UC San Diego Statistical Visual Computing Lab

Black-Box Test-Time Shape REFINEment for Single View 3D Reconstruction

Project Website & Dataset Download: http://www.svcl.ucsd.edu/projects/OOWL/CVPRW2022_REFINE.html

Brandon Leung, Chih-Hui Ho, Nuno Vasconcelos

CVPR JUNE 19-24 2022 L3D-IVU Workshop

Workshop on Learning with Limited Labelled Data for Image and Video Understanding

Introduction

- Recent progress has been made in reconstructing 3D object shape from images, i.e. single view 3D reconstruction.
- But due to the difficulty of collecting large datasets in the wild with 3D ground truth, it's still very challenging for generalize across domain, viewpoint, and class.
- To address this we propose REFINE, a postprocessing mesh refinement step easily integratable into the pipeline of any black-box method.
- At test time, REFINE optimizes a network per mesh instance, to encourage consistency between the mesh and the given object view. This helps restore details and improve reconstruction accuracy.
- A new hierarchical multiview, multidomain image dataset with 3D meshes called 3D-ODDS is also proposed as a uniquely challenging benchmark.

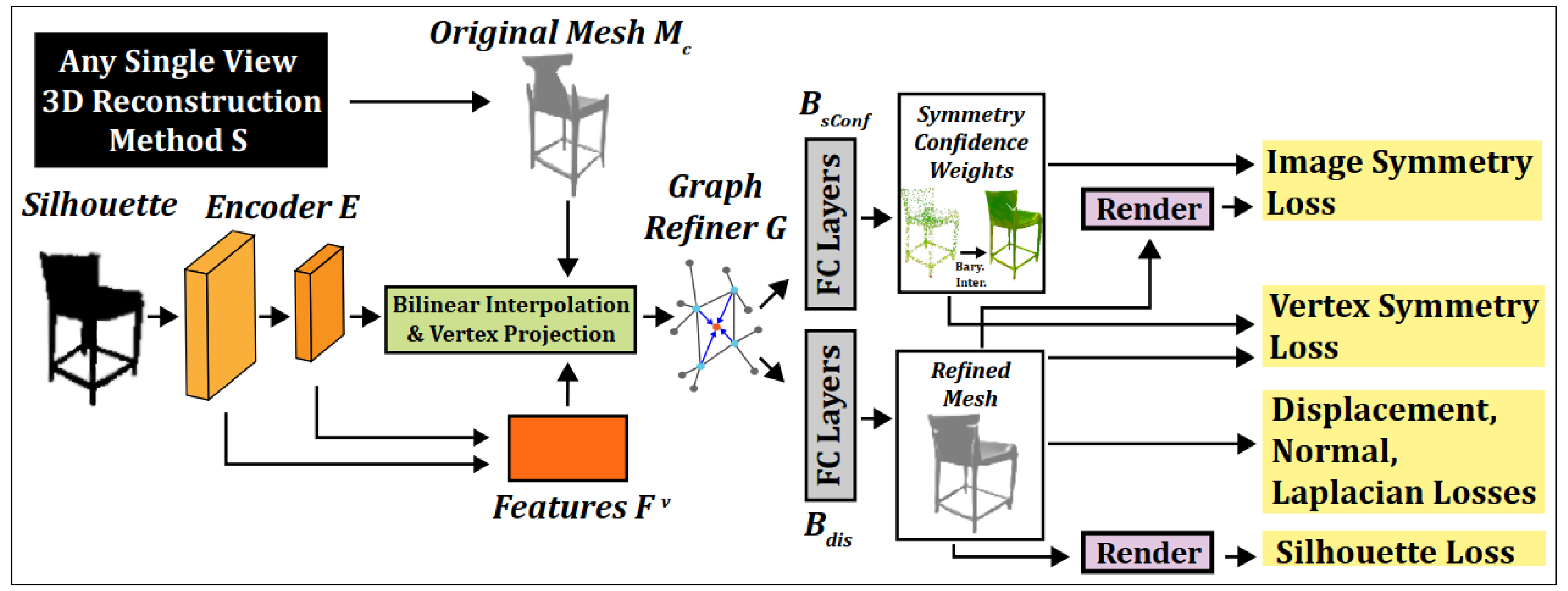
Previous Related Work

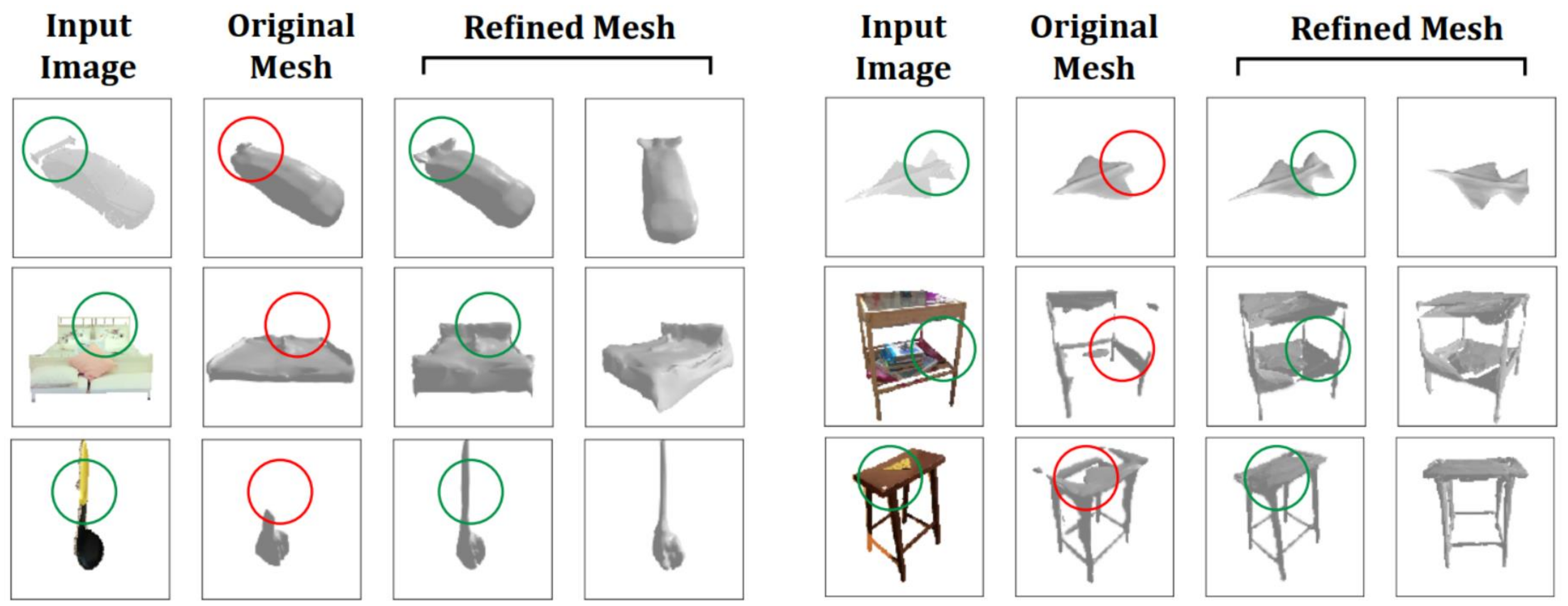
- Many single view 3D reconstruction methods exist; they can use varying 3D modalities (e.g. mesh, voxel, implicits) and use varying levels of supervision (e.g. supervised, weakly supervised, unsupervised).
- Regardless of method, important image details (circled in green) are frequently lost by state-of-the-art methods (circled in red).

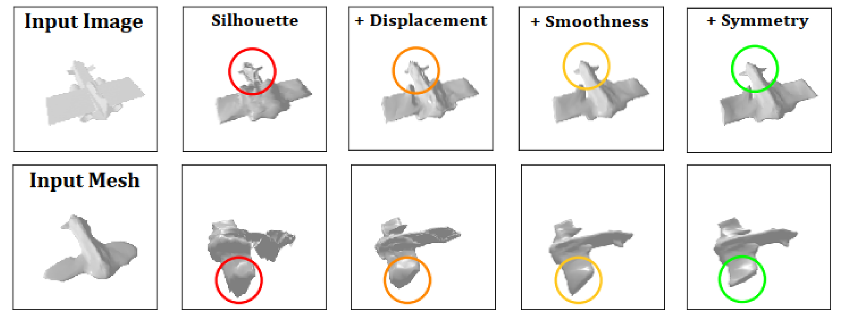
Kanazawa et al. | Li et al. | Mescheder et al. | Wang et al. | Xu et al.

Experiments & Findings

- REFINE neural network learns to displace vertices of reconstructions, by exploiting silhouette and viewpoint information.
- Main signal is leveraging rendered 3D mesh match input image silhouette.
- Symmetry, smoothness, and displacement based losses provide regularization to prevent degenerate solutions.
- Operates at test time, on-the-fly per mesh. Weights reinitialized & reoptimized per mesh, taking about 400 iterations to converge (about 30 seconds per mesh).



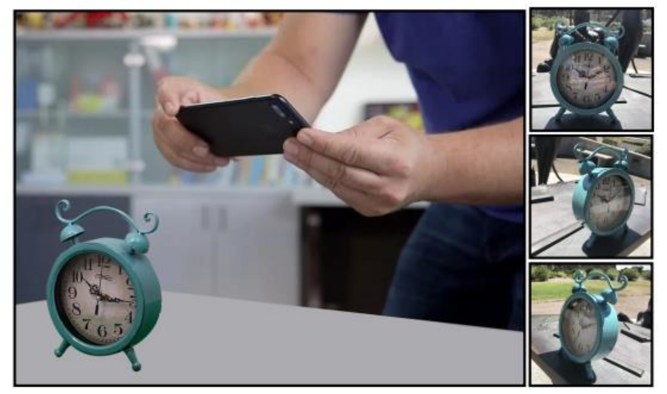




	EMD ↓	CD-l_2 ↓	F-Score ↑	Vol. IoU ↑
REFINEd OccNet [34]	4.3 → **3.3** (**-1.0**)	34.0 → **22.5** (**-11.5**)	80 → **84** (**+4**)	33 → **35** (**+2**)
REFINEd Pix2Mesh [51]	4.8 → **3.5** (**-1.3**)	38.0 → **23.1** (**-14.9**)	67 → **78** (**+11**)	22 → **27** (**+5**)
REFINEd AtlasNet [13]	6.2 → **4.9** (**-1.3**)	62.5 → **32.9** (**-29.6**)	56 → **72** (**+16**)	8 → **13** (**+5**)
REFINEd Pix2Vox [57]	4.5 → **3.3** (**-1.2**)	37.3 → **21.8** (**-15.5**)	70 → **80** (**+10**)	27 → **34** (**+7**)

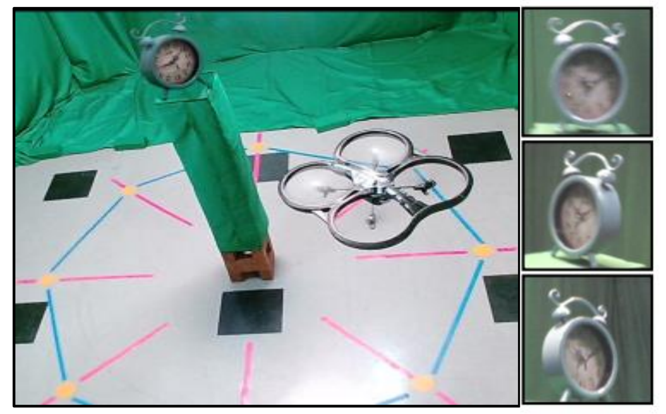
3D-ODDS Dataset

- The **3D Object Domain Dataset Suite** (3D-ODDS) is a hierarchical real-world dataset ideal as a benchmark for rigorously testing invariance.
- Contains 200,000 images and 331 corresponding 3D meshes.
- 3 disentangled factors of variation: class, viewpoint, and domain.
- About 20 different classes, 25 object instances per class.
- Domains originate from 3 types of data collection: turntable, drone, and in the wild.
- 8 viewpoints per object instance (in azimuth increments of 45º)



a) OTURN Domain
In the lab, turntable & DSLR Camera to create 3D meshes.

b) OOWL Domain
In the lab, flying drone camera



c) OWILD Domain
In the wild with real indoor & outdoor locations, smartphone camera