MVSFormer++: Revealing the Devil in transformer's details for multi-view stereo

PNU CSE AI Department

오지현

Introduction

- Multi-View Stereo (MVS)란?
  - 여러 시점의 이미지를 통해 2D->3D 구조로 복원하는 것

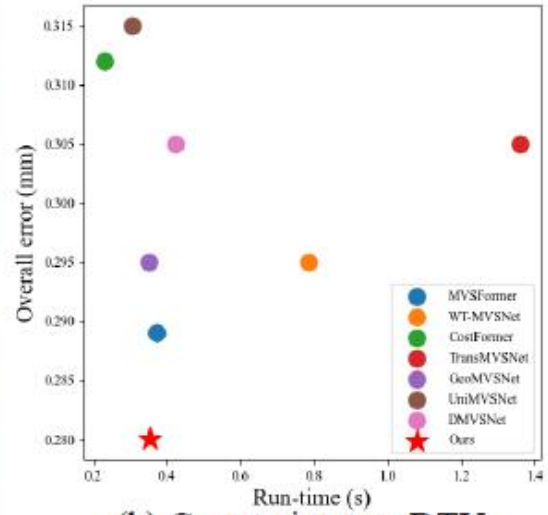
Introduction

- 이미지 복원에서 Transformer를 쓰는 이유
  - 긴 거리의 관계 정보를 잘 포착함 (self-/cross-attention)
- 기존 Transformer MVS의 한계
  - 다양한 모듈 (Feature encoder vs. Cost volume)에 동일한 Attention 사용
  - Cross-view 정보의 부족
  - 입력 크기 변화에 대한 일반화 어려움

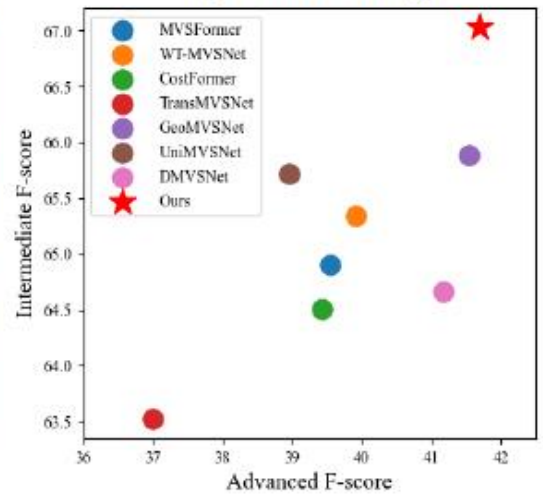
Contributions

- 서로 다른 MVS 모듈에 특화된 Attention 메커니즘 적용
- DINOv2 기반의 Pretrained ViT에 SVA (Side View Attention) 추가
- Cost Volume에는 CVT + FPE + AAS 사용
- 다중 해상도 대응 및 더 나은 generalization
- SOTA 성능 달성

(a) Point cloud results between MVSFormer and MVSFormer++ on DTU and Tanks-and-Temples.

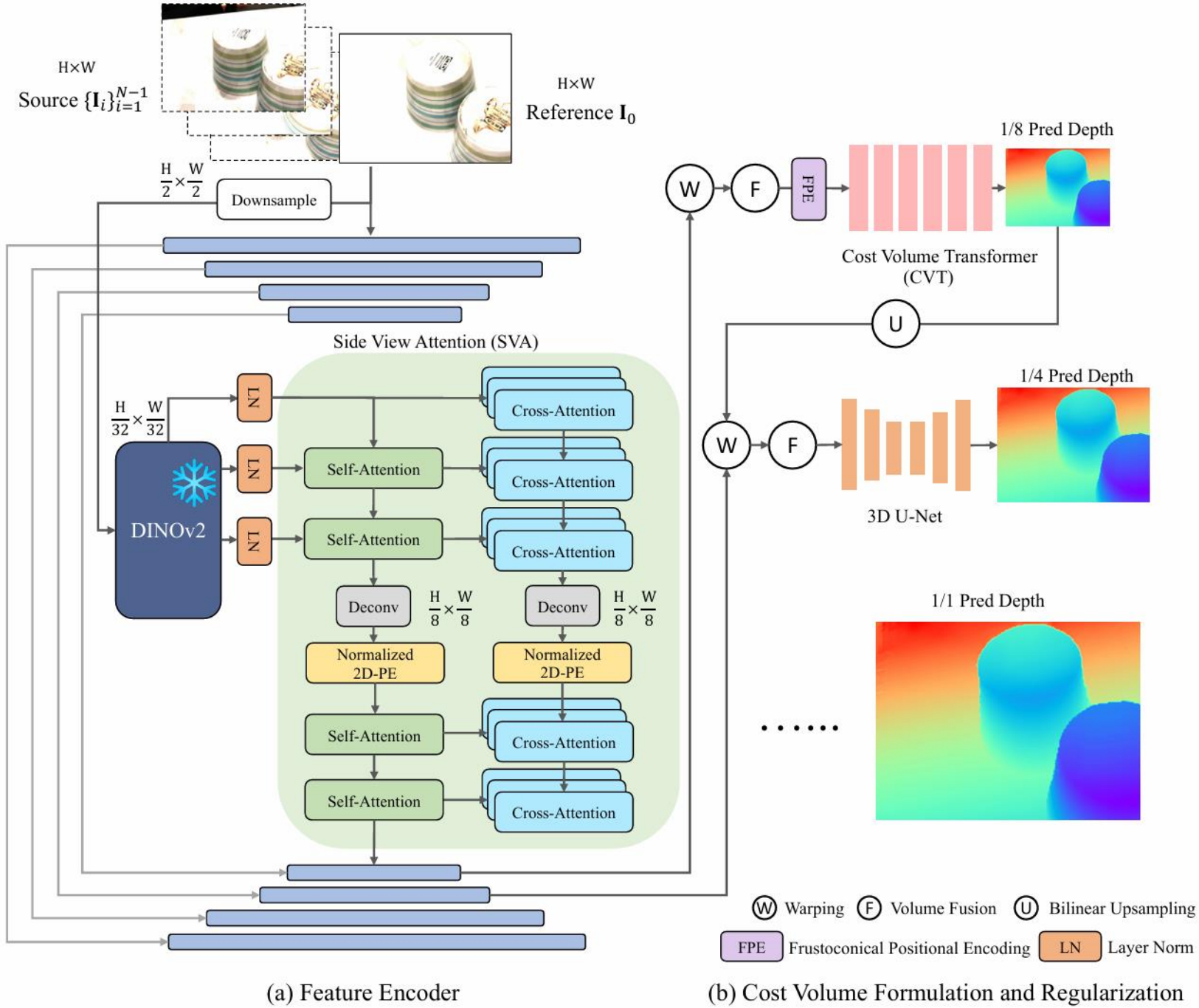


(b) Comparison on DTU (Overall error↓)



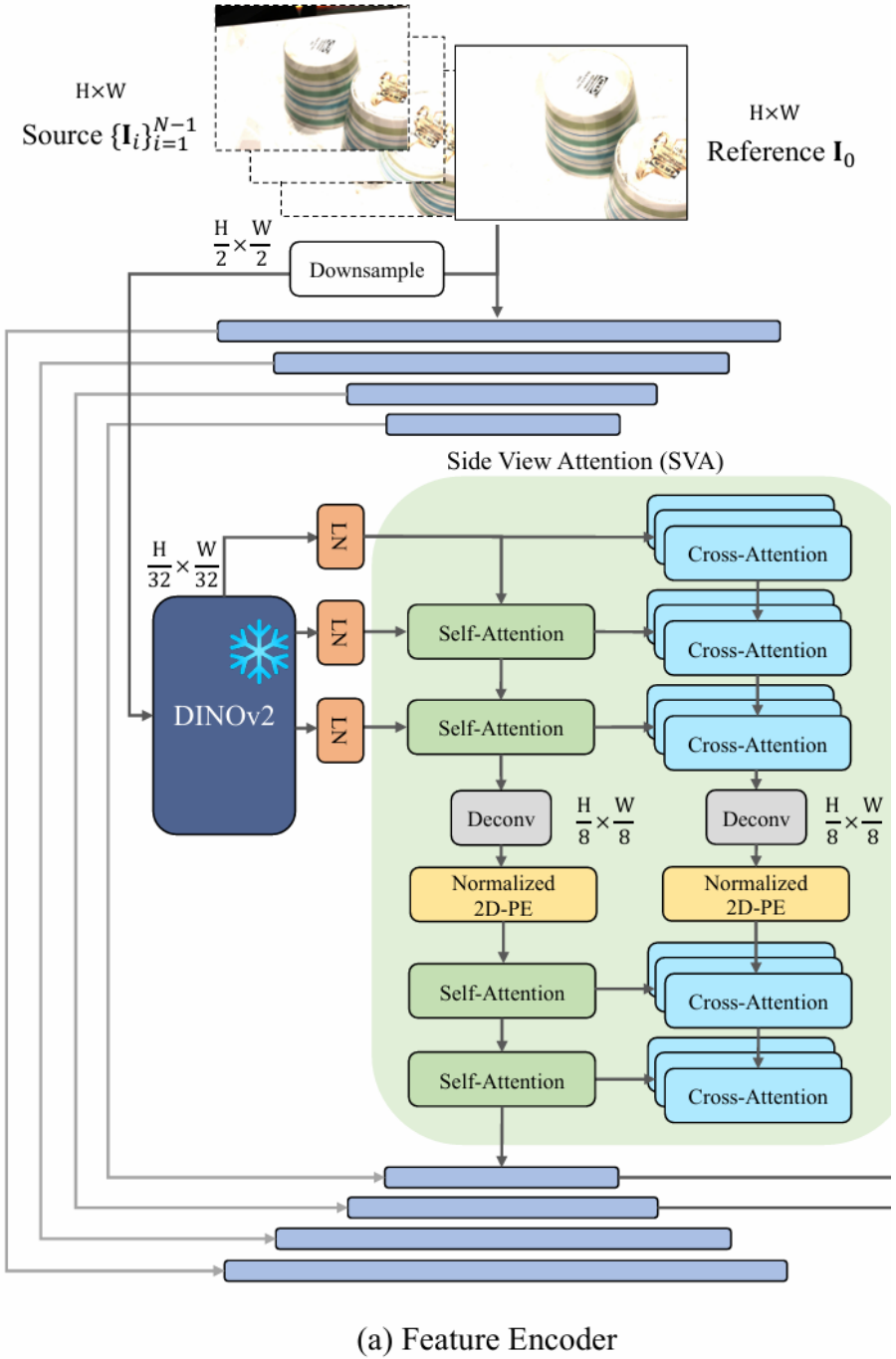
(c) Comparison on Tanks-and-Temples (F-score↑)

Figure 1: (a) Point cloud results compared between MVSFormer (Cao et al., 2022) and the proposed MVSFormer++ on DTU and Tanks-and-Temples. Results of state-of-the-art MVS methods on (b) DTU and (c) Tanks-and-Temples benchmark.



(a) Feature Encoder

(b) Cost Volume Formulation and Regularization



(a) Feature Encoder

(a) Feature Encoder

- **입력 (Input)**
  - Reference Image (Io): 기준 뷰 이미지
  - Source Images {Ii}: 다른 뷰에서 촬영된 이미지들
    - > 이들을 이용해 3D 장면의 depth를 예측.

- **Backbone Feature Extractor**
  - 일반적인 CNN 기반 pyramid 구조로 여러 해상도에서 feature를 추출
  - DINOv2를 통해 고수준의 전역 특징 추출을 수행

- **Side View Attention (SVA)**
  - Self-Attention과 Cross-Attention을 사용하여 이미지 간의 뷰 정보 통합 (source view 간의 관계)
  - Deconv (업샘플링) & positional encoding (위치 정보) 포함

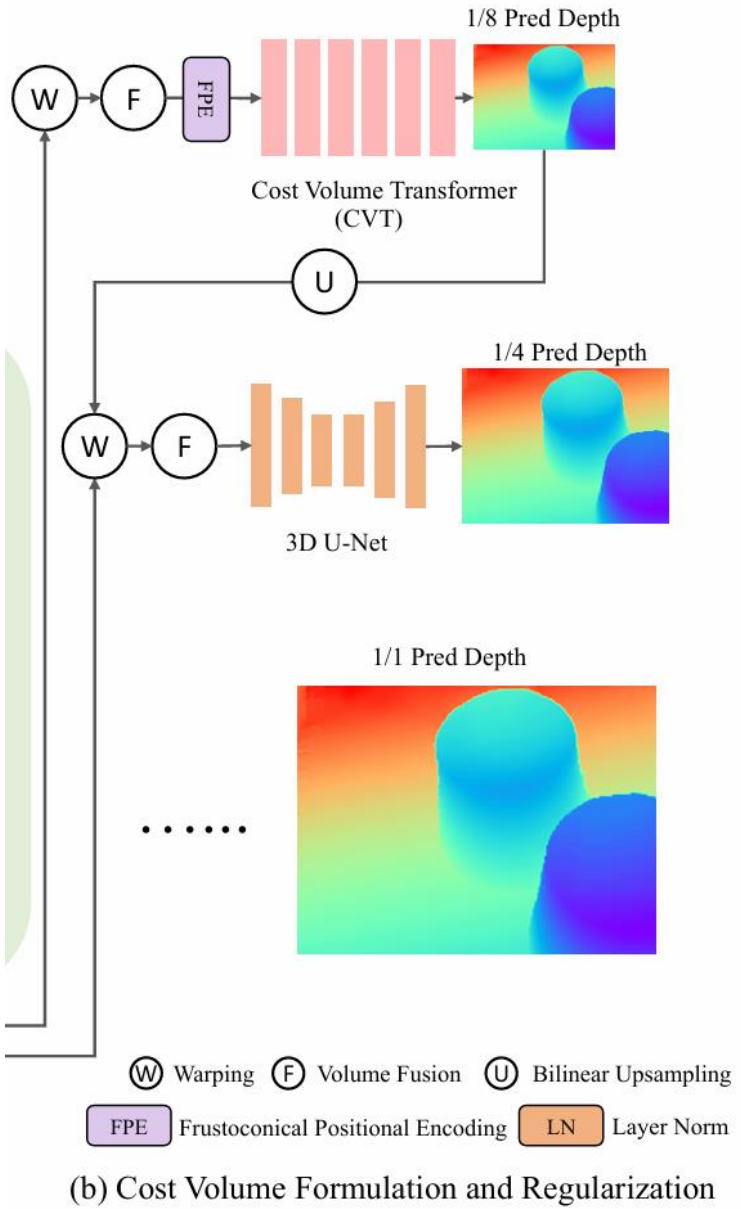
(b) Cost Volume Formulation and Regularization

- **Warping (W)**
  - Source image들의 feature들을 reference view에 정렬
  - Disparity 를 기반으로 정렬된 cost volume을 구성

- **Volume Fusion (F)**
  - 1/8 scale에서 깊이 예측을 transformer 기반의 블록으로 수행
  - FPE를 활용

- **3D U-Net**
  - ¼ scale에서의 깊이 예측을 CNN 구조인 U-Net으로 수행하여 정교하게 복원

- **Bilinear Upsampling (U)**
  - 최종적으로 고해상도 1/1 깊이 맵을 얻기 위해 업샘플링



(b) Cost Volume Formulation and Regularization

Key Components – Feature Encoder

- SVA: Pretrained DINOv2에 cross-view 정보를 주입
- Normalized 2D-PE, Adaptive Layer Scaling
- 효과: 고해상도 일반화에 강하고 수렴 안정성 향상

Key Components – Cost Volume Regularization

- CVT: 전통적인 3DCNN 대신 Transformer 기반 처리
- FPE: 3D 공간 정보 인코딩
- AAS: Attention dilution 문제 보완
- 정밀한 Depth map 추정 가능

Experiments

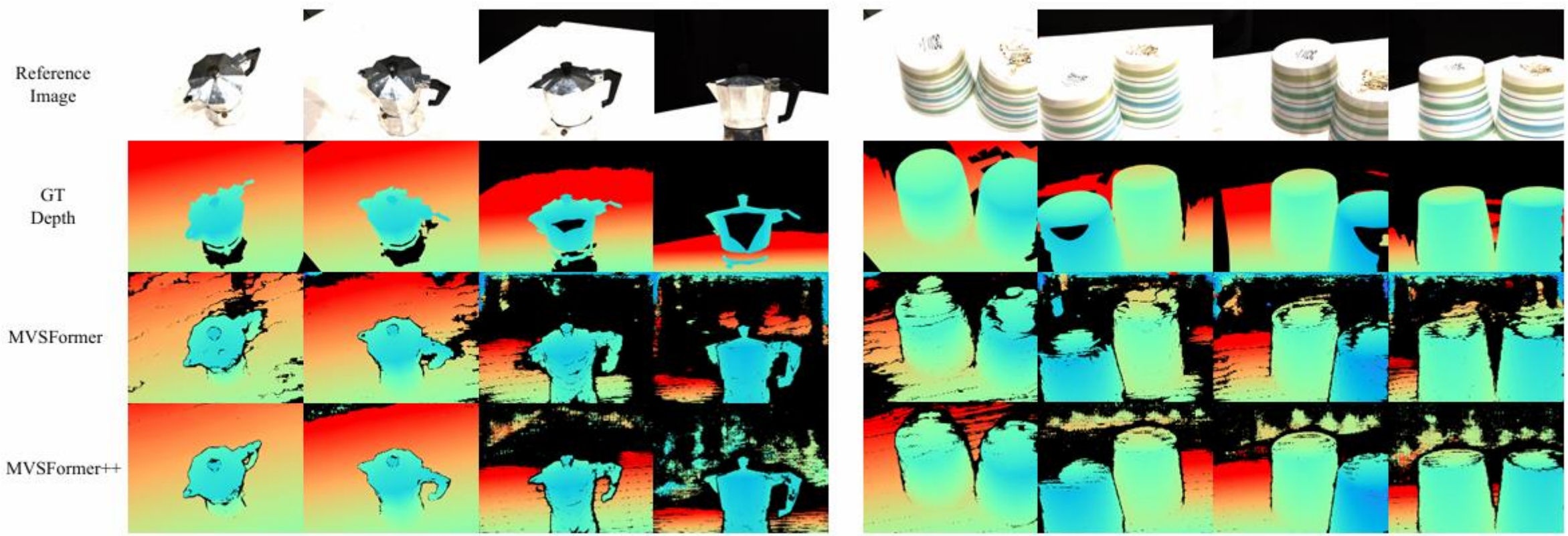


Figure 9: Qualitative depth comparisons on DTU between MVSFormer (Cao et al., 2022) and MVSFormer++.

Table 2: Quantitative point cloud results (mm) on DTU (lower is better). The best results are in bold, and the second ones are underlined. *All scenes share the same threshold for the post-processing.*

Methods	Accuracy↓	Completeness ↓	Overall↓
Gipuma (Galliani et al., 2015a)	**0.283**	0.873	0.578
COLMAP (Schönberger et al., 2016)	0.400	0.664	0.532
CasMVSNet (Gu et al., 2020)	0.325	0.385	0.355
AA-RMVSNet (Wei et al., 2021)	0.376	0.339	0.357
UniMVSNet (Peng et al., 2022)	0.352	0.278	0.315
TransMVSNet (Ding et al., 2022)	0.321	0.289	0.305
WT-MVSNet (Liao et al., 2022)	0.309	0.281	0.295
CostFormer (Chen et al., 2023)	<u>0.301</u>	0.322	0.312
RA-MVSNet (Zhang et al., 2023b)	0.326	0.268	0.297
GeoMVSNet (Zhang et al., 2023c)	0.331	0.259	0.295
MVSFormer (Cao et al., 2022)	0.327	**0.251**	<u>0.289</u>
MVSFormer++ (ours)	0.3090	<u>0.2521</u>	**0.2805**

Figure 5: Qualitative results compared with state-of-the-art models on scan77 in DTU.

Conclusion

- Transformer 기반 MVS 설계의 세부
- 각 모듈에 특화된 설계를 도입하여 성능을 향상할 수 있다.