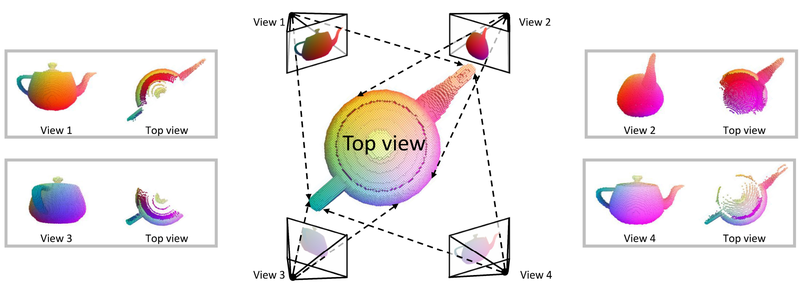
MVPNet: Multi-View Point Regression Networks for 3D Object Reconstruction from A Single Image

Jinglu Wang (MSRA) Bo Sun (PKU) Yan Lu (MSRA)

Motivation

Why multi-view representation? Existing representations for 3D reconstruction are mainly of four categories. While 3D volumetric grids suffer high computational complexity, unordered point sets require to solve point-wise mapping, meshes are difficult for CNNs to encode and decoder, multi-view based representation are convolution-favored, ordered and can depict dense and detailed surface.

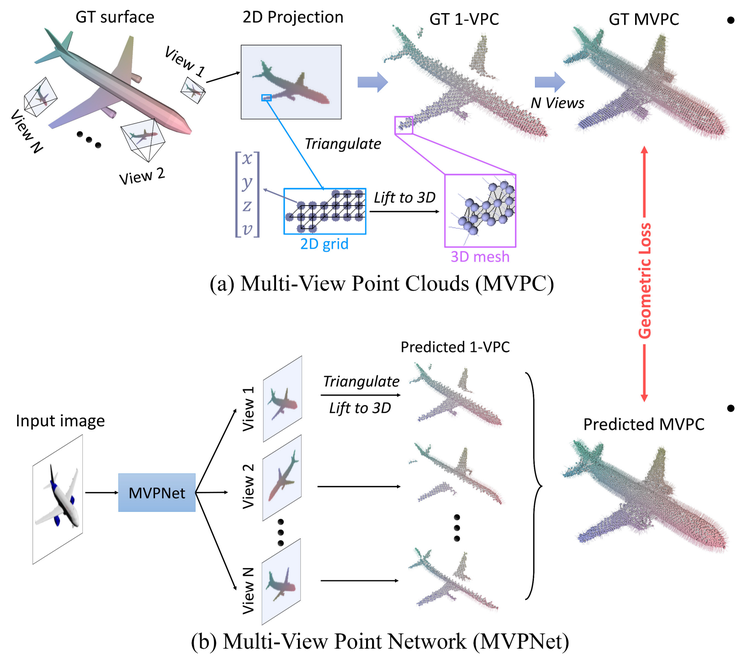
Overcome limitations of multi-view representation. View based representation handing features in 2D projective space can neglect information loss through dimension reduction from 3D to 2D. The proposed MVPC constructs meshes in 3D from 2D grids. MVPC allows us to discretize integrals of surface variations over the constructed triangular mesh and to enforce multi-view consistency with view correspondences.



Multi-View Point Clouds:

Overview

We reconstruct an object's surface from a single image by 1) representing a 3D surface as MVPC, 2) regressing MVPC with MVPNet, and 3) proposing a geometric loss to interpret discrepancy over 3D surfaces.



(a) Multi-View Point Clouds (MVPC)

(b) Multi-View Point Network (MVPNet)

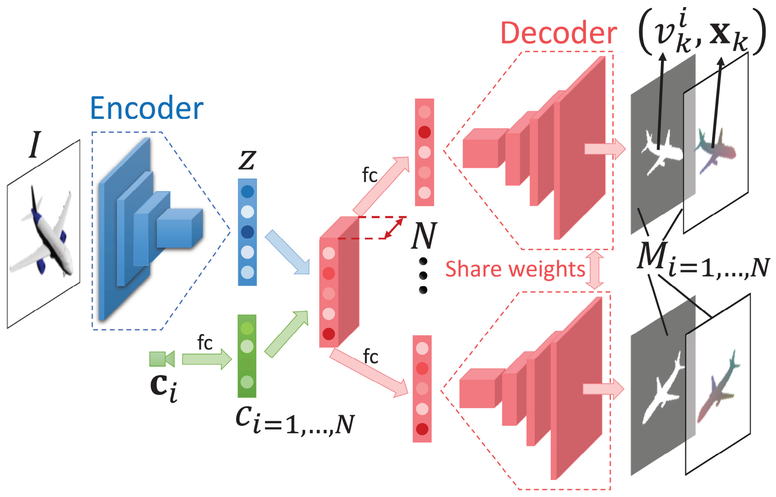
- **Multi-View Point Clouds:** A surface is represented by Multi-View Point Cloud (MVPC). Each pixel in a 1-VPC stores the backprojected surface point *(x, y, z)* from this pixel and its visibility *v*. The stored 3D points are triangulated according to the 2D grid on the image plane and their normals are shown to indicate surface orientation.
- **Muti-View Point Network.** Given an RGB image, the MVPNet generates a set of 1-VPCs and their union forms the predicted MVPC. The geometric loss measures discrepancy between predicted and ground truth MVPC.

Network Architecture

The MVPNet is an encoder-decoder generative network incorporating camera parameters to generate view-dependent point clouds.

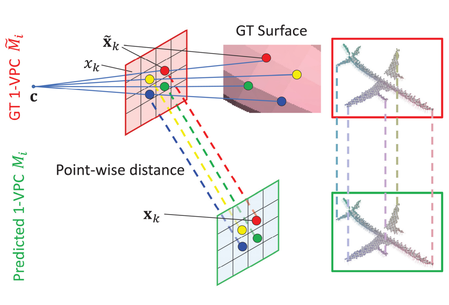
- The **encoder** maps an image *I* to an embedding space to obtain a feature *Z*. Each camera matrix $\boldsymbol{c}_i$ is first mapped to a feature c_i, serving as a view indicator, and then is concatenated with *Z* to get (Z, c_i).
- The **decoder** converting (Z, c_i) to a 1-VPC M_i indicated by c_i learns the projective transformation and space completion. The decoder shares weights among *N* branches.

The output MVPC is of shape *N* x *H* x *W* x 4. The last channel corresponds to a 3D coordinate $\boldsymbol{x}_k = (x_k, y_k, z_k)$ and visibility v^i_k of a point $\boldsymbol{x}_k$.



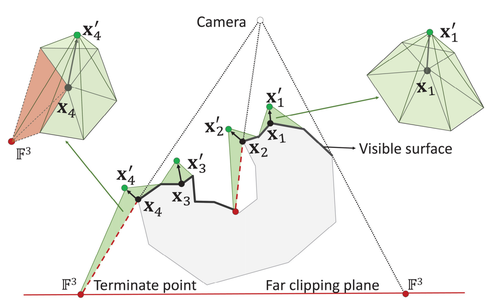
Geometric Loss

We propose a geometric loss (GeoLoss) that is able to capture variances over 3D surfaces rather than over sparse point sets or 2D projective planes. The GeoLoss is made up of three components: $\mathcal{L}_{Geo} = \mathcal{L}_{ptd} + \alpha\mathcal{L}_{vol} + \beta\mathcal{L}_{mv}$



Point-wise distance term. The points in ground truth and predicted 1-VPC have a one-to-one mapping, since 2D pixels with equal 2D coordinates are defined to store the same surface point induced by the same viewpoint. The sum of point-wise distances for ground truth and predicted 1-VPC is the L2 loss.

$$\mathcal{L}_{ptd} = \sum_i^N \sum_{x \in M_i} ||M_i(x) - \widetilde{M}_i(x)||_2$$

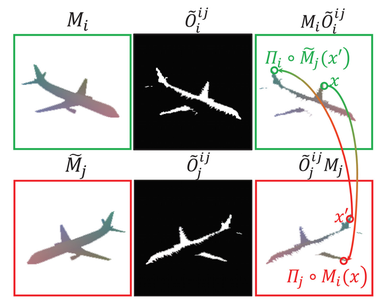


Quasi-volume term. Inspired by the volume-preserving constraints used in variational surface deformation, we propose a quasi-volume discrepancy metric to describe the surface discrepancy, characterizing details and handling occluding contours.

$$\mathcal{L}_{vol}(\mathcal{S}, \widetilde{\mathcal{S}}) = \int_{\mathcal{S}} (\mathbf{x} - \widetilde{\mathbf{x}}) \cdot \mathbf{n} d\mathbf{x}$$

It is discretized in the MVPC's mesh as:

$$\mathcal{L}_{vol} = \sum_i^N \sum_{x \in M_i} \widetilde{V}_i(x)(M_i(x) - \widetilde{M}_i(x)) \cdot \widetilde{\mathbf{N}}_i(x)$$

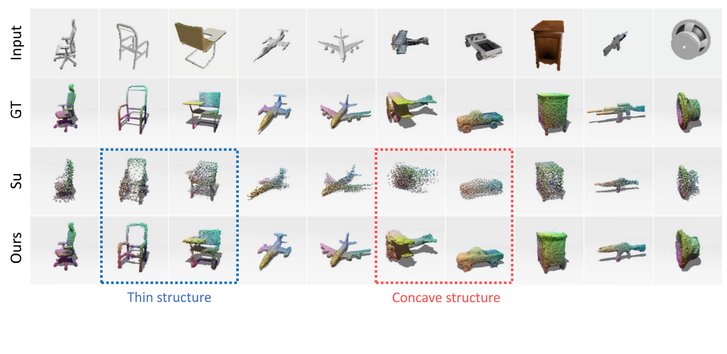


Multi-view consistency term. Some 1-VPCs may have overlap. Corresponding points should be close. We minimize sum of two distances between stored 3D points in two corresponding pixels and their reprojected pixels in another 1-VPC.

$$\mathcal{L}_{mv} = \sum_{i,j}(\sum_{x \in \widetilde{O}_i^{ij}} ||M_i(x) - \widetilde{M}_j(\Pi_j \circ M_i(x))||_2 + \sum_{x \in \widetilde{O}_j^{ij}} ||\widetilde{M}_j(x) - M_i(\Pi_i \circ \widetilde{M}_j(x))||_2)$$

Results

We present qualitative and quantitative results of the reconstruction, comparing the proposed method two collections of state-of-the-art methods according to result representations, i.e., point clouds and volumetric grids. Also, we show the generative representation of the learned features with interpolation, arithmetic, classification and clustering.



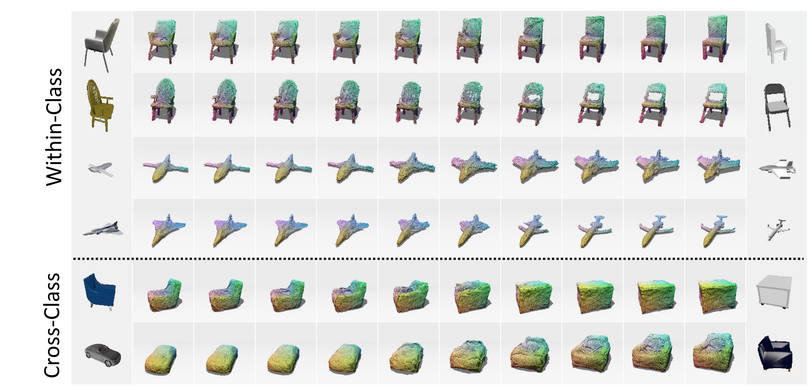
Qualitative comparison to point generation method [1].

		plane	bench	cabinet	car	chair	display	lamp	speaker	firearm	couch	table	phone	vessel	mean
voxel	R2N2(Choy et al. 2016)(1 view)	0.513	0.421	0.716	0.798	0.466	0.468	0.381	0.662	0.544	0.628	0.513	0.661	0.513	0.56
	R2N2(Choy et al. 2016)(5 views)	0.561	0.527	**0.772**	**0.836**	0.550	0.565	0.421	0.717	0.600	0.706	0.580	0.754	0.610	0.630
	PTN-Comb(Yan et al. 2016)	0.584	0.508	0.711	0.738	0.470	0.547	0.422	0.587	0.610	0.653	0.515	0.773	0.551	0.590
	CNN-Vol(Yan et al. 2016)	0.575	0.514	0.697	0.735	0.445	0.539	0.386	0.548	0.603	0.647	0.514	0.769	0.5445	0.578
point	Soltani(Soltani et al. 2017)	0.587	0.524	0.698	0.743	0.529	0.679	0.480	0.586	0.635	0.59	0.593	0.789	0.604	0.618
	Su(Fan, Su, and Guibas 2017)	0.601	0.55	0.771	0.831	0.544	0.552	0.462	**0.737**	0.604	**0.708**	0.606	0.749	0.611	0.640
Depth	GeoLoss(N=4)	0.655	0.578	0.664	0.709	0.546	0.653	0.486	0.573	0.676	0.630	0.561	0.783	0.633	0.627
	GeoLoss(N=6)	0.624	0.579	0.677	0.719	0.543	0.636	0.498	0.578	0.682	0.636	0.548	0.800	0.643	0.628
	GeoLoss(N=8)	0.622	0.576	0.691	0.724	0.540	0.643	0.501	0.590	0.684	0.647	0.534	0.788	0.640	0.629
MVPNet	PtLoss(N=6)	0.474	0.459	0.573	0.704	0.436	0.558	0.375	0.496	0.519	0.567	0.432	0.691	0.558	0.526
	GeoLoss(N=4)	0.666	0.622	0.693	0.786	0.616	0.653	0.510	0.599	**0.696**	0.690	0.635	0.811	**0.663**	0.665
	GeoLoss(N=6)	**0.678**	**0.623**	0.685	0.788	**0.627**	**0.681**	**0.523**	0.602	0.693	0.701	**0.652**	**0.814**	0.659	**0.671**
	GeoLoss(N=8)	0.667	0.610	0.686	0.782	0.609	0.667	0.507	0.596	0.688	0.686	0.641	0.809	0.661	0.662

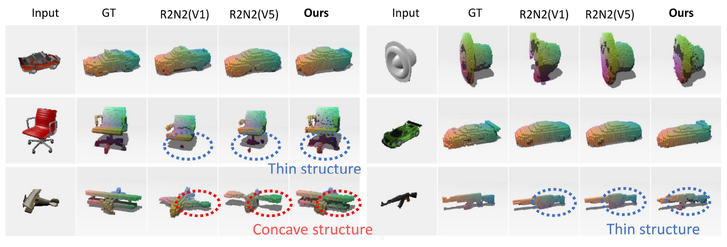
Quantitative comparison with voxel IoU.

	plane	bench	cabinet	car	chair	display	lamp	speaker	firearm	couch	table	phone	vessel	mean
Su(Fan, Su, and Guibas 2017)	1.395	1.899	2.454	1.927	2.121	2.127	2.280	3.000	1.337	2.688	2.052	1.753	2.064	2.084
Lin(Lin, Kong, and Lucey 2018)	1.418	1.622	1.443	1.254	1.964	1.640	3.547	2.039	1.400	1.670	1.655	1.569	1.682	1.761
Soltani(Soltani et al. 2017)	0.167	0.165	0.122	**0.026**	0.277	0.085	1.814	0.163	0.107	0.138	0.226	0.258	0.102	0.28
MVPNet(N=4)	0.045	0.084	0.063	0.042	0.086	0.065	0.561	0.163	0.104	0.082	0.070	0.046	0.060	0.113
MVPNet(N=6)	**0.041**	**0.079**	0.060	0.041	**0.085**	0.053	**0.421**	**0.152**	**0.093**	**0.070**	**0.069**	**0.038**	**0.050**	**0.096**
MVPNet(N=8)	0.044	0.085	**0.058**	0.040	0.103	**0.050**	0.494	0.153	0.113	0.083	0.075	0.039	0.059	0.107

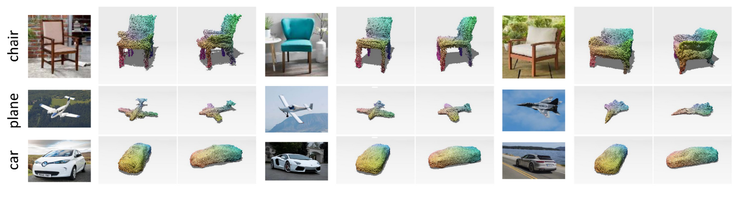
Quantitative comparison with chamfer distance metric.



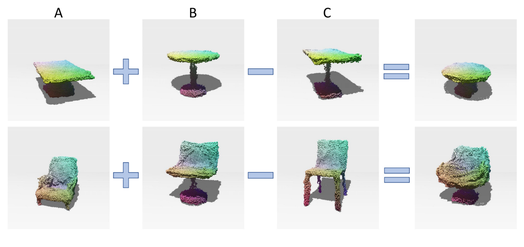
Interpolation. Reconstructions of linear interpolation of two features.



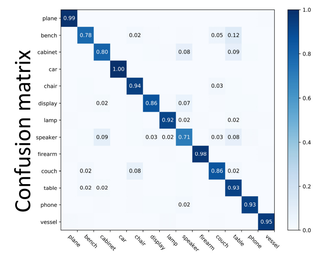
Qualitative compare to voxel-based method [2].



Reconstruction results on real word data.



Arithmetic. The shapes of last column are obtained by decoding the fature (A+B-C).



Classification.

Reference

[1] Fan, H.; Su, H.; and Guibas, L. 2017. A point set generation network for 3d object reconstruction from a single image. In ICCV.

[2] Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In ECCV.

Microsoft Research 微软亚洲研究院

PEKING UNIVERSITY 1898