# NViST : In the Wild New View Synthesis from a Single Image with Transformers

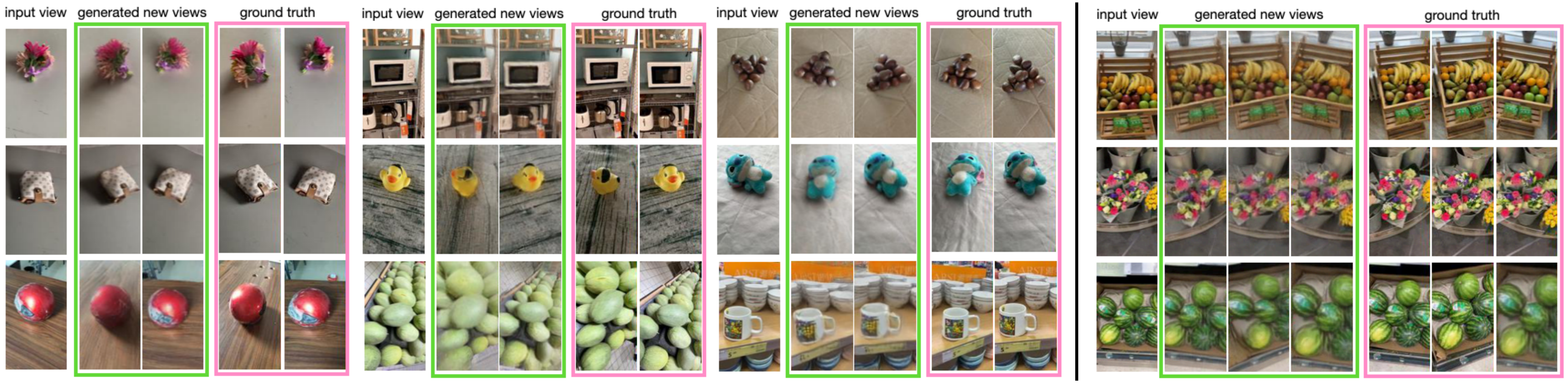
Wonbong Jang    Lourdes Agapito

**University College London**

CVPR SEATTLE, WA JUNE 17-21, 2024

Paper, code, and additional results can be found at
https://wbjang.github.io/nvist_webpage/

## 1. NViST turns in-the-wild single images into implicit 3D functions with a single pass using Transformers



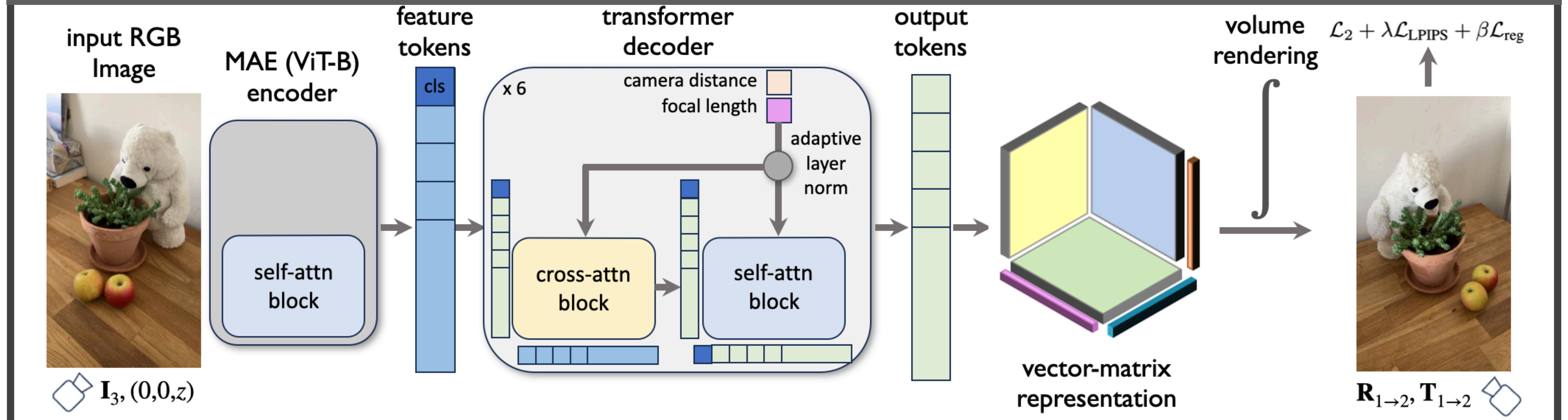
MVImgNet (Test scenes)

Phone Captures (Out of Distribution)

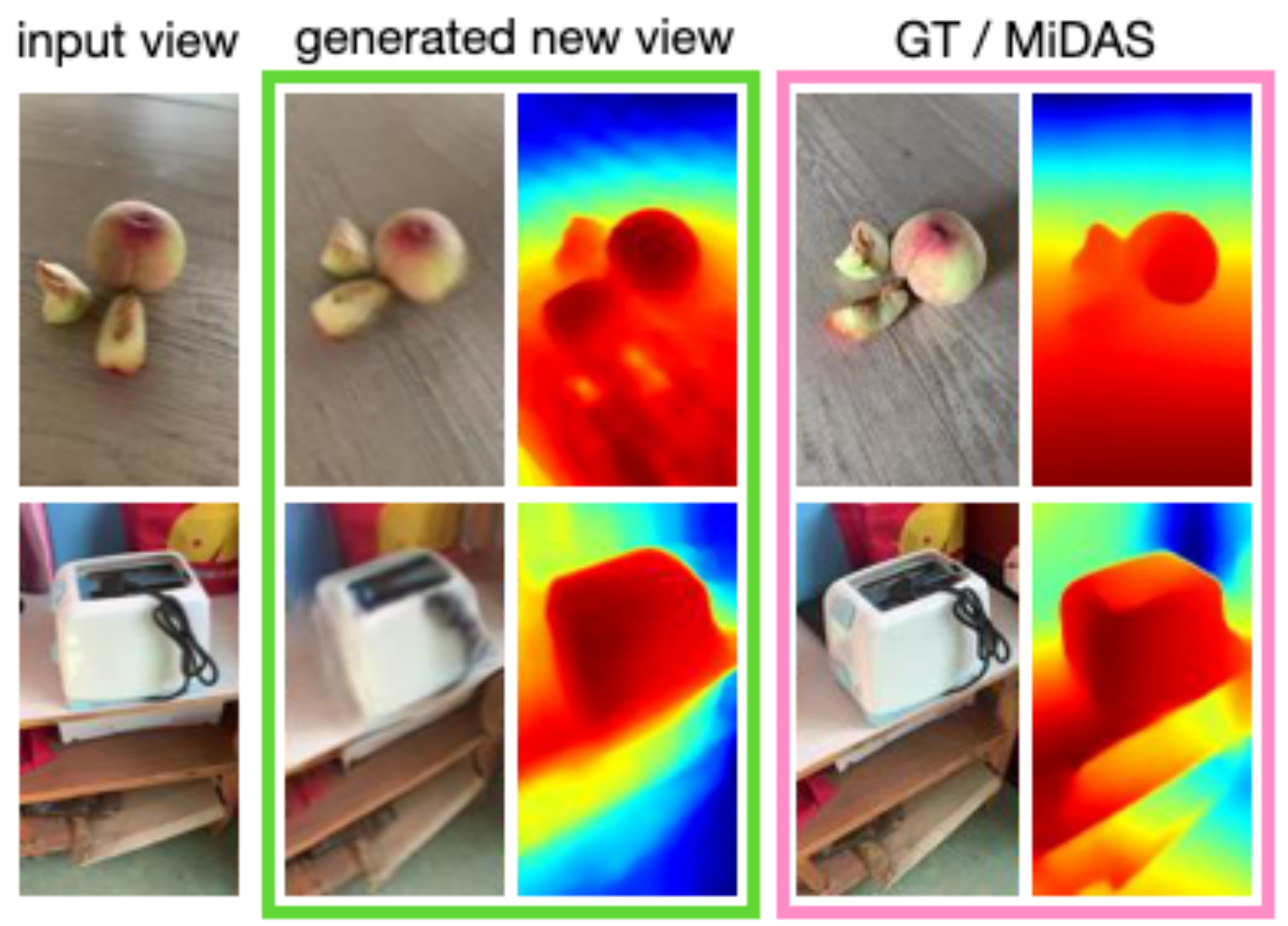
## 2. Contributions

- NViST **only requires relative pose** and doesn't require a canonicalized dataset
- Our novel decoder maps **MAE features to 3D tokens via cross-attention** and **AdaLN conditioned on camera parameters**.
- Ours is **category-agnoistic** , and **generalizes well** over test scenes and even phone captures.
- NViST assumes **6DoF camera** and does **not require the masked images**, unlike previous 3D-aware diffusion models.

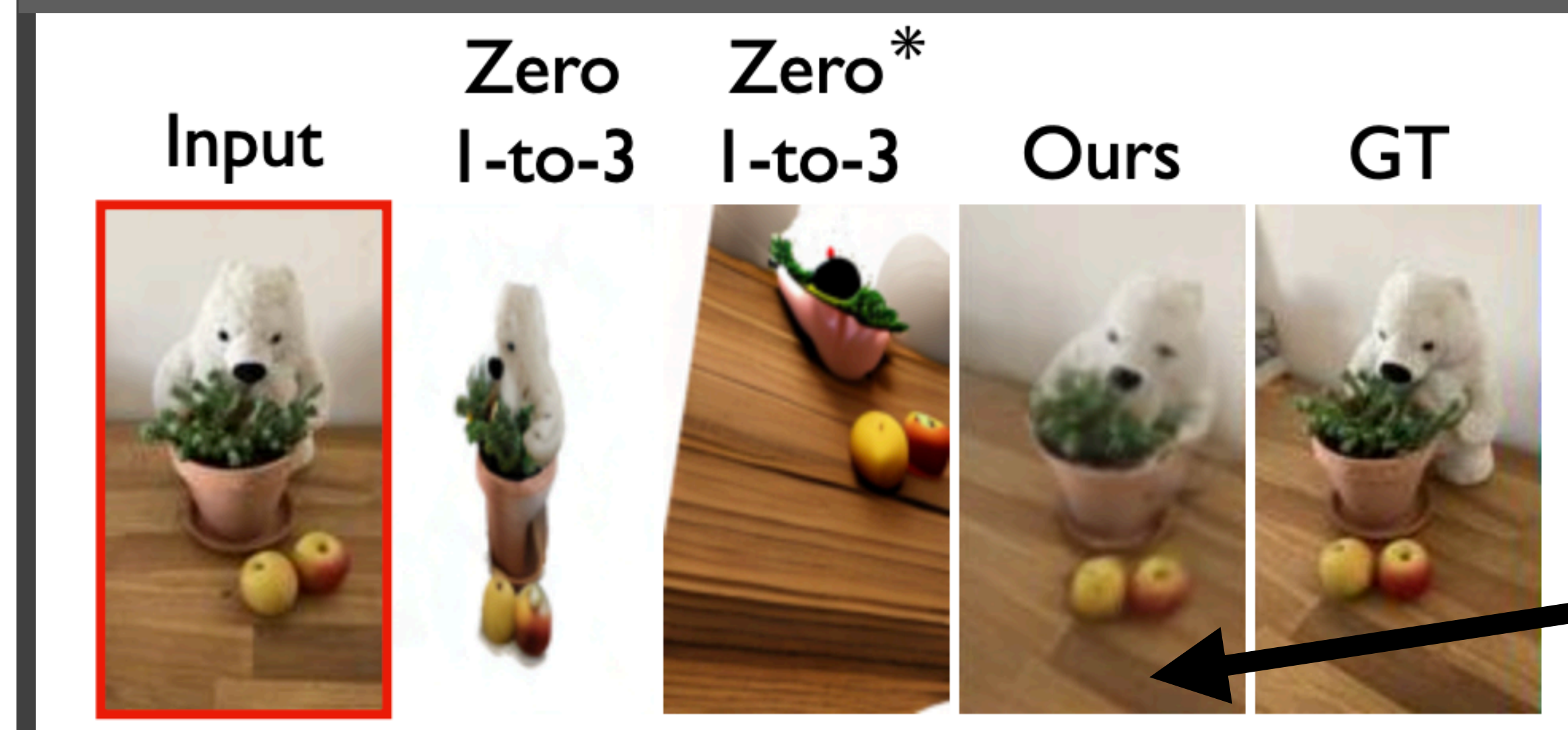
## 3. Methodology



## 4. Depth Maps



## 5. Comparison / Ablation



We can **capture the background**, and do not assume the simplified camera.

| | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| PixelNeRF | 17.02 | 0.41 | 0.54 |
| VisionNeRF | 19.82 | 0.51 | 0.47 |
| Ours | **20.83** | **0.57** | **0.29** |

| | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| Ours | **20.83** | **0.57** | **0.29** |
| w/o LPIPS | 20.72 | 0.54 | 0.46 |
| w/o Camera Conditioning | 20.24 | 0.49 | 0.36 |
| w/o VM Representation | 19.60 | 0.49 | 0.44 |
| w/o Updating Encoder | 18.54 | 0.47 | 0.49 |