ICG-MVSNet: Learning Intra-view and Cross-view Relationships for Guidance in Multi-View Stereo

Yuxi Hu[1] Jun Zhang[1] Zhe Zhang[2] Rafael Weilharter[1]
Yuchen Rao[1] Kuangyi Chen[1] Runze Yuan[1] Friedrich Fraundorfer[1]

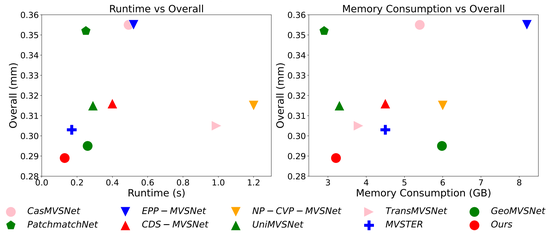
[1]Institute of Visual Computing, Graz University of Technology, Austria

[2]School of Electronic and Computer Engineering, Peking University, China

Contributions

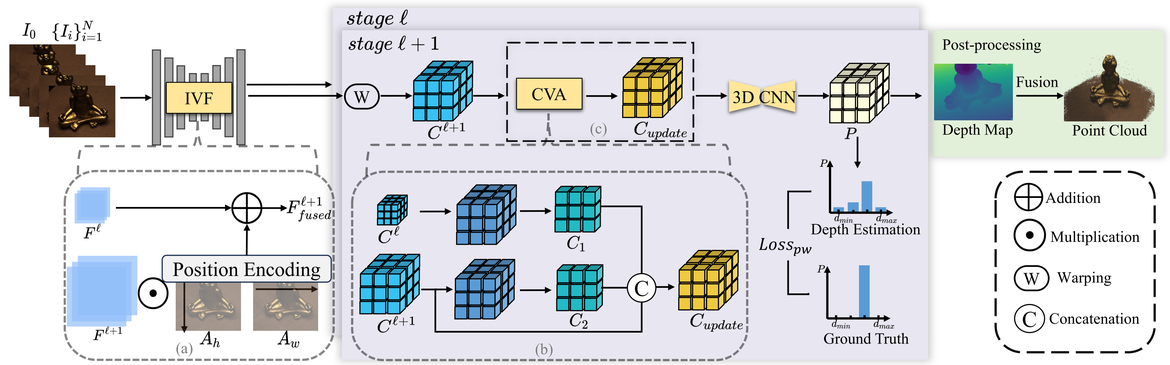
Multi-View Stereo (MVS) aims to estimate depth and reconstruct 3D geometry from multiple images captured at different viewpoints. Our method focuses on improving feature fusion and aggregation for better accuracy and efficiency.

- We propose a learning-based MVS with an **Intra-View Fusion (IVF)**, which embeds positional information along two coordinate directions into feature maps within a single image.
- We introduce **Cross-View Aggregation (CVA)**, a lightweight scheme that efficiently leverages the prior from previous correlations.
- Our **ICG-MVSNet** achieves competitive performance with an optimal balance between effectiveness and computational efficiency.



Comparison with state-of-the-art methods in runtime and GPU consumption on DTU.

Pipeline



The overall architecture. Our method is a coarse-to-fine framework that estimates depths from low resolution ($stage\ \ell$) to high resolution ($stage\ \ell+1$), where $\ell = 0, 1, 2$, resulting in a total of 4 stages.

1. Extract multi-scale features $\{\boldsymbol{F}_i\}_{i=0}^{N}$ from images using a feature pyramid network enhanced by **Intra-View Fusion (IVF, (a))**, which encodes the coordinate information within each view.
2. For each source image, warp its feature map $\{\boldsymbol{F}_i\}_{i=1}^{N}$ into the reference camera frame I_0 using D depth hypotheses, and compute pair-wise correlations to measure cross-view similarity.
3. Aggregate multi-view correlations into cost volume $\boldsymbol{C}$ and update it using proposed **Cross-View Aggregation (CVA, (b)(c))**, which injects prior matching information and contextual cues.
4. Regularize cost volumes using a 3D CNN to obtain probability volumes and select depth hypotheses.
5. Fuse multi-view depth maps into 3D point clouds in a non-learnable manner.

Quantitative comparison

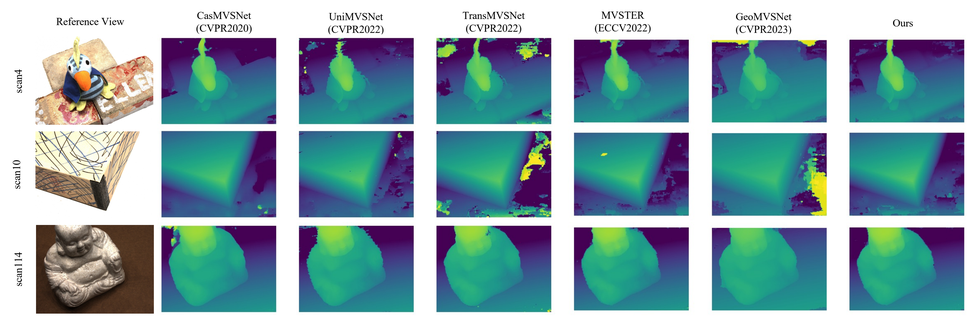
	Method	Acc.↓ (*mm*)	Comp.↓ (*mm*)	**Overall↓(mm)**	**Time↓(s)**	**GPU↓(GB)**
learning-based	Gipuma	0.283	0.873	0.578	-	-
	COLMAP	0.400	0.664	0.532	-	-
	R-MVSNet	0.383	0.452	0.417	-	-
	CasMVSNet	0.325	0.385	0.355	0.49	5.4
	CVP-MVSNet	0.296	0.406	0.351	-	-
	PatchmatchNet	0.427	0.277	0.352	0.25	2.9
	EPP-MVSNet	0.413	0.296	0.355	0.52	8.2
	CDS-MVSNet	0.352	0.280	0.316	0.40	4.5
	NP-CVP-MVSNet	0.356	0.275	0.315	1.20	6.0
	UniMVSNet	0.352	0.278	0.315	0.29	3.3
	TransMVSNet	0.321	0.289	0.305	0.99	3.8
	MVSTER*	0.340	0.266	0.303	0.17	4.5
	GeoMVSNet	0.331	0.259	0.295	0.26	5.9
	Ours	0.327	0.251	0.289	0.13	3.2

Quantitative comparison on DTU. $*$ means MVSTER is trained on full-resolution images. The colors indicate rankings, with red representing the top position, orange indicating second place, and yellow marking third.

Method	ade ↓	tde(1) ↓	tde(2) ↓	tde(4) ↓	tde(8) ↓	tde(16) ↓
MVSNet	14.7356	28.22	20.01	16.19	14.00	12.18
CasMVSNet	8.4086	25.41	17.80	14.06	11.42	8.97
CVP-MVSNet	6.9875	26.75	18.80	14.10	10.22	6.75
TransMVSNet	15.1320	26.09	18.03	15.17	13.18	11.57
MVSTER	9.3494	26.94	19.36	15.84	13.41	11.03
UniMVSNet	11.5872	28.90	19.66	13.73	11.02	8.92
GeoMVSNet	11.6586	24.06	16.53	13.28	11.08	9.21
Ours	6.6450	24.38	16.81	13.20	10.55	8.05

Depth map errors on DTU. The ade represents the average absolute depth error (mm), while $tde(X)$ indicates the percentage of pixels with an error above X mm.

Depth Map



Qualitative comparison with other state-of-the-art methods on DTU.

The depth map estimated by our method has a more complete and continuous surface and also has clearer outlines at the edges.

Key Formulations

Intra-View Fusion (IVF, see (a) in Pipeline)
Goal: Encode long-range dependencies and feature channel relationships with coordinate information to enhance feature expressiveness.

$$\boldsymbol{F}_{fused}^{\ell+1} = \boldsymbol{F}_{\uparrow}^{\ell} \oplus \boldsymbol{F}_{A}^{\ell+1}\ , \tag{1}$$

$$\boldsymbol{F}_{A}^{\ell+1} = \boldsymbol{A}_h \odot \boldsymbol{A}_w \odot \boldsymbol{F}^{\ell+1}\ , \tag{2}$$

Cross-View Aggregation (CVA, see (b)(c) in Pipeline)
Goal: Aggregate cross-view correlations to capture contextual relationships across stages, depth assumptions, and feature channels.

$$\boldsymbol{C}_1 = Conv2D(\boldsymbol{C}^{\ell}), \boldsymbol{C}_2 = Conv2D(\boldsymbol{C}^{\ell+1}) \tag{3}$$

$$\boldsymbol{C}_{update} = Concat(\boldsymbol{C}, \boldsymbol{C}_1, \boldsymbol{C}_2) \tag{4}$$

Point Cloud

Reconstructed point clouds on DTU.