EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF
SCIENCE

Real Virtual Humans

# Benchmarking Multi-View Image Generation for 3D Consistency Without Ground Truth

Philipp Bauer, Julia Schneider

# Motivation


[1] *Zero-1-to-3: Zero-shot 3D view synthesis from a single image (Liu et al., Columbia & TRI).*

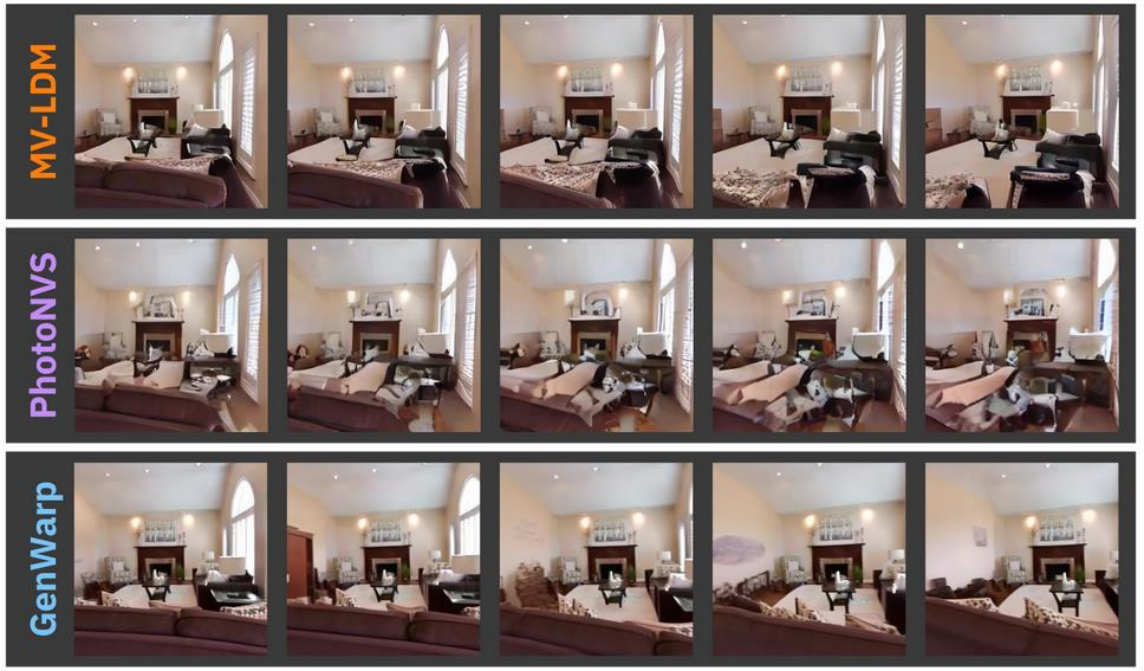
**How should we evaluate the generated views?**

# Related Work

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

## Generative models

- learn priors to synthesize novel views from limited images
- not rely on explicit 3D reconstruction, unlike traditional multi-view stereo or structure-from-motion

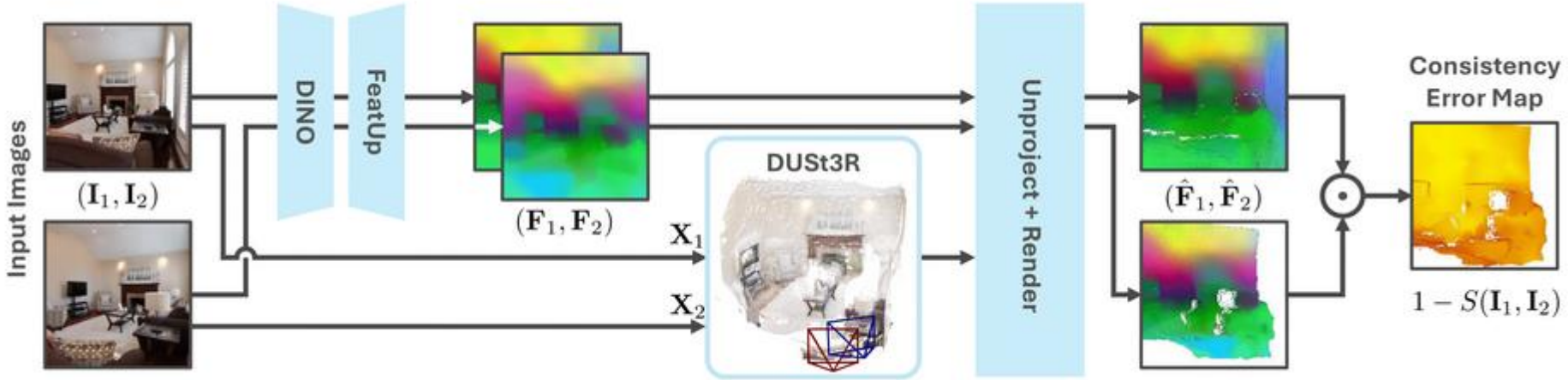



*[2] MEt3R: Measuring Multi-View Consistency in Generated Images (Asim et al., MPI-INF & ETH Zurich)*

# Related Work

## Met3r

- evaluates consistency between generated multi-view images
- Uses DUSt3R to perform dense 3D reconstruction from image pairs
- One view is warped into the other using the 3D reconstruction
- Feature maps from original and warped views are compared to compute a similarity score

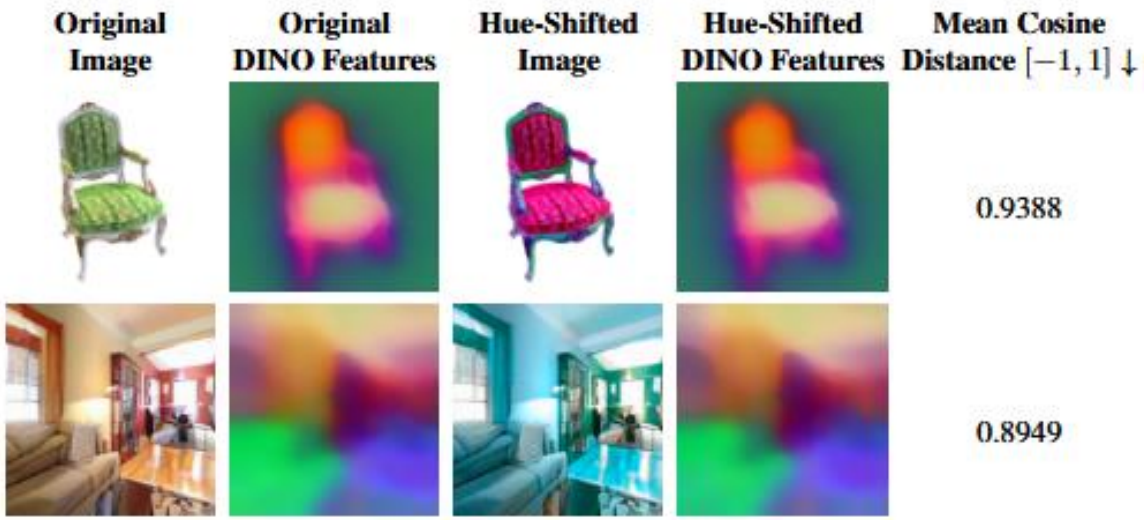



*[2] MEt3R: Measuring Multi-View Consistency in Generated Images (Asim et al., MPI-INF & ETH Zurich)*

# Methodology

**DINO Features – Texture Insensitivity**

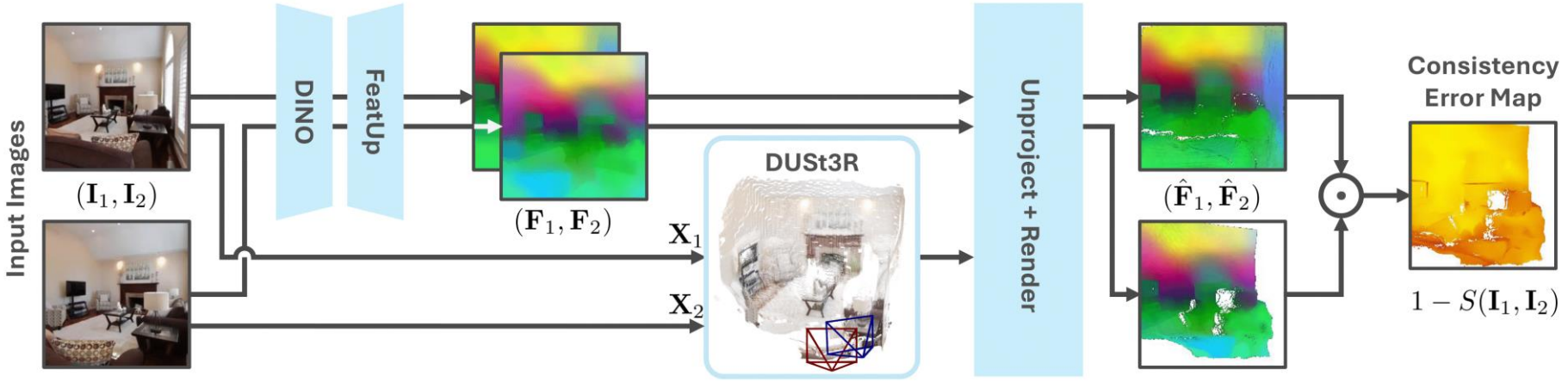
- DINO features capture **geometry and structure**, **not texture**
- Global **hue shifts** applied to images cause minimal change in DINO embeddings
- Cosine similarity remains high → indicates **color and texture invariance**
- Implies limited suitability for evaluating **appearance consistency**



# Methodology

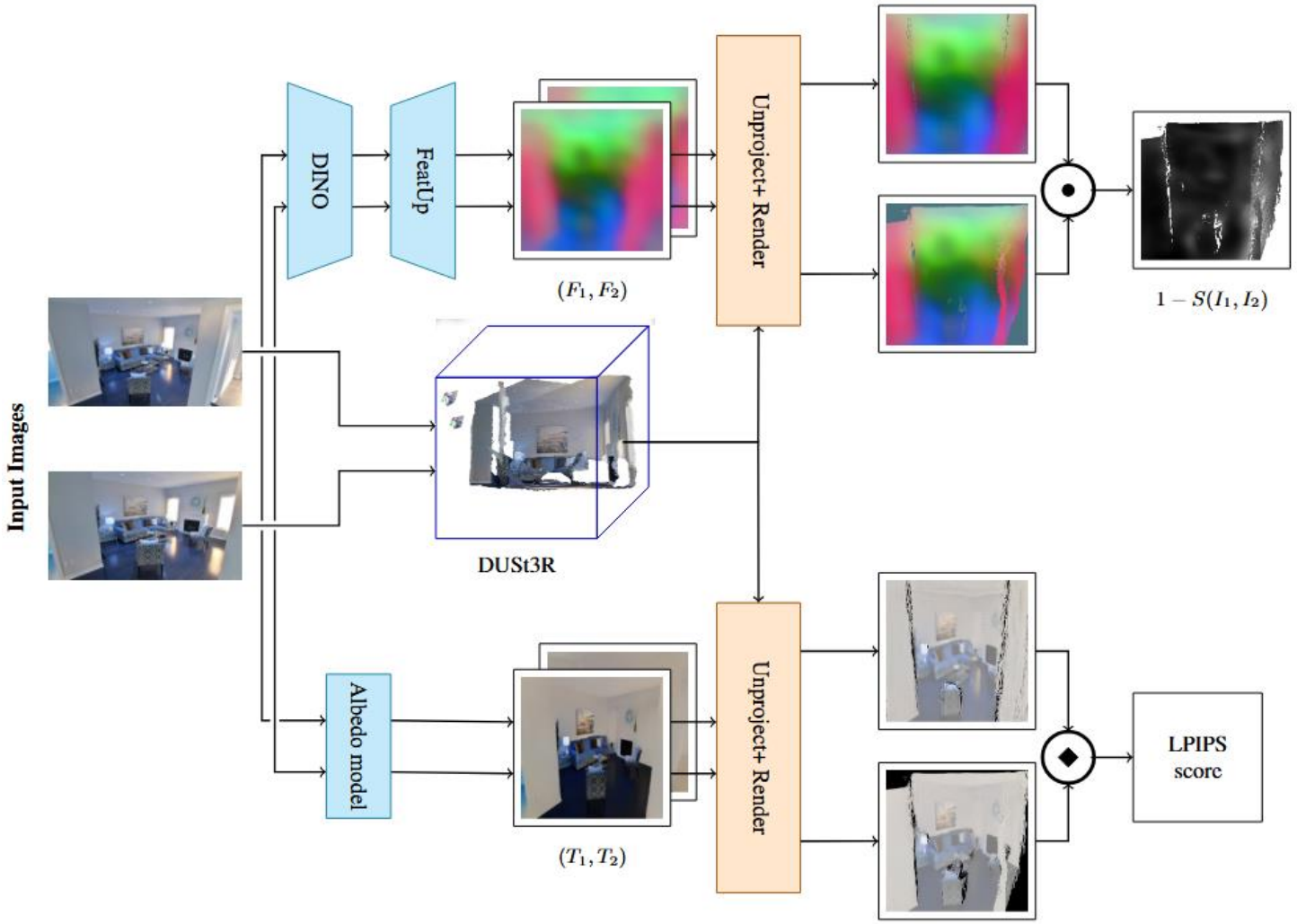
## Limitations of MEt3R

- MEt3R relies on DINO features → biased toward **geometric alignment**
- **Insensitive to texture and color mismatches** across views
- Cannot detect appearance inconsistencies in multi-view generation
- Motivates the need for **dual-stream evaluation** (geometry + texture)




*[2] MEt3R: Measuring Multi-View Consistency in Generated Images (Asim et al., MPI-INF & ETH Zurich)*

# Methodology




**Met3er – Tex Architecture**

# Evaluation


| | | | | | | |
|---|---|---|---|---|---|---|
| **MEt3R ↓** | 0.4024 | 0.0994 | 0.3952 | 0.3798 | 0.2291 | 0.7384 |
| **Texture ↓** | 0.7486 | 0.3302 | 0.6710 | 0.6452 | 0.3422 | 0.7338 |

# Evaluation


| | | | | | |
|---|---|---|---|---|---|
| **MEt3R ↓** | 0.4024 | 0.0994 | 0.3952 | .2291 | 0.7384 |
| **Texture ↓** | 0.7486 | 0.3302 | .710 | .3422 | 0.7338 |

# Evaluation


| | | | | | |
|---|---|---|---|---|---|
| MEt3R ↓ | 0.4024 | 0.09 | 0.3798 | 0.2291 | 0.7384 |
| Texture ↓ | 0.7486 | 0.33 | 0.6432 | 0.3422 | 0.7338 |

# Conclusion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Proposed an **enhanced evaluation framework** for multi-view image generation
- **Extended MEt3R** with a **texture consistency stream**
- Showed that **DINO features** capture geometry but **miss texture discrepancies**
- Introduced **intrinsic image decomposition** to isolate **view-independent albedo**
- Enables more robust assessment of texture across views
- **Limitations**:
  - Pairwise only
  - Sensitive to rendering artifacts
  - Depends on decomposition quality