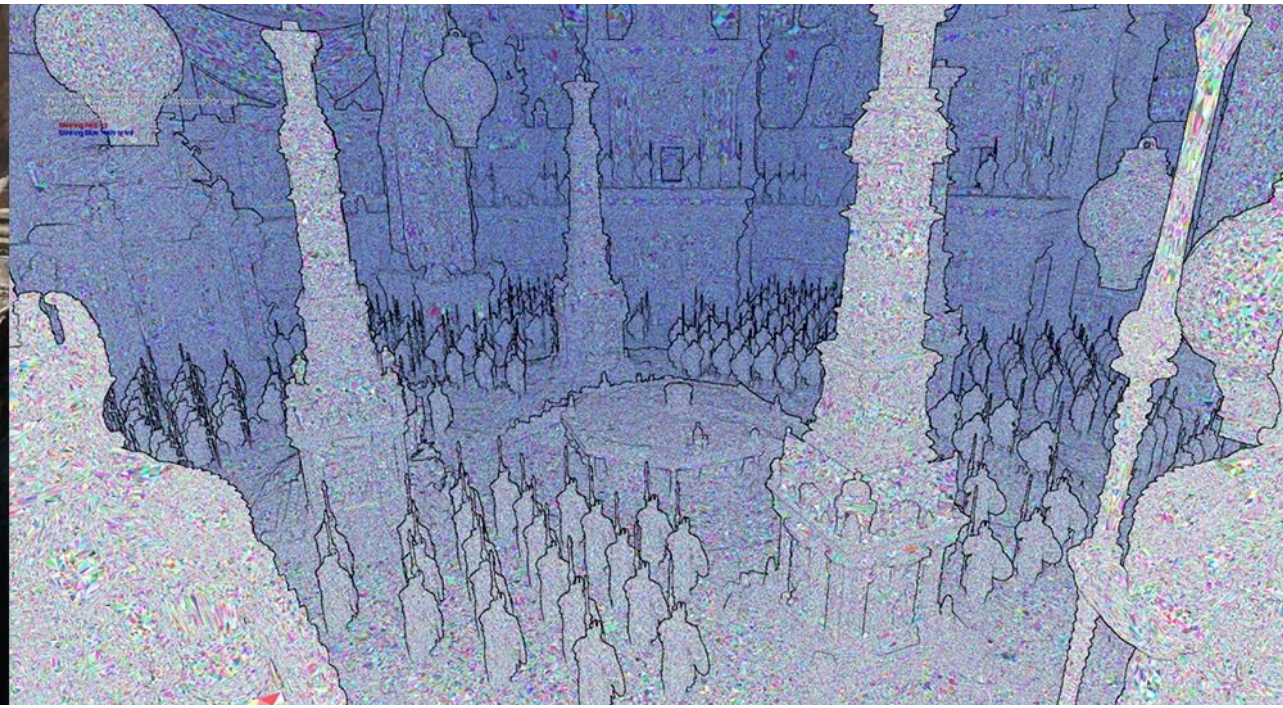
Outline

- **Introduction and motivation**
- Background on learning-based view synthesis
- Plane-sweep volume representation
- Sphere-based dynamic human rendering
- Generative AI and future works

Generating realistic images (Computer graphics)

Photorealistic rendering

Curated scene geometry



Source: UE 5 Documentation

Pro: full control of the scene parameters such as cameras, lights, motion, geometry, etc.
Cons:

- Requires lots of manual works
- Long rendering time

=> **Neural rendering to the rescue !**

Motivation for neural rendering

Transparency objects

Glossy objects

Digital human (skin, cloth, faces)

Neural rendering is a deep neural networks that can synthesize images and videos and allows:

- Full control of scene parameters
- End-to-end training
- Interactive inference/ rendering

Question: How to train and represent the 3D scenes ?

Source: NV, NeX

Outline

- Introduction and motivation
- **Background on learning-based view synthesis**
- Plane-sweep volume representation
- Sphere-based dynamic human rendering
- Generative AI and future works

General neural rendering pipeline

A collection of 2D images

Neural Scene Representation

Target views

Differentiable Neural Renderer

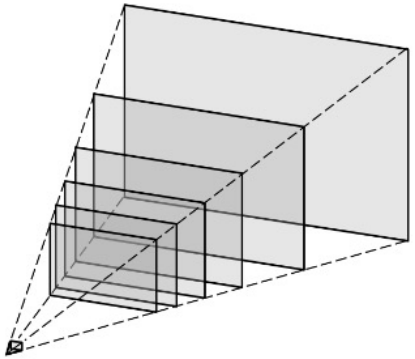
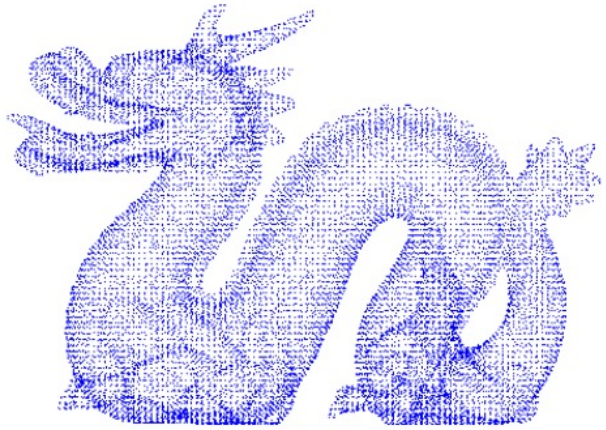
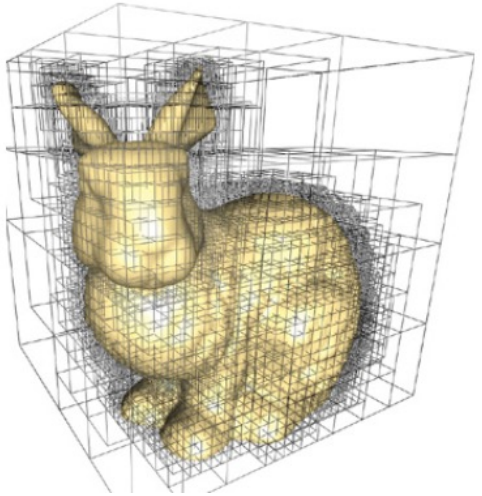
Rendered novel views

Image Loss

- Neural scene representations are learned features of the scene from 2D posed images
- A differentiable renderer generates novel views at an arbitrary viewpoint.

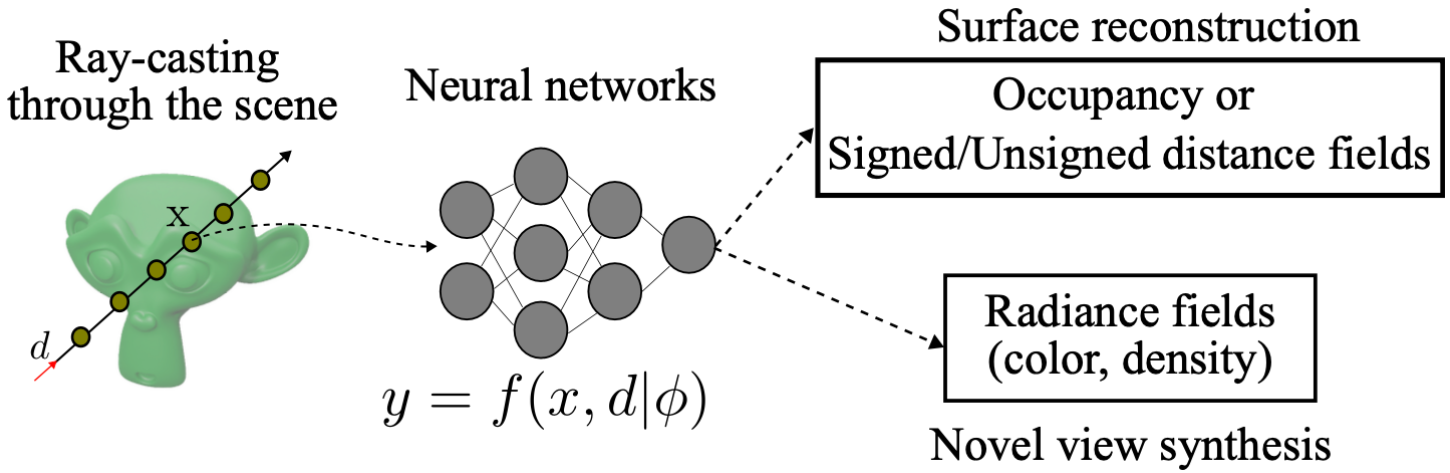
Source: NeRF

Explicit scene representations

	Multiple plane images	Point-cloud	Voxel-grid (sparse occtree)
Scene representation			
Rendering function	Alpha compositing	Rasterization	Volumetric Ray-based

The main issue of explicit representation is the scaling problem, the bigger the scene the larger the learned representation !

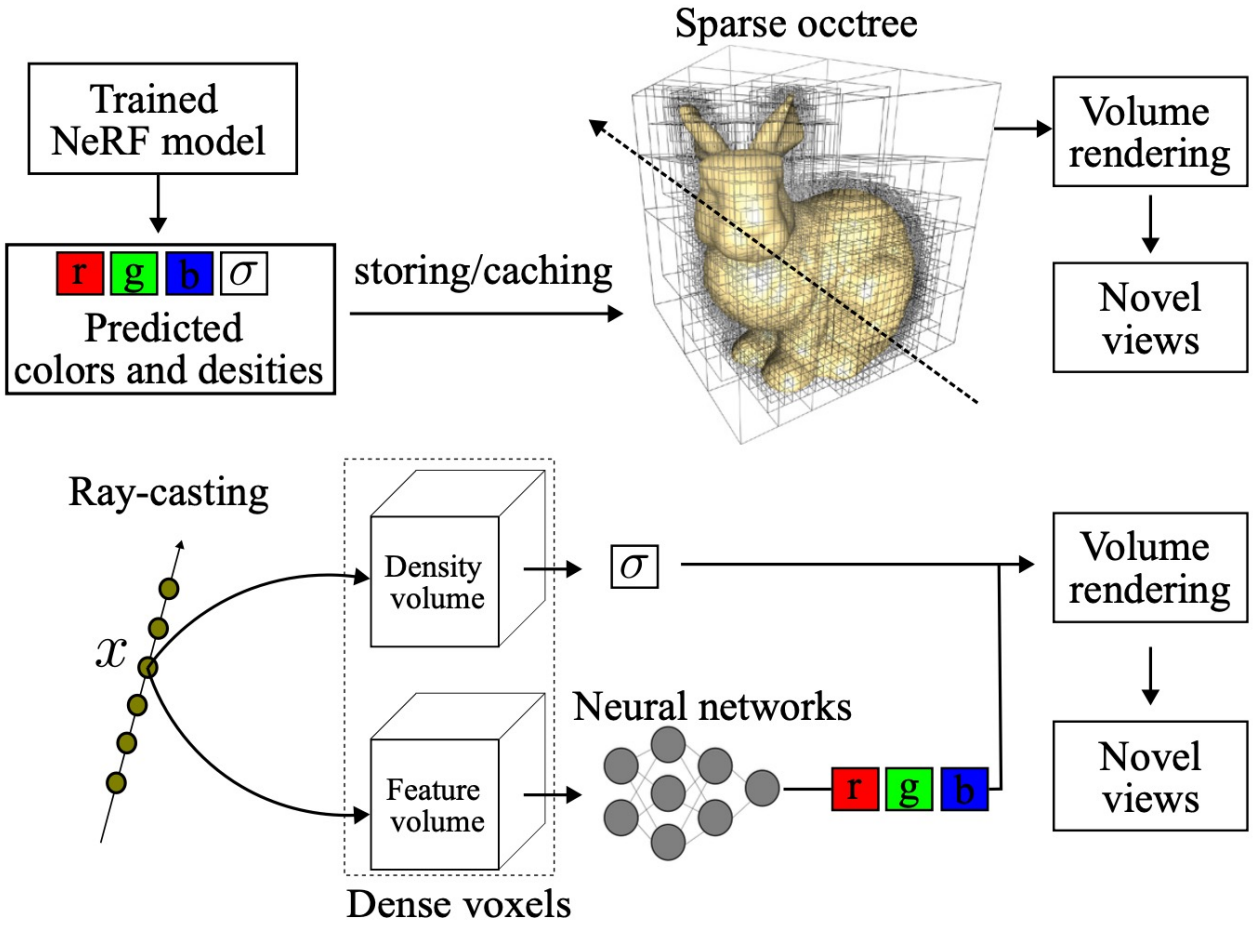
Coordinate-based representations



- A mapping function ***f*** between 3D coordinates ***x***, viewing direction ***d*** can serve as an implicit scene representation.
- The rendered novel views are high quality, and the extracted geometry is accurate.
- Rendering speed is significantly slower than explicit representations.

Source: NeuS

Hybrid representations



Instant-NGP, Siggraph 2022

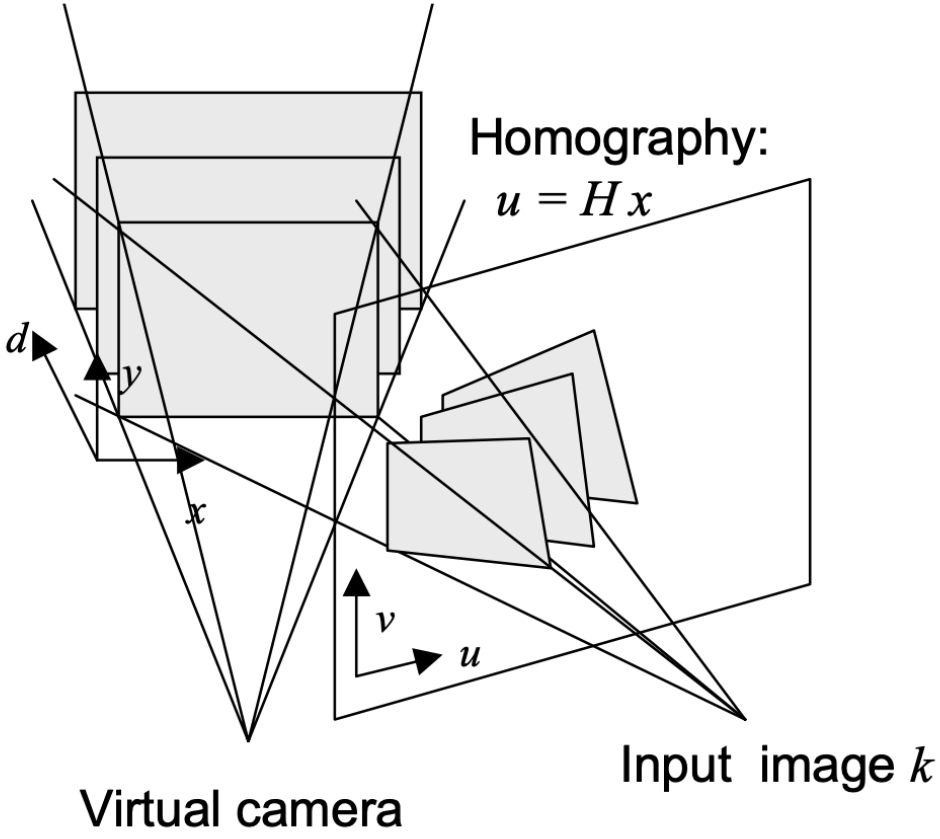
Hybrid approaches address the slow training/rendering issue by proposing different explicit scene representations such as:

- Sparse octree
- Tensor grid
- Multi-resolution hashgrids

Outline

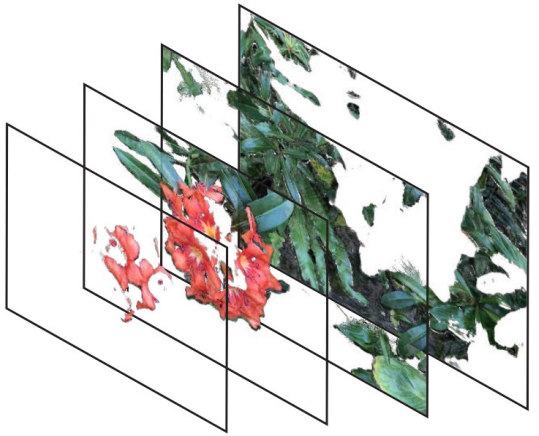
- Introduction and motivation
- Background on learning-based view synthesis
- **Plane-sweep volume representation**
- Sphere-based dynamic human rendering
- Generative AI and future works

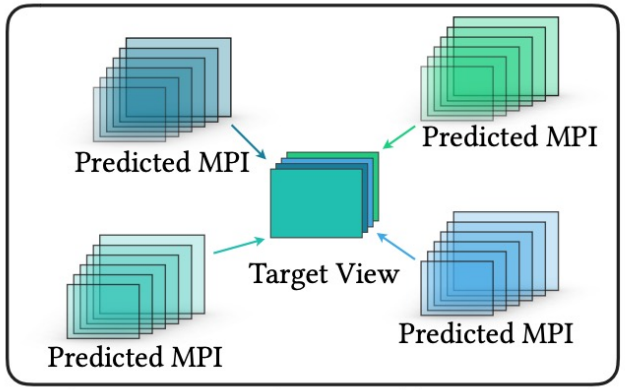
Plane Sweep Volume (PSV) and Multiple-plane images (MPI)



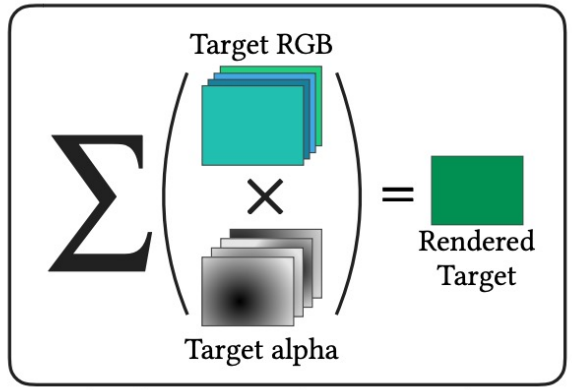
Input view

MPI inference





Render target view from neighboring MPIs by homography warping and alpha compositing



Blend RGBA renderings together to render final output image

Source: [Local Light Field Fusion]

- The basic idea of constructing a **PSV** is to back-project the input image onto successive virtual planes of the target view.
- The projected **PSVs** are then fed to a deep network to infer **MPI** and render the novel views.

Multiple-plane images (MPIs)

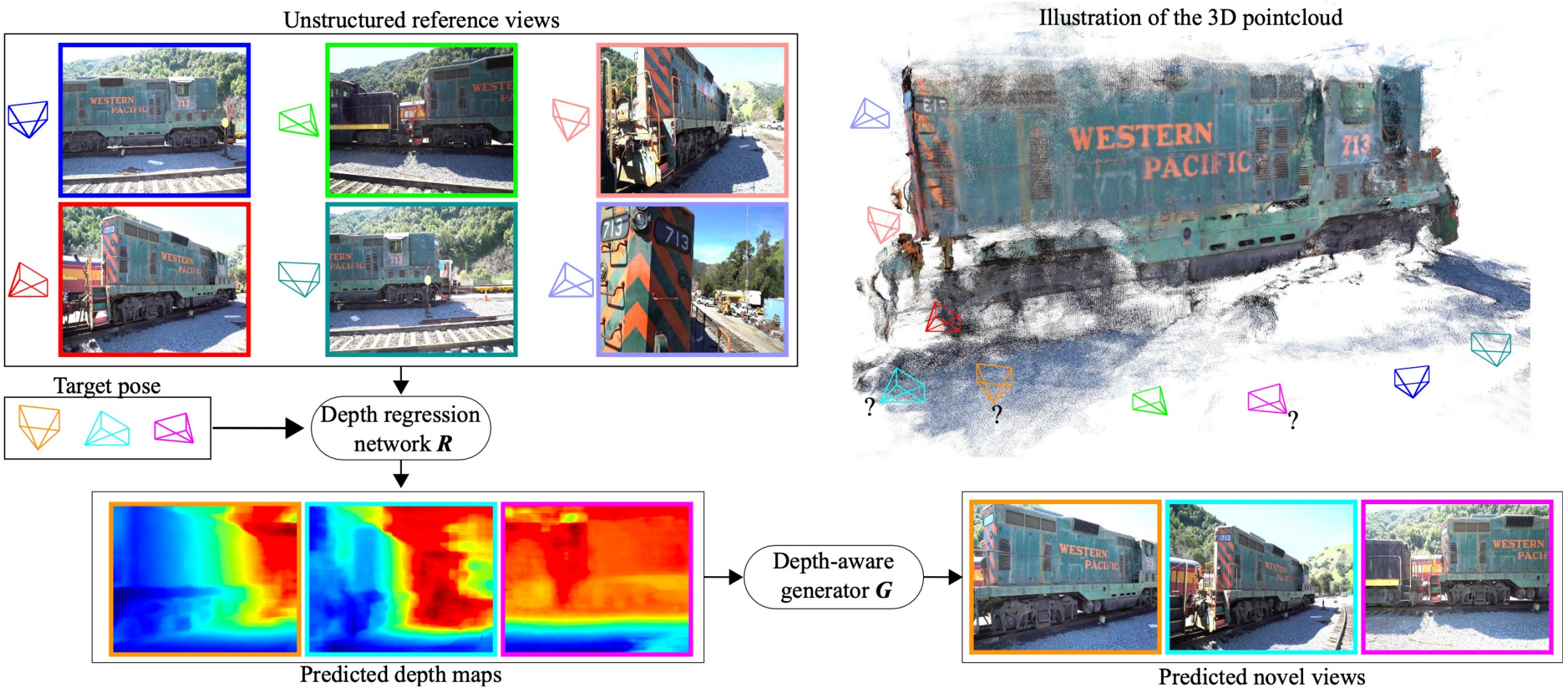
Pros:

- Produce high quality near-by novel views
- Fast to render

Cons:

- The learned represenation is 2.5D
- Require a lot of GPU memory to store MPIs

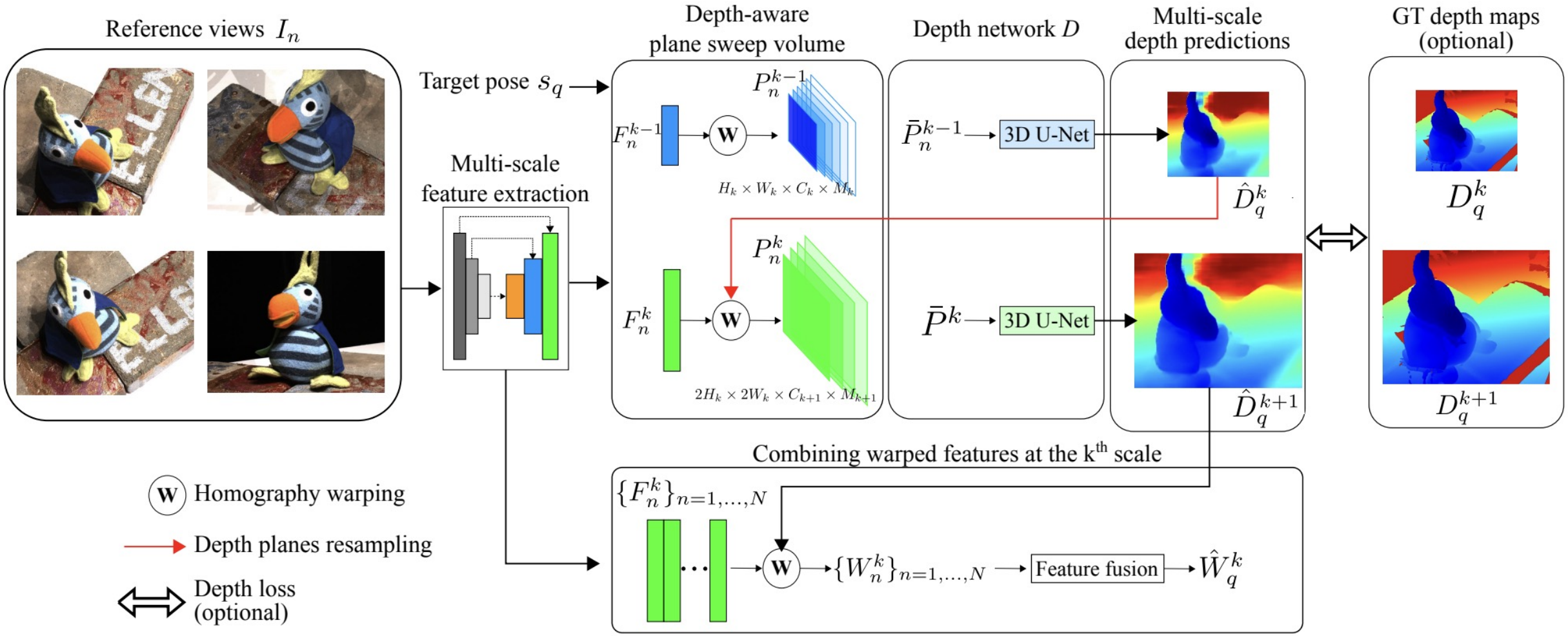
RGBD-Net: Predicting color and depth images for novel views synthesis



Overall pipeline of RGBD-Net:

- Instead of directly estimating novel views, RGBD-Net first predicts the target depth map using a regression network.
- A depth-aware generator synthesizes novel views by combining depth-based warping features.

Depth regression network



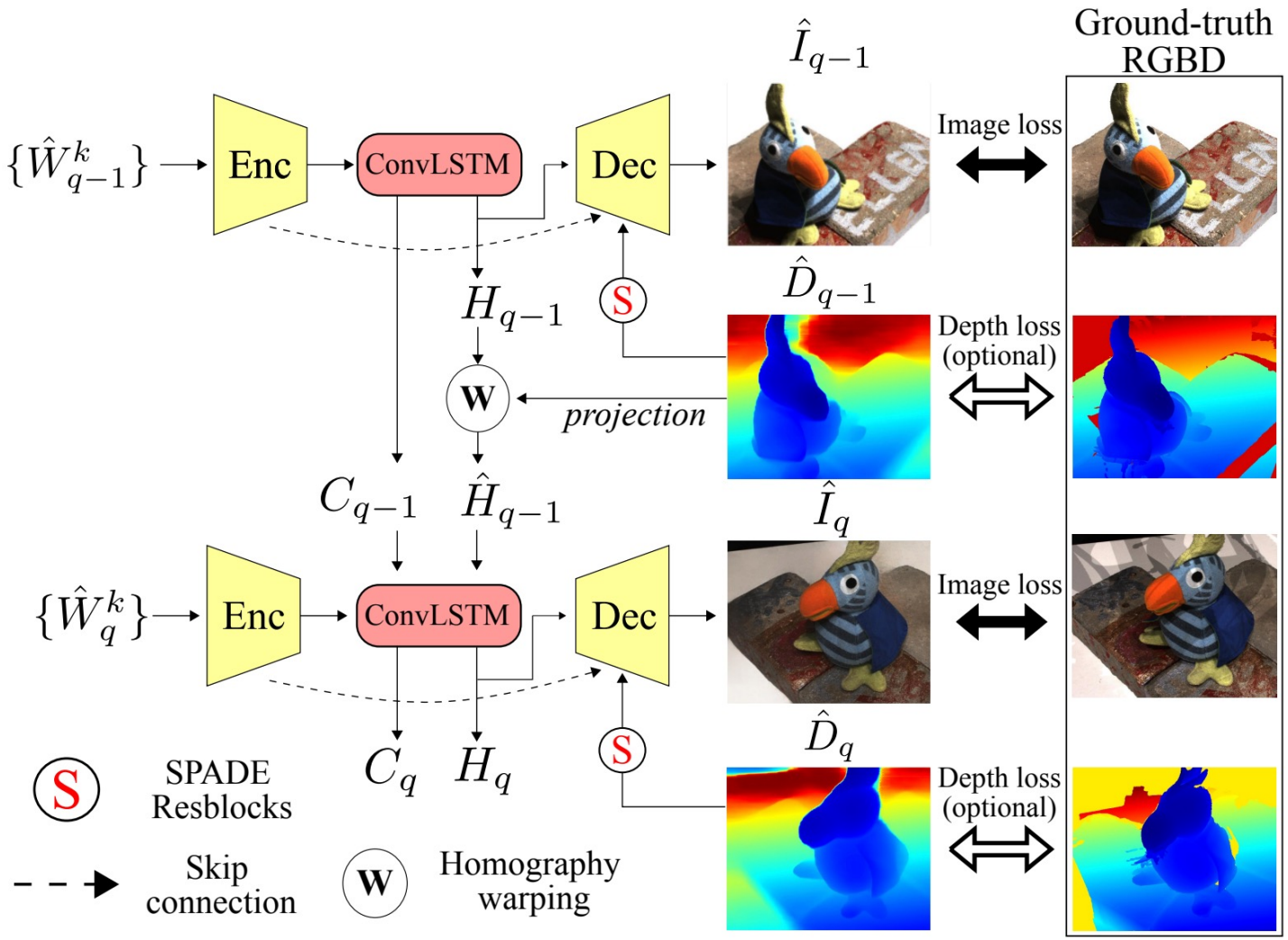
- Multi-scale depth predictions are used to warp multi-scale feature maps to the novel views
- The fusion weights are based on the inverse z-distance between the estimated 3D points in each reference viewpoint.

Depth-aware generator network

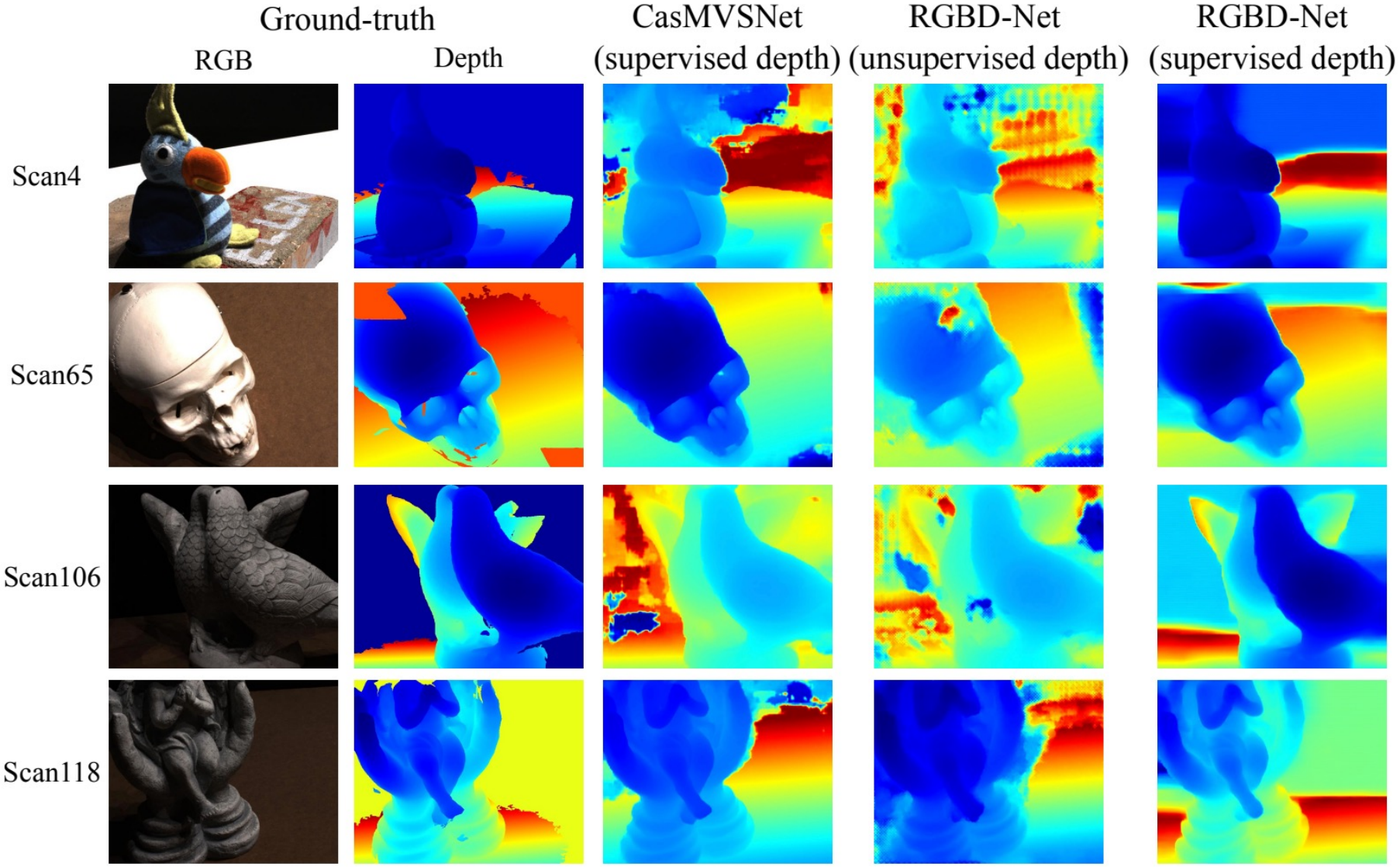
- Generating videos along smooth camera paths is potentially subject to temporally inconsistent. A ConvLSTM network is proposed to model spatial-temporal relations between rendered novel views

  => **requires a long training time to achieve 3D consistent results.**

- Training losses: L1 loss, Perceptual loss, hinge GAN loss, depth loss (optional).

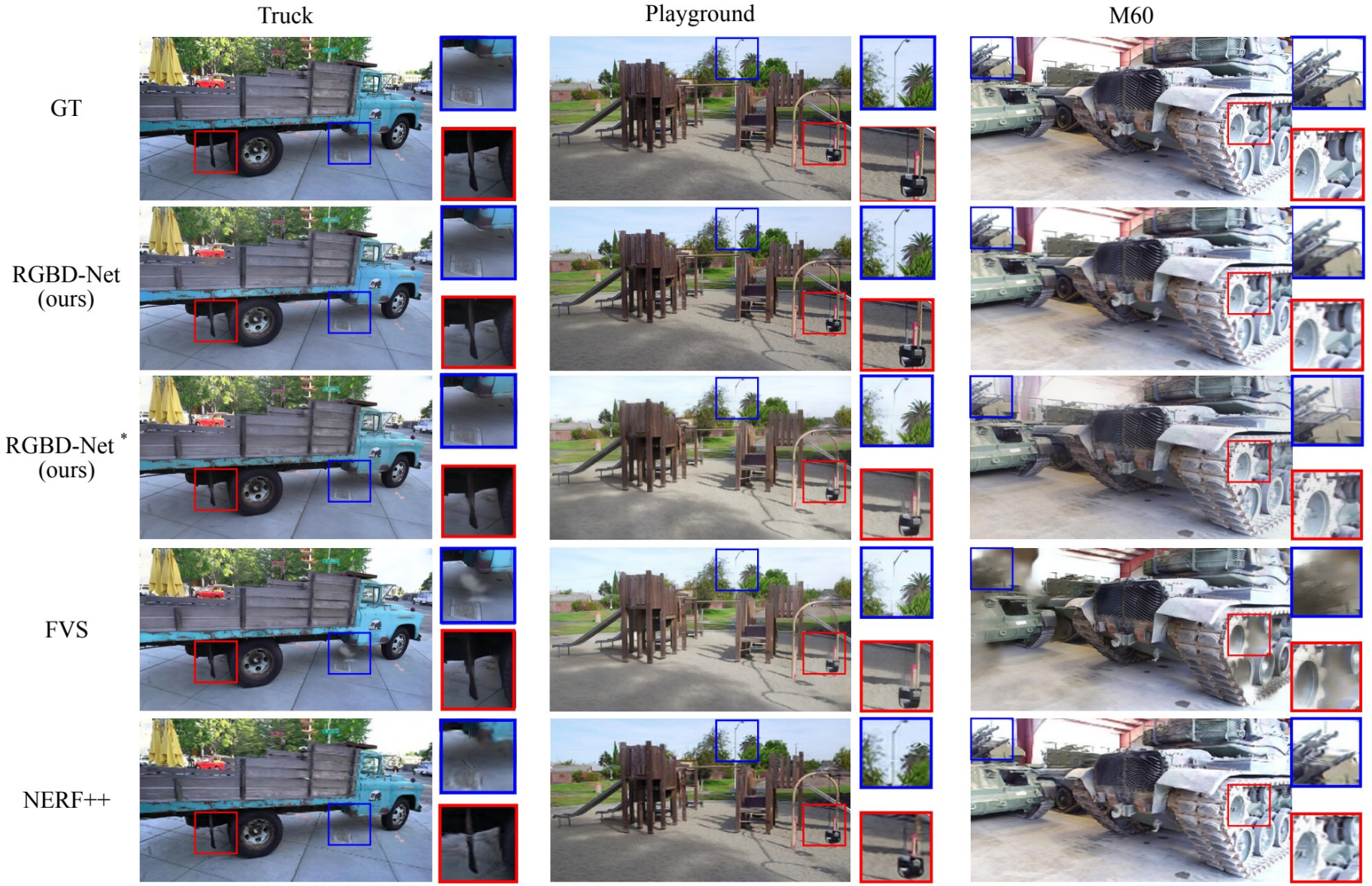


Synthesis results of RGBD-Net



With or without depth supervision, RGBD-Net outperforms CasMVSNet which was trained using the GT depth loss.

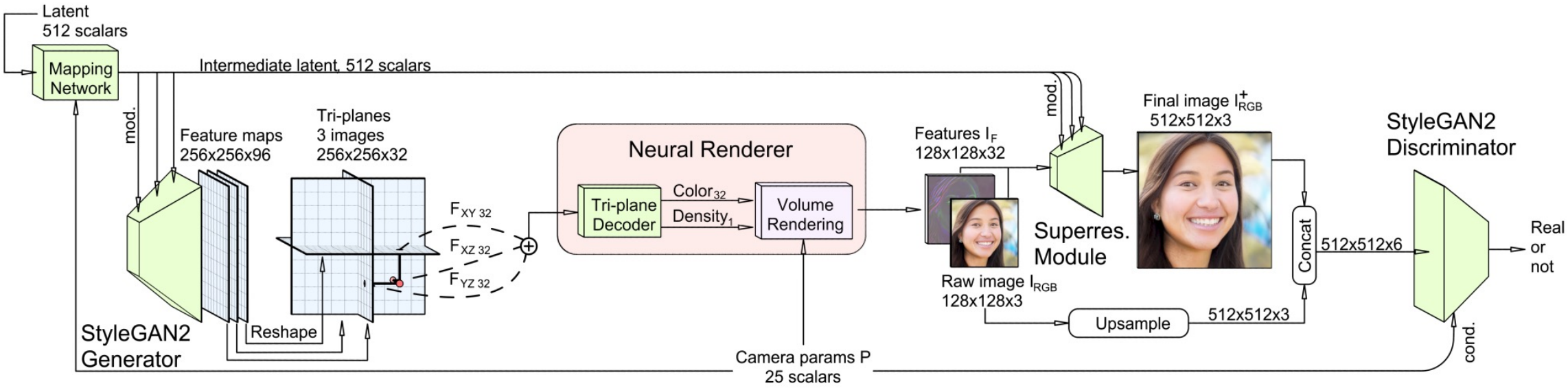
Synthesis results of RGBD-Net



Question: Can we improve the rendering results without relying on the depth supervision ?

Neural radiance fields to the rescue !

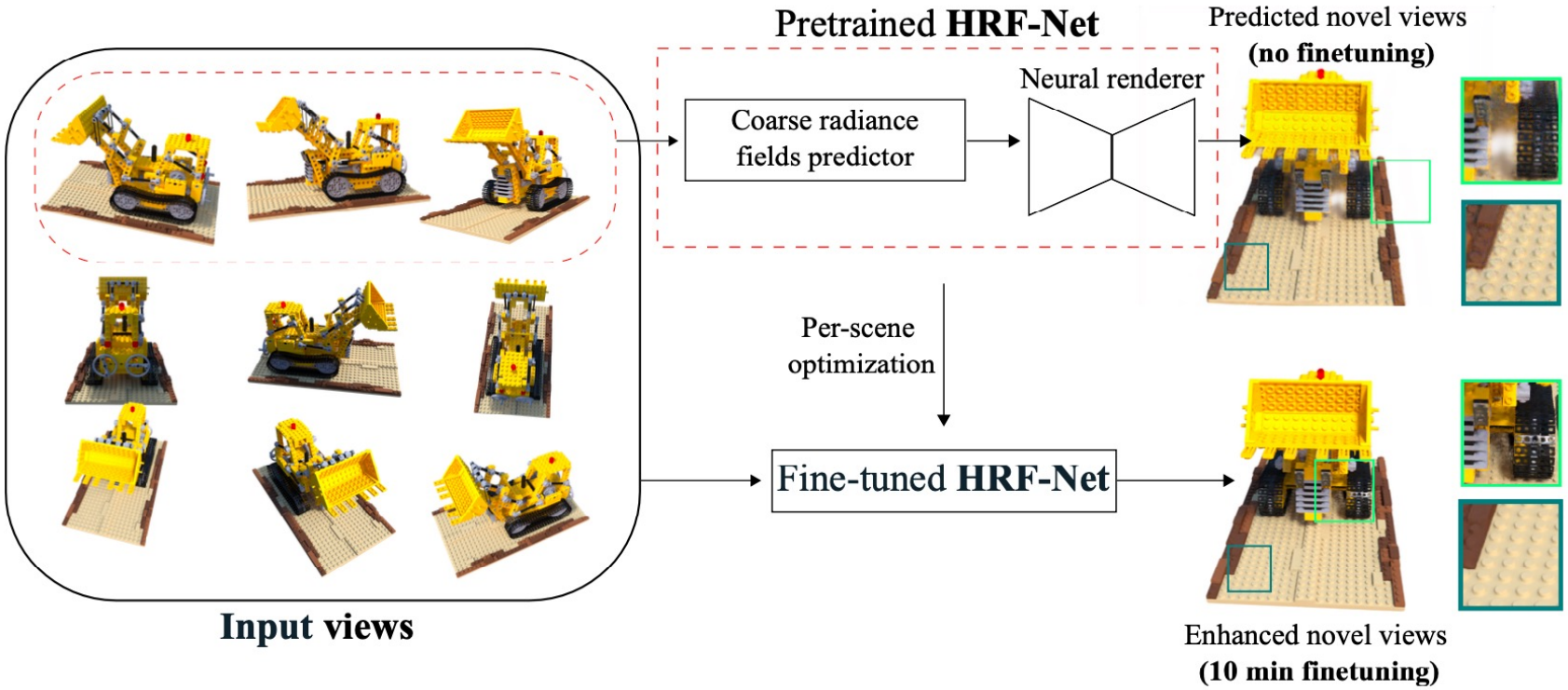
EG3D: Efficient Geometry-aware 3D Generative Adversarial Networks



Synthesized low and high-res images

Interpolation results

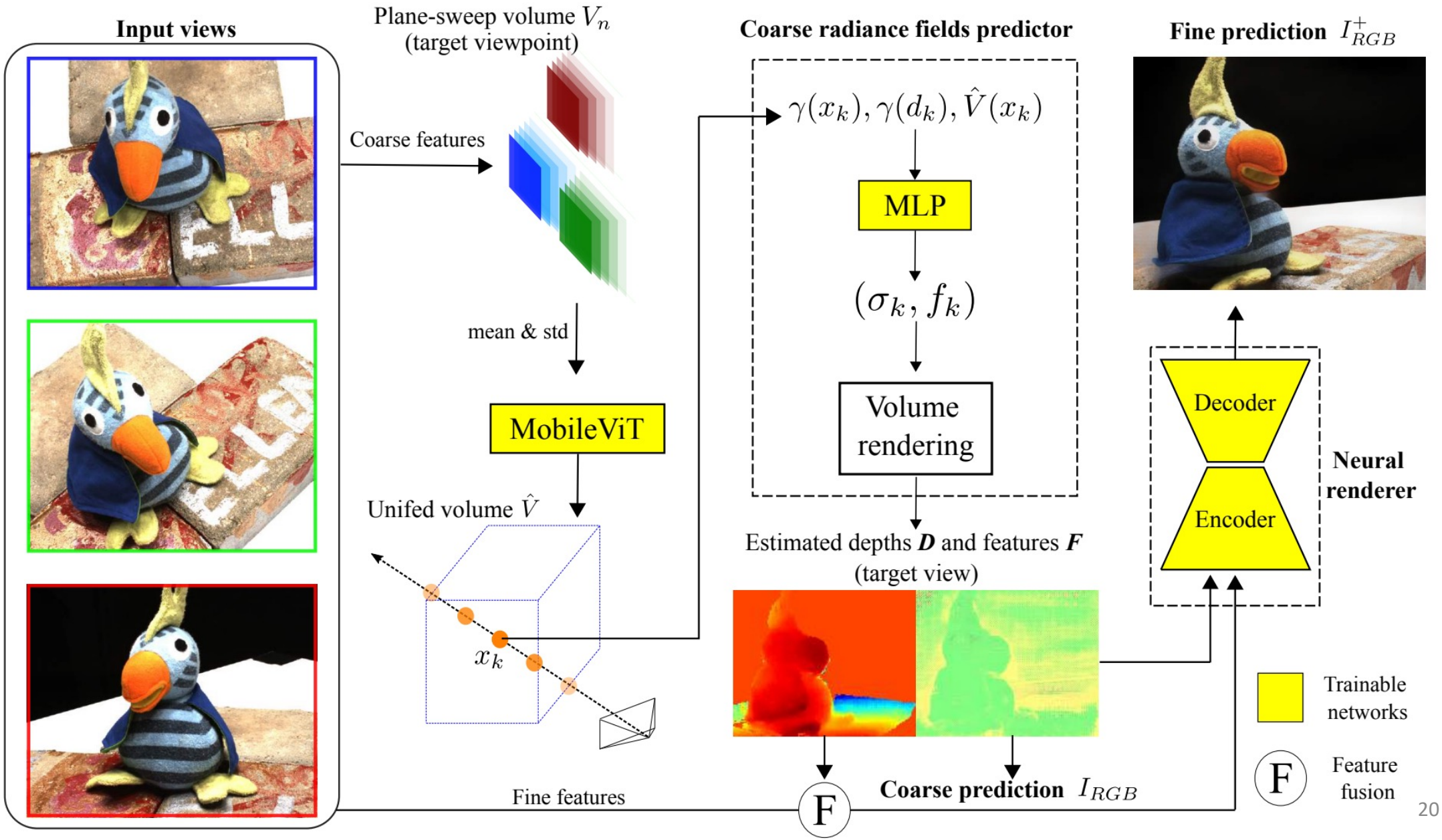
Holistic Radiance Fields Network (HRF-Net)



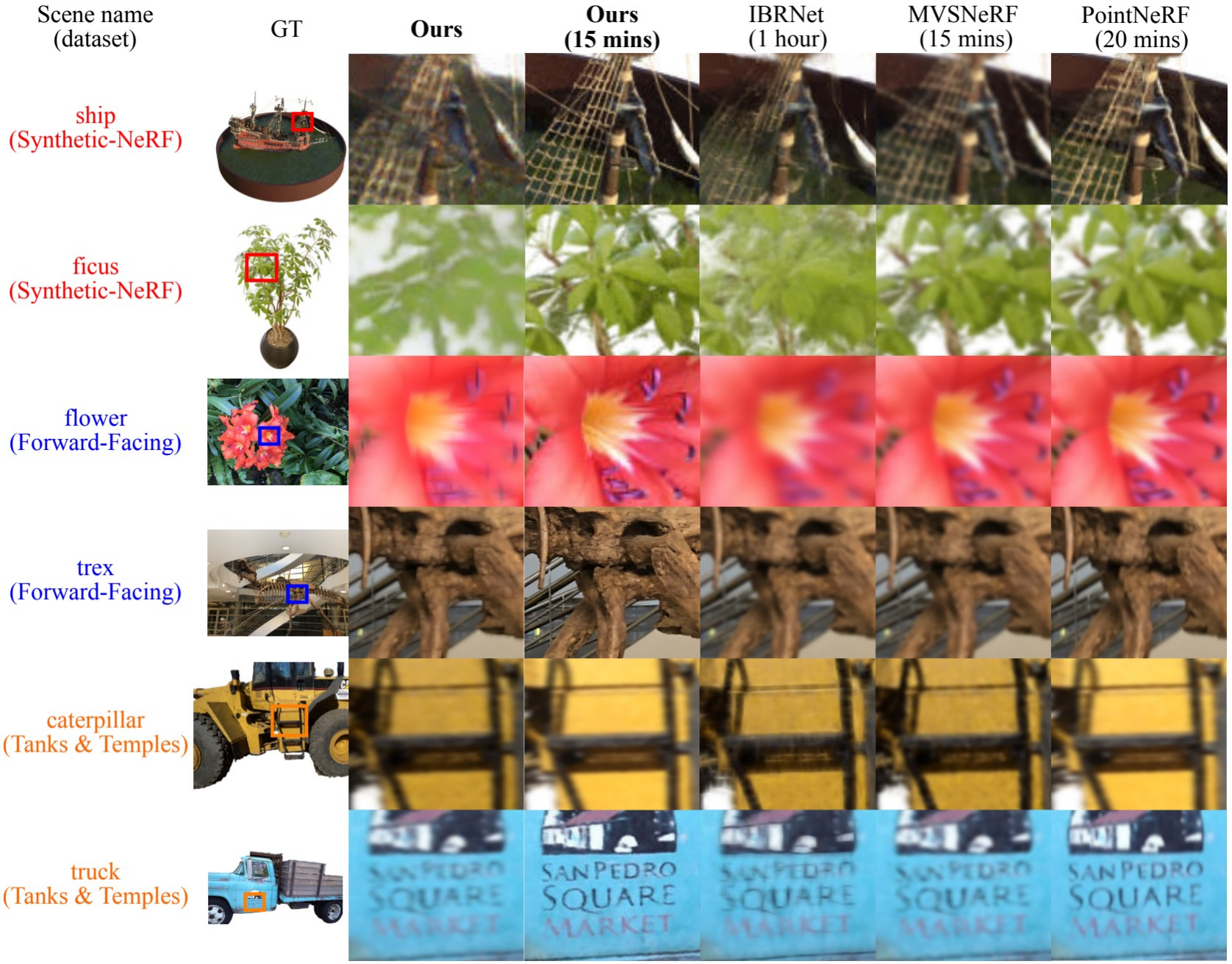
Key observations:

- The learned NeRF representation produces 3D consistent geometry regardless of input image resolutions.
- Recent work on 3D-aware generative model (EG3D) has shown that a super-resolution module can be used to obtain consistent novel views.

Overview of HRF-Net



Synthesis results of HRF-Net



Key results:

- The predicted novel views generated by HRF-Net (with or without finetuning) is much sharper than other baselines.
- The rendering speed of an 800x800 image is 3-10 times faster using the same V100 GPU.

Synthesis results of HRF-Net

HRF-Net is able to render free-view point video given a set of sparse input views

Outline

- Introduction and motivation
- Background on learning-based view synthesis
- Plane-sweep volume representation
- **Sphere-based dynamic human rendering**
- Generative AI and future works

Motivation

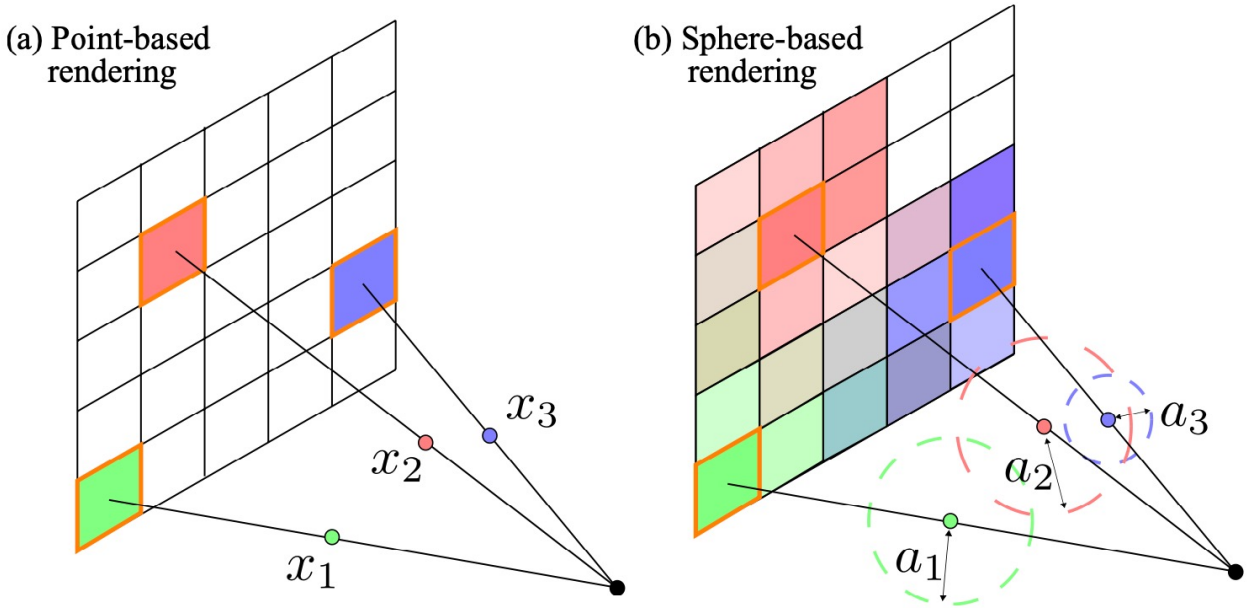
- **Input**: a sparse RGBD image of the "Main" view.
- **Output**: RGB at the novel camera.
- **Goal**: High fidelity and temporal consistent novel views.

Related Work: LookingGood (LK)

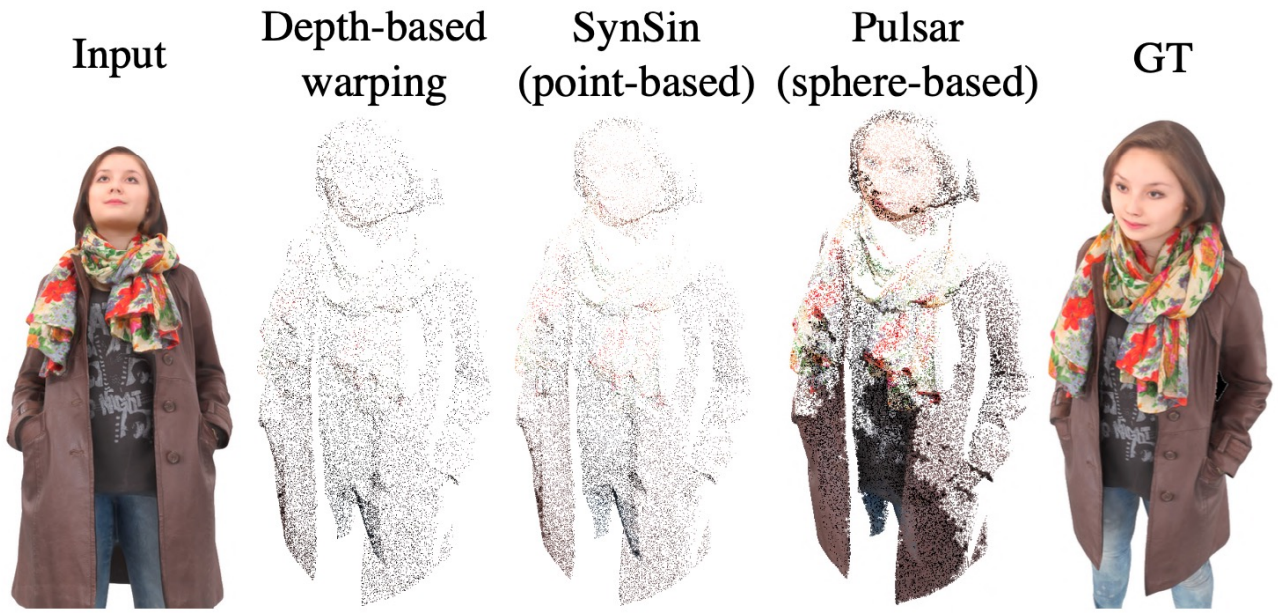
- LK use 16 IR cameras and 8 RGB cameras to obtain a temporal-consistent full-body mesh.
- LK use a high-res RGB camera for the frontal "Main" view.
- Render novel views from a single RGBD image using *a dense depth map* (extracted from the mesh).

LookinGood: enhancing performance capture with real-time neural re-rendering, ACM Transactions on Graphics, 2018

Sphere-based rendering



Point vs **Sphere-based** rendering

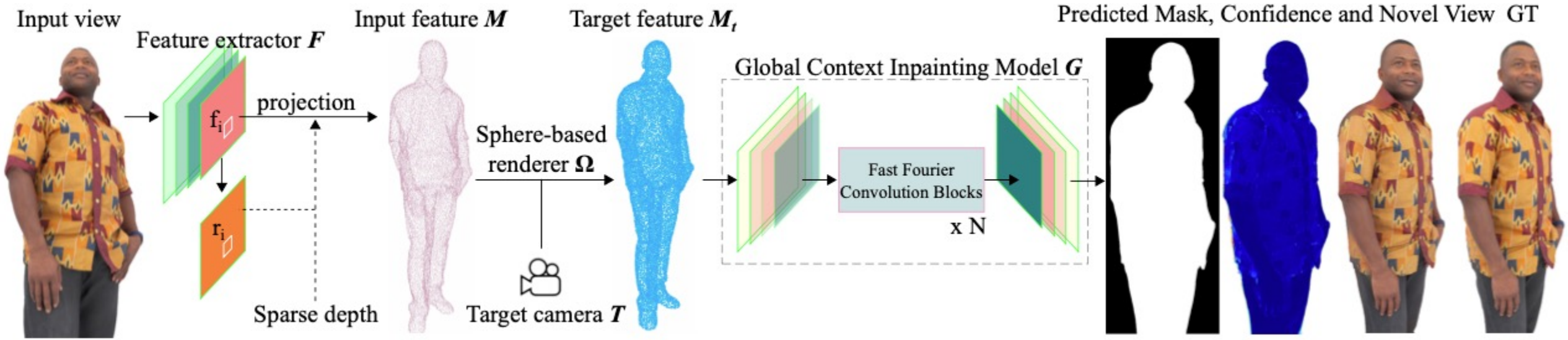


Warped novel views from a single RGBD image

Key observations: Sphere-based rendering produces much denser features maps.

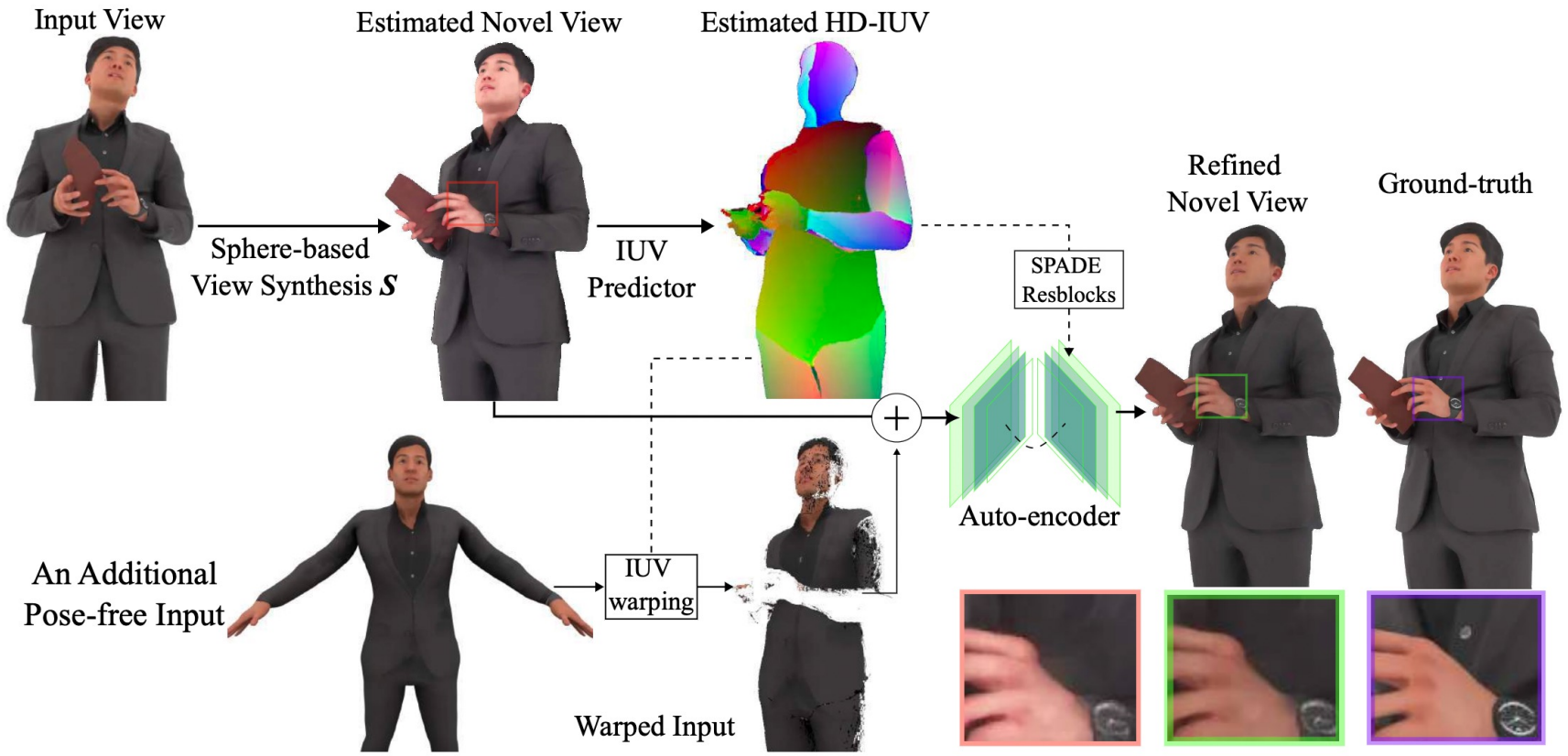
=> Less pixels to inpaint.

Sphere-based view synthesis network



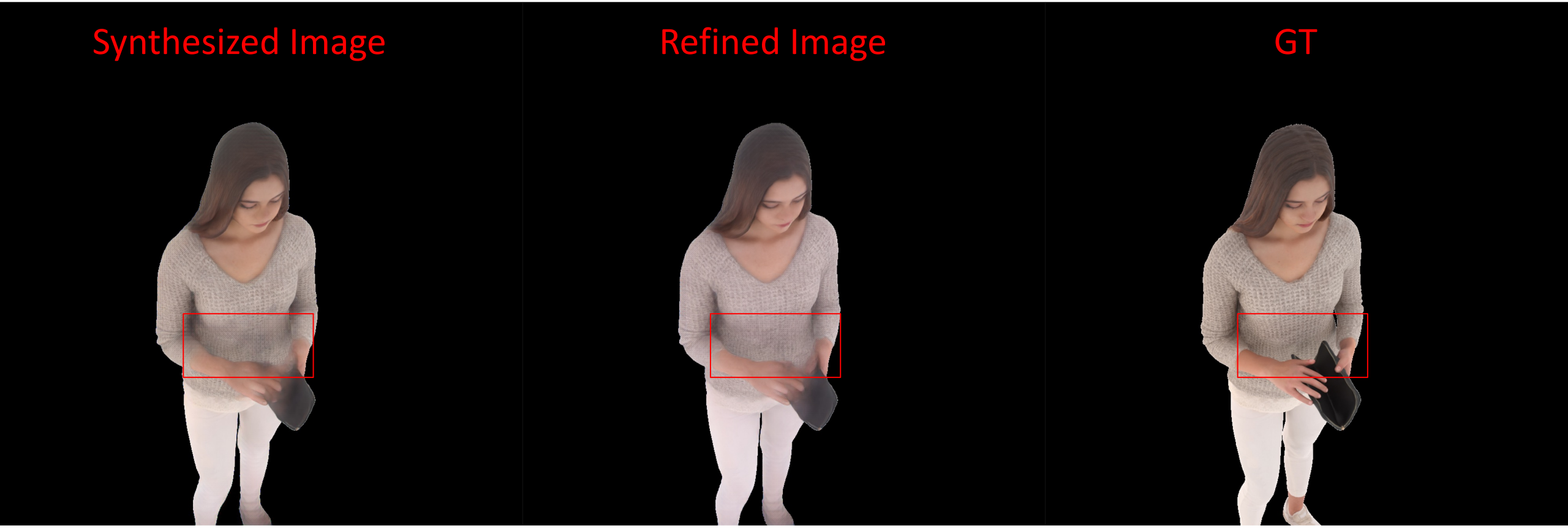
- Instead of warping RGB values to the novel views, we warp deep features of the input views.
- Per-pixel radius is estimated using a single CNN layer with sigmoid activation.
- Our model predicts not only color of the novel view but also foreground, confidence masks.

Enhancer network



- Using an additional occlusion-free input, we enhanced the “hard” occluded region which are not visible in the input view.
- We propose a HD-IUV network that predict the dense human correspondence.
- Finally, we refine the initial estimates with the warped input using predicted IUV.

A comparison between synthesized image and a refined one



Notice how the model is able to generate more realistic texture in the self-occluded area.

Input views

Generated novel views

Input views

Generated novel views
(dynamic trajectory #1)

Generated novel views
(dynamic trajectory #2)

Live Demo

Our method is able to render 512x512 novel views at 30fps using NVIDIA RTX 2080 TI

Outline

- Introduction and motivation
- Background on learning-based view synthesis
- Plane-sweep volume representation
- Sphere-based dynamic human rendering
- **3D Generative AI and future works**

THE *GENERATIVE AI* STARTUP LANDSCAPE

ANTLER

TEXT

Smartwriter.ai · Hypertype · Lately · Autobound · Writesonic · Jasper · cogram · genei · YOU · AI21labs · letterdrop · copysmith · Creatext · jenni · mavenoid · anyword · [PERSADO] · frase · regie.ai · Soloce Vision · Mutiny · Linguix · Hypotenuse AI · WRITER · OTHERSIDE AI · copy.ai · copymatic · COMPOSE AI

IMAGE

ClipDrop · p·e·n·c·i·l · beautiful.ai · PhotoRoom · BRIA · Facet · Poly · CSM · Blend · Let's Enhance.io · HYPAR · stability.ai · KAEDIM · ROSEBUD.AI · maket · Autoenhance.ai · BOTIKA · soul machines · Sloyd · MODULIZE · Re:cast AI · uizard · imagen · Hexo AI

AUDIO

MURF AI · REPLICA · notably · Endel · WELLSAID · AssemblyAI · DEEPGRAM · krisp · Speechify · RESEMBLE.AI · AIMI · KAIZAN · coqui · Mubert · Neural Space · soundful · PAPERCUP · moises · VOICEMOD · Listnr · LOVO · Vocal Clarity · Dubverse · AD AURIS

CODE

Debuild · tabnine · Codiga · Locofy · AIXcoder · Mintlify · māyā · MutableAI · Codis · durable · The.com · bloop. · replit · ENZYME · DhiWise · codota · anima · CODACY · warp · Metabob

CHATBOTS

lang.ai · PolyAI · Tymely · Incentivai · Kasisto · ushur · CRESTA · Elise AI · verloop.io · Replika · ultimate.ai · Cohere · Sapling · haptik · ada · Forethought · OBSERVE·AI · XOXind · Balto · Certainly.

VIDEO

ZUBTITLE · TERRA · Peech · VOCHI · Maverick · recut · VEED.IO · Basch.io · inworld · WOMBO · tavus · deepdub.ai · FATHOM · runway · Mario · XEMBLY · PICTORY · vidyo.ai · Rephrase.ai · lumen5 · Steve AI · windsor.io · YEPIC · deepdub.ai · Colossyan · METAPHYSIC · Potion

ML PLATFORMS

slai · jina · symbl.ai · Adept · aporia · GANTRY · deepset · Synthesis.ai · Archistar · Galileo · featureform

SEARCH

glean · Looria · Hebbia · consensus · Air · vectara · Pinecone · qdrant

GAMING

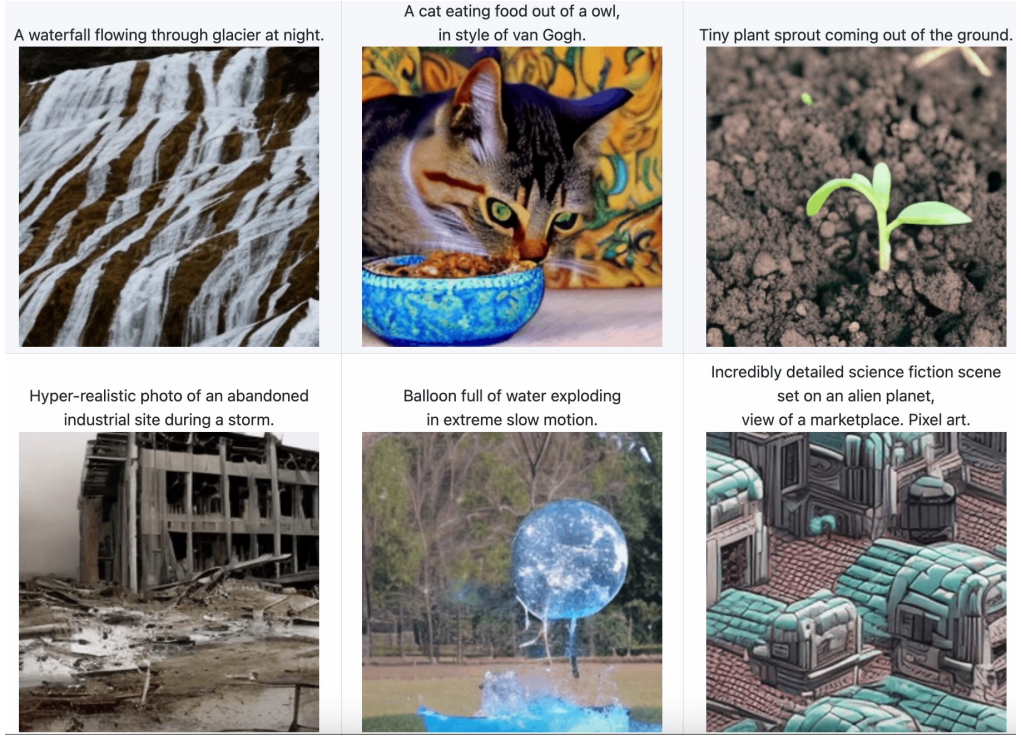
charisma.ai · hidden door · LATITUDE · Spellbrush

DATA

Pilot · gretel · DATAHERALD · SYNTEGRA Mirry.AI · BIFROST · datagen

Source: https://www.antler.co/blog/generative-ai

Text-to-video diffusion model



Large Language Models: GPTs from OpenAI, LLAMA from Meta, Bard from Google, etc.

Image generation: DALL-E, Mid Journey, Stable Diffusion, etc.

Video generation: Video fusion

Question: What about 3D, 4D and beyond ?
=> Again, neural rendering to the rescue !

Magic3D: High-Resolution Text-to-3D Content Creation

Chen-Hsuan Lin* Jun Gao* Luming Tang* Towaki Takikawa* Xiaohui Zeng*
Xun Huang Karsten Kreis Sanja Fidler† Ming-Yu Liu† Tsung-Yi Lin

*† : equal contributions

NVIDIA Corporation

IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2023

HIGHLIGHT

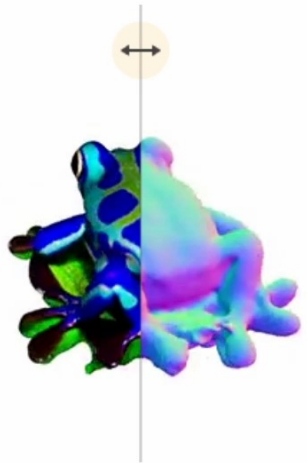
High-Resolution 3D Meshes

Magic3D can create high-quality 3D textured mesh models from input text prompts. It utilizes a coarse-to-fine strategy leveraging both low- and high-resolution diffusion priors for learning the 3D representation of the target content. Magic3D synthesizes 3D content with 8× higher-resolution supervision than DreamFusion while also being 2× faster.

[...] indicates helper captions added to improve quality, e.g. "A DSLR photo of".

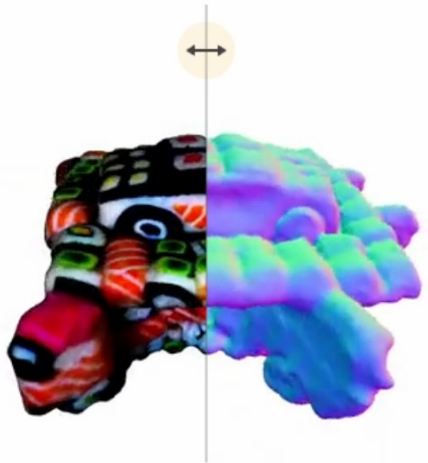
Reveal 3D mesh!

A beautiful dress made out of garbage bags, on a mannequin. Studio lighting, high quality, high resolution.



Reveal 3D mesh!

A blue poison-dart frog sitting on a water lily.



Reveal 3D mesh!

[...] a car made out of sushi.

GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models

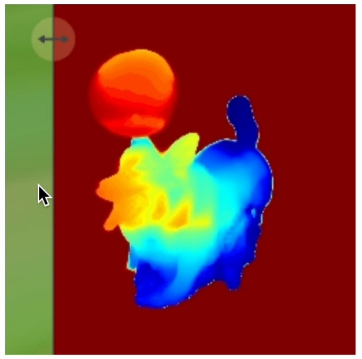
https://nvlabs.github.io/genvs/

Text-To-4D Dynamic Scene Generation

Uriel Singer* Shelly Sheynin* Adam Polyak*
Oron Ashual Iurii Makarov Filippos Kokkinos Naman Goyal
Andrea Vedaldi Devi Parikh Justin Johnson Yaniv Taigman

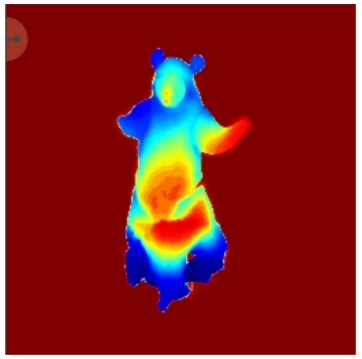
*Equal Contribution

Meta AI



Load mesh

An emoji of a baby panda reading a book.

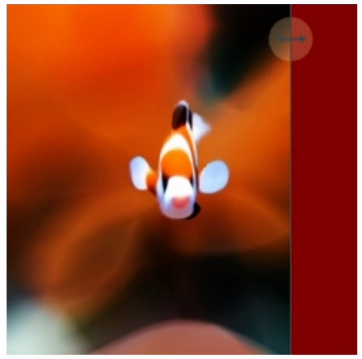


Load mesh

A dog riding a skateboard.

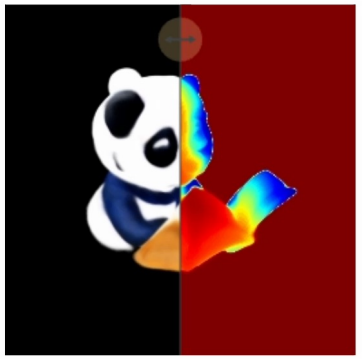
Load mesh

3D rendering of a fox playing videogame.

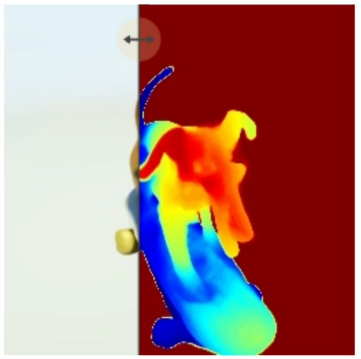


Load mesh

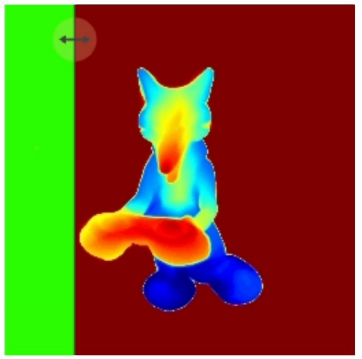
A squirrel riding a motorcycle.



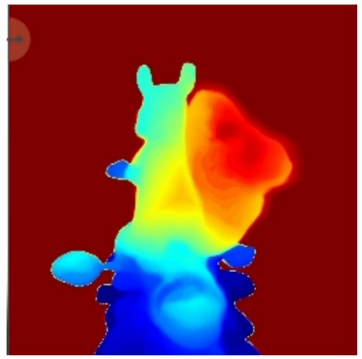
Load mesh



Load mesh



Load mesh



Load mesh

Scaling up GANs for Text-to-Image Synthesis

Minguk Kang[1,3], Jun-Yan Zhu[2], Richard Zhang[3], Jaesik Park[1], Eli Shechtman[3], Sylvain Paris[3], Taesung Park[3]
[1]POSTECH, [2]Carnegie Mellon University, [3]Adobe Research
in CVPR 2023 (Highlight)

GigaGAN: Large-scale GAN for Text-to-Image Synthesis

Can GANs also be trained on a large dataset for a general text-to-image synthesis task? We present our 1B-parameter GigaGAN, achieving lower FID than Stable Diffusion v1.5, DALL·E 2, and Parti-750M. It generates 512px outputs at 0.13s, orders of magnitude faster than diffusion and autoregressive models, and inherits the disentangled, continuous, and controllable latent space of GANs. We also train a fast upsampler that can generate 4K images from the low-res outputs of text-to-image models.

Be aware that GAN is not out-of-the-picture yet !


рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
gracias
хвала
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
thank you
asante
manana
bedankt
bayarlalaa
gràçie
hvala
köszönöm
aċiū
dankon
akun
obrigada
murakoze
tenki
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
děkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
arigatô
takk
dakujem
trugarez
ありがとう
tanemirt
grazie
didi madloba
najis tuke
terima kasih
rahmet
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
diolch
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci