GC-MVSNet: Multi-View, Multi-Scale, Geometrically-Consistent Multi-View Stereo

Vibhas K. Vats [1] Sripad Joshi [1] David J. Crandall [1] Md. Alimoor Reza [2] Soon-Heung Jung [3]

[1]Indiana University, Bloomington, Indiana [2]Drake University, Des Moines, Iowa [3]Electronics and Telecommunications Research Institute, Korea

Multi-View Stereo Problems

- Depth map-based MVS algorithms estimate the reference view depth maps using multiple RGB inputs (Reference + Source views)
- A consistent scene requires geometric consistency of depth estimates across multiple views

Two broader approaches are undertaken to ensure geometric consistency in estimated depth maps:

- Repeated application of geometric constraints during the depth estimation process → Traditional MVS Algorithms
- Geometric constraints applied as a post-processing step → Learning-based MVS Algorithms

GC-MVSNet is a learning-based algorithm with geometric constraints applied during the learning process.

Learning-Based MVS Algorithms

A learning-based MVS method:

- Extract multi-level features using CNNs
- Creates a matching 3D cost volume using features
- Regularize cost volume using 3D-CNN
- Filter geometrically consistent points to generate 3D point-cloud

They only use Geometric Constraints as a post-processing step for filtering multi-view consistent points. It leads to:

- Limited geometric cues during the learning process
- Require more training iterations to learn to reason about geometry

Hypothesis

GC-MVSNet:

- Explicitly models cross-view geometric constraints during learning
- It penalizes geometrically inconsistent estimates during learning

With such explicit geometric constraint modeling, GC-MVSNet should:

- Develop a better understanding of multi-view geometry → Improved quantitative results
- Learn quickly to reason about scene geometry → Require less training iterations

Forward-Backward-Reprojection

Inputs: $D_0, c_0, D_i^{gt}, c_i^{gt}$
Output: $D''_{P''_0}, P''_0$

$K_R, E_R \leftarrow c_0; K_S, E_S \leftarrow c_i^{gt}$

$D_{(R\rightarrow S)} \leftarrow K_S \cdot E_S \cdot E_R^{-1} \cdot K_R^{-1} \cdot D_0$ ▷ Project

$X_{D_{(R\rightarrow S)}}, Y_{D_{(R\rightarrow S)}} \leftarrow D_{(R\rightarrow S)}$

$D_{S_{remap}} \leftarrow REMAP(D_i^{gt}, X_{D_{(R\rightarrow S)}}, Y_{D_{(R\rightarrow S)}})$ ▷ Remap

$D''_{P''_0} \leftarrow K_R \cdot E_R \cdot E_S^{-1} \cdot K_S^{-1} \cdot D_{S_{remap}}$ ▷ Back project

$P''_0 \leftarrow (X_{D''_{P''_0}}, Y_{D''_{P''_0}})$

Other Modifications

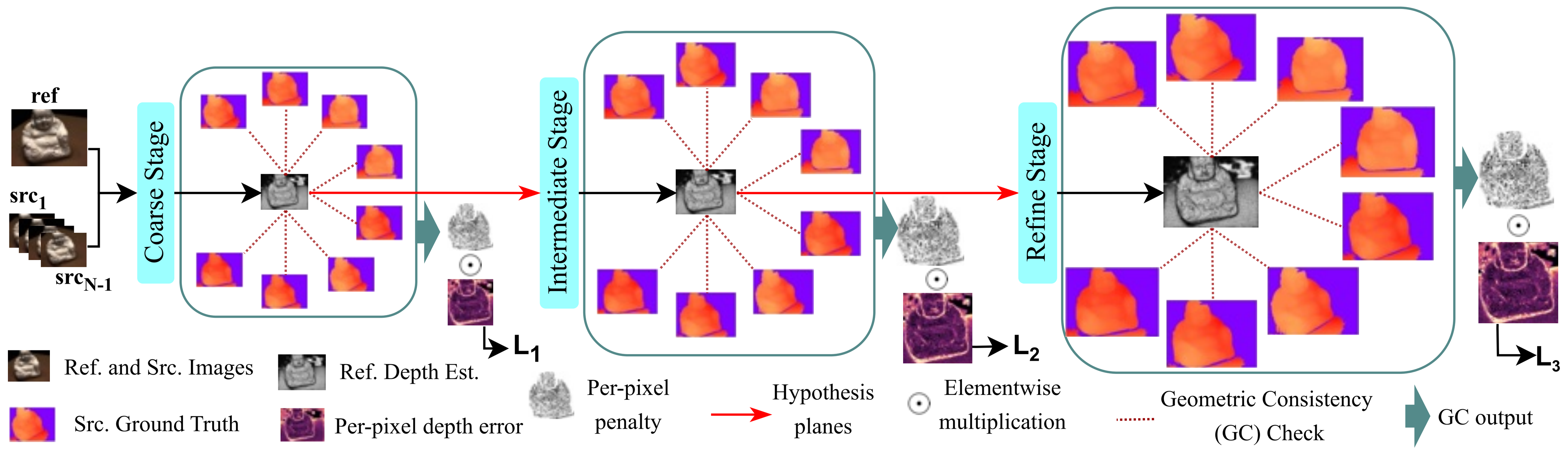
Two additional modifications were to stabilize the model's performance.

- Keeping the feature-extraction-network as Feature Pyramid Network, replaced the regular conv-layers with deformable conv-layers
- Replaced BatchNorm-layers with GroupNorm-layers as BatchNorm is not well suited for small batch-size

Method

Geometric-Consistency (GC) Module:

- Applied at the end of each stage to check cross-view consistency of the reference view depth maps
- Generates penalty for geometrically inconsistent estimates for each stage



Geometric-Consistency Module

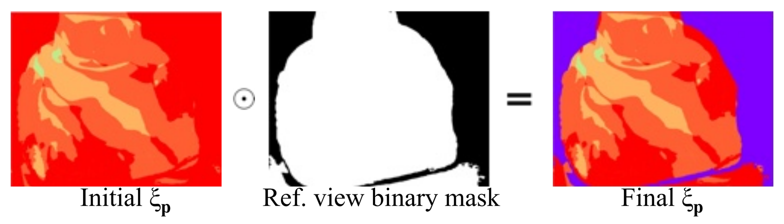
Complete GC-Algorithm

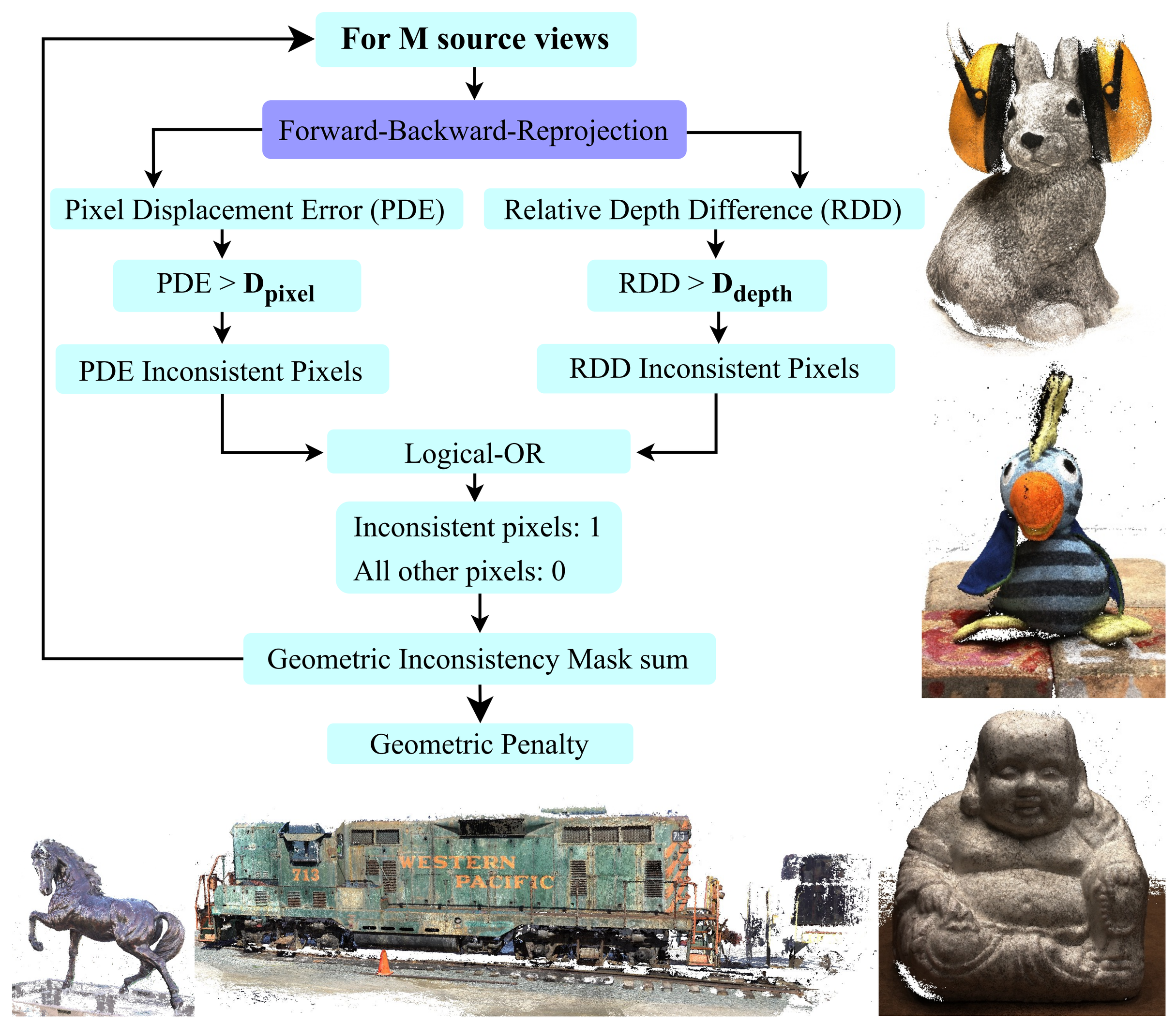
Initialize Mask-Sum → 0

For each Src. depth map:

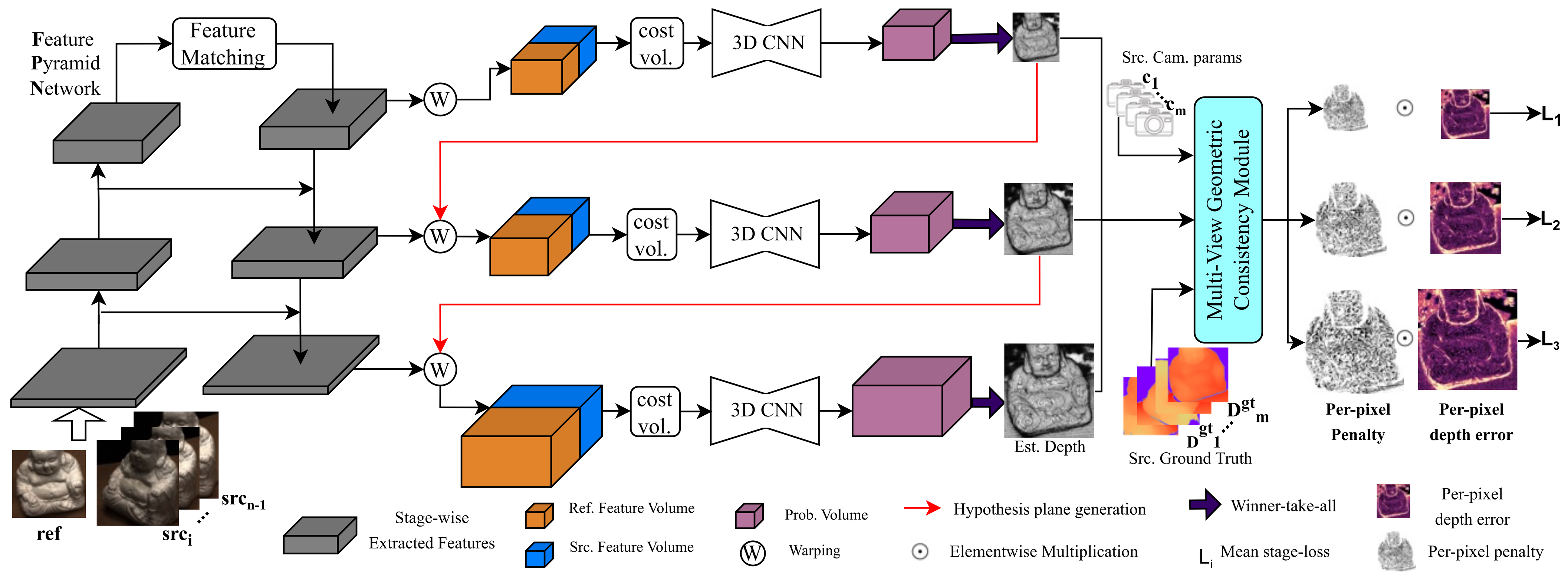
1. forward-backward-reprojection to get PDE and RDD
   - PDE $\leftarrow ||P_0 - P''_0||_2$
   - RDD $\leftarrow {}^1/_{D_0}||D''_{P''_0} - D_0||_1$
2. Select geometrically inconsistent pixels
   - $PDE_{mask} > D_{pixel}$
   - $RDD_{mask} > D_{depth}$
3. Combine inconsistent pixels from both masks
   - Logical-OR (PDE_{mask}, RDD_{mask})
4. Current-Mask ← Assign penalty to each pixel
   - Inconsistent pixels → 1
   - All other pixels → 0
5. Add Current-Mask to initial Mask-Sum

Geometric penalty (ξ_p) ← average Mask-Sum
Apply reference view binary mask to generate final ξ_p





GC-MVSNet Architecture



Quantitative Result on DTU Dataset

	Method	Acc ↓	Comp ↓	Overall ↓
Traditional	Furu [9]	0.613	0.941	0.777
	Tola [36]	0.342	1.190	0.766
	Gipuma [10]	**0.283**	0.873	0.578
	COLMAP [33]	0.400	0.664	0.532
Learning-based	SurfaceNet [16]	0.450	1.040	0.745
	MVSNet [48]	0.396	0.527	0.462
	P-MVSNet [25]	0.406	0.434	0.420
	R-MVSNet [49]	0.383	0.452	0.417
	Point-MVSNet [2]	0.342	0.411	0.376
	CasMVSNet [12]	0.325	0.385	0.355
	CVP-MVSNet [47]	<u>0.296</u>	0.406	0.351
	UCS-Net [3]	0.338	0.349	0.344
	AA-RMVSNet [41]	0.376	0.339	0.357
	UniMVSNet [30]	0.352	0.278	0.315
	TransMVSNet [6]	0.321	0.289	0.305
	GBi-Net* [28]	0.312	0.293	<u>0.303</u>
	MVSTER [39]	0.350	<u>0.276</u>	0.313
	GC-MVSNet (ours)	0.330	**0.260**	**0.295**
	GBi-Net [28]	0.315	0.262	0.289
	GC-MVSNet (ours)	0.323	**0.255**	0.289

Our method achieve State-of-the-art result on two datasets:

- DTU and BlendedMVS

GC: A Plug-in Module

GC module is designed as a plug-in module

- Plug into any depth map-based MVS method
- Retraining the network with GC-module provides:
  - Improved quantitative results to its previous performance
  - Require less training iterations to achieve optimal performance

We demonstrate this on two different methods:

- CasMVSNet and TransMVSNet

Methods	Loss	Other	GC	Overall↓	Epoch
CasMVSNet [2]	L_1	×	×	0.355	16
	L_1	✓	×	0.357	16
	L_1	×	✓	0.335	11
TransMVSNet [1]	FL	×	×	0.305	16
	FL	✓	×	0.322	16
	FL	×	✓	0.303	8

Table 1. GC-module as a plug-in module in TransMVSNet and CasMVSNet

References

[1] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8585–8594, 2022.

[2] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2495–2504, 2020.

Connect with us

Provide feedback:

Scan to visit our project page

Acknowledgement: This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. 23ZH1200, The research of the fundamental media contents technologies for hyper-realistic media space.