Towards Reconstruction of 3D Shapes in a Realistic Environment

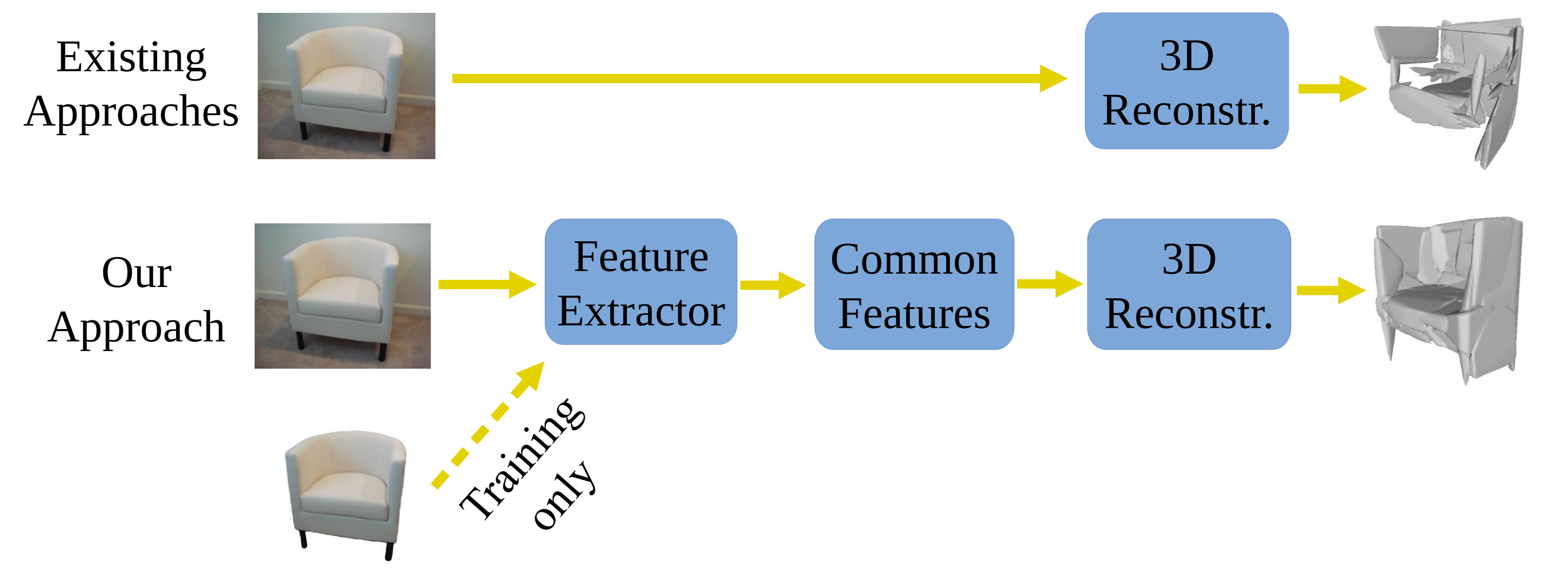
Mohammad Zohaib[1,2], Matteo Taiana[1], Alessio Del Bue[1]

[1] PAVIS, Istituto Italiano di Tecnologia, Genoa, Italy
[2] Elect., Electron. and Telecom. Eng. and Naval Arch. Department, University of Genoa, Genoa, Italy
{mohammad.zohaib, matteo.taiana, alessio.delbue}@iit.it

ICIAP 2021

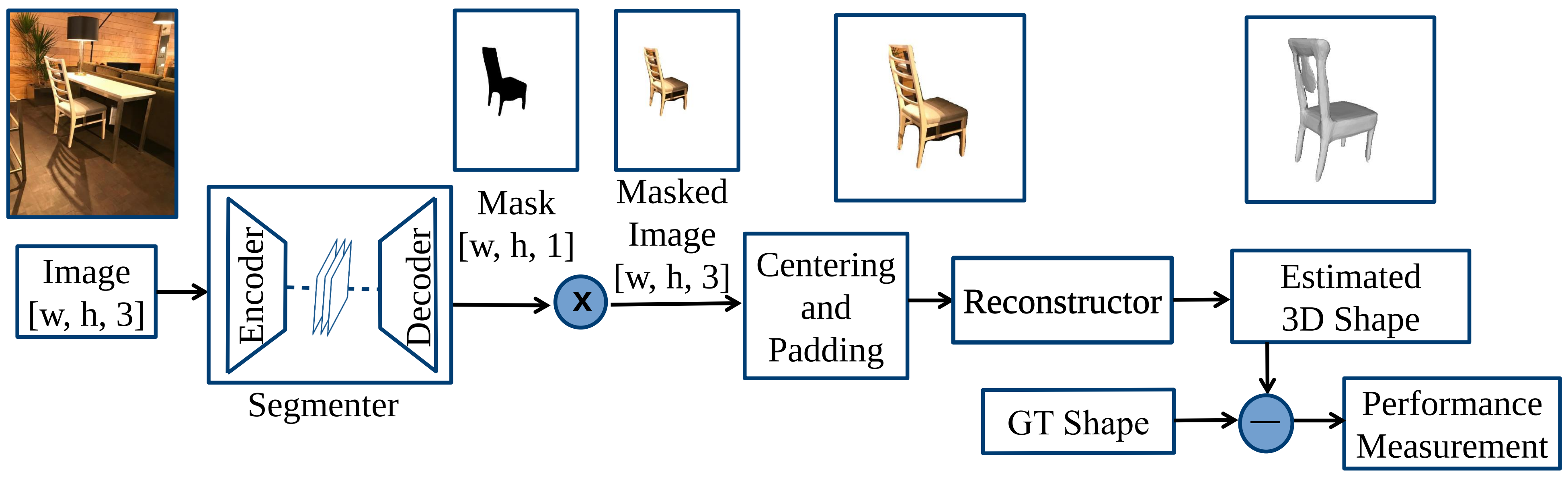
Paper ID: 67

MOTIVATION



Object's shape estimation approaches with high accuracy [1, 2] do not perform well when applied to natural images directly (top). In comparison, the proposed approach reconstructs accurate 3D shape by estimating common features for realistic and white background images (bottom).

BASELINE SOLUTION USING CvxNet [1] and ONet [2]



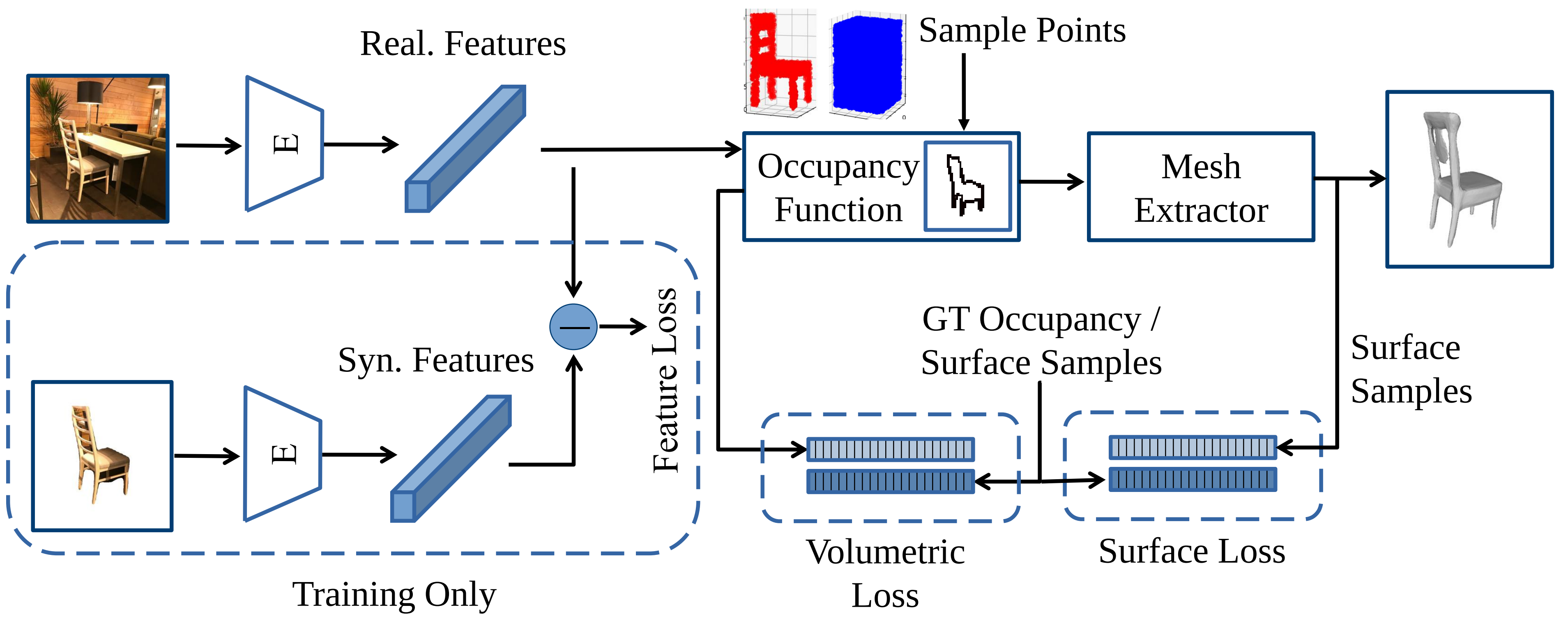
The approaches [1, 2] can produce good results for real images if the input image is processed appropriately; separating an object by applying an instance segmentation algorithm, pasting it on the center of the white image, and padding the image in order to make it similar to synthetic (image without background).

PROPOSED METHOD

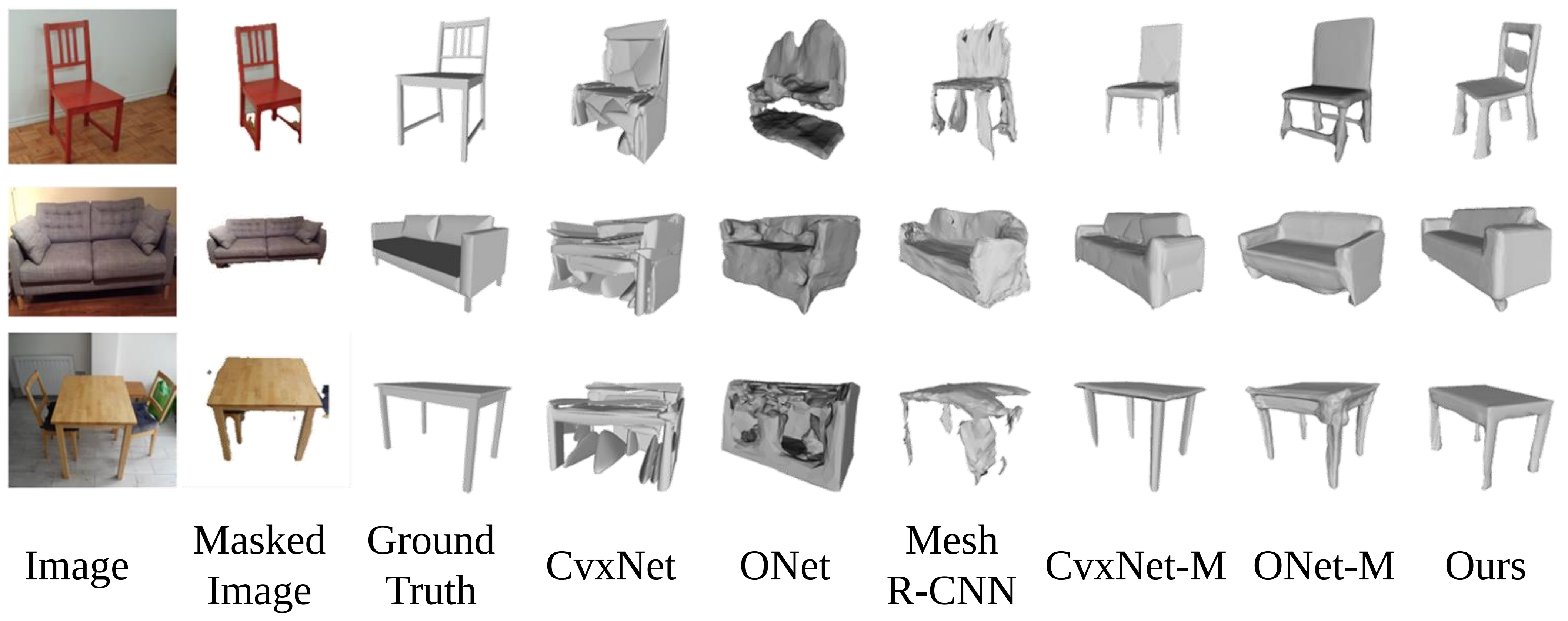
During training, an image with and without background is fed to the encoder in parallel, producing two feature vectors; synthetic and realistic. The vectors are compared in order to enable the encoder to extract common features from both image versions. The shape predictor module based on an occupancy function uses the feature vector (coming from a realistic image) for object boundary estimation. The boundary is evaluated by computing volumetric and surface loss. At inference time, only the real image is fed to the system.

$$\mathcal{L}_{vol} = \mathcal{L}_{BCE}(\mathcal{O},\mathcal{Q}), \qquad \mathcal{O} = \mathcal{F}(f_{com}\,, P_i \mid i\!:1\ to\ N)$$

$$\mathcal{L}_{surf} = \frac{1}{n_X}\sum_{x\in X}\min_{y\in Y}|x-y| + \frac{1}{n_y}\sum_{y\in Y}\min_{x\in X}|y-x|$$



QUALITATIVE RESULTS



Qualitative comparison of the proposed approach (ours) with the baselines. The masked images are obtained by removing the background, centering the object and padding. CvxNet-M and ONet-M use masked images, whereas the rest of the approaches i.e., CvxNet, ONet, Mesh R-CNN [3], and ours use natural images. Here are our findings:

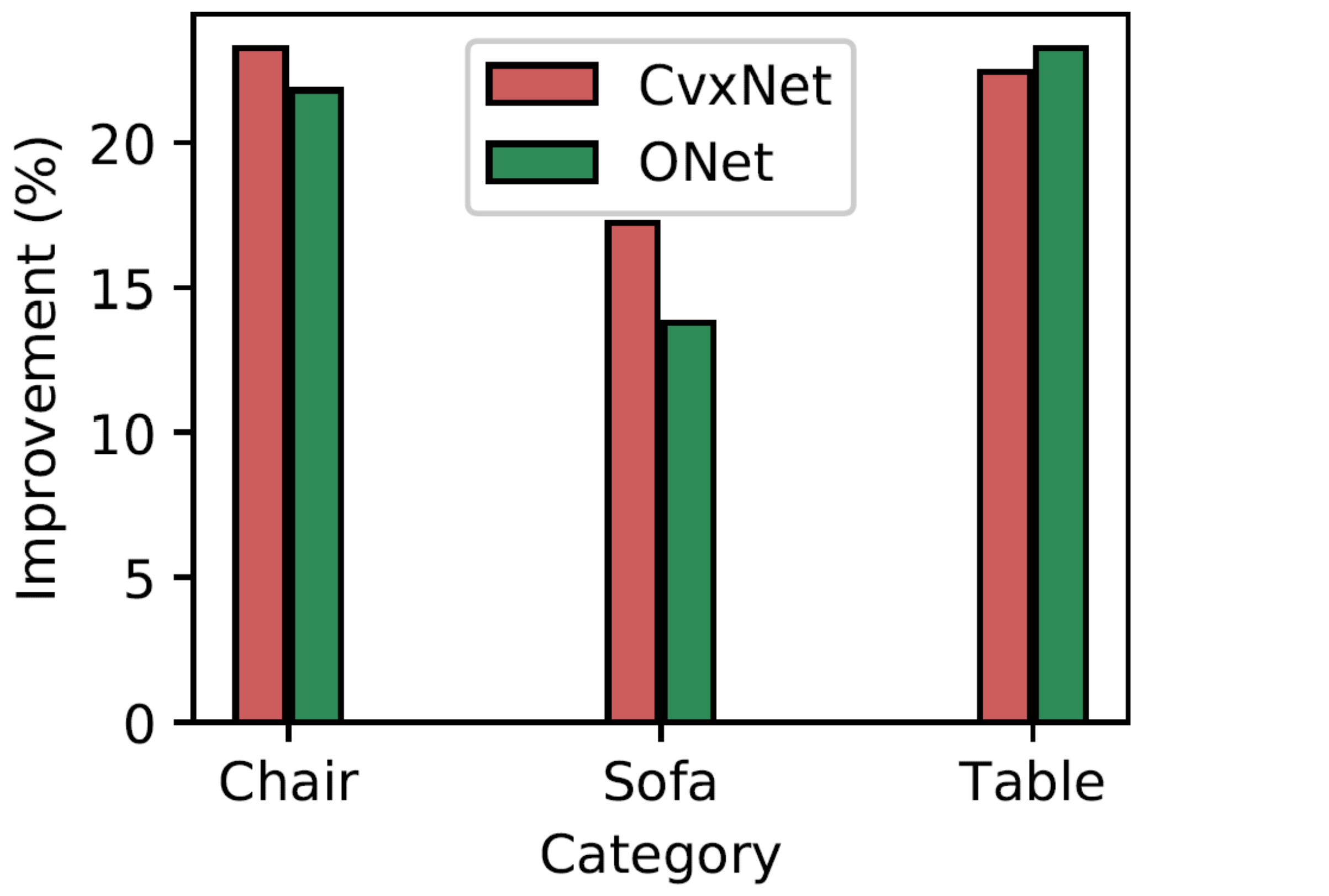
- The results of the Cvxnet-M and ONet-M are more accurate than those of CvxNet and ONet.
- Estimations by the Mesh R-CNN are not complete, specially in occluded scenarios.
- In comparison, the presented approach outperforms by estimating sharp and smoother surface and without requiring any preprocessing (segmentation) on the images.

QUANTITATIVE RESULTS

Quantitative comparison between the baselines and the approach on the Pix3D dataset. Our approach achieves better reconstruction accuracy on a scale of F1 score in all the cases, while ONet-M retains an advantage for the masked version of the chair category for Chamfer L1 distance. The best values are highlighted in bold.

Category	CvxNet	ONet	Mesh R-CNN	CvxNet-M	ONet-M	Ours
			F_1 Score (%) ↑ / Chamfer L_1 Distance ↓			
Chair	35.43/2.73	34.21/2.45	37.63/1.99	46.88/1.91	46.24/**1.54**	**47.16**/1.82
Sofa	41.99/1.94	42.45/1.91	53.61/1.76	58.35/1.68	53.73/1.75	**61.58/1.63**
Table	33.15/5.07	28.79/5.01	48.12/2.41	44.98/3.72	45.19/2.79	**48.91/2.14**
Average	36.86/3.25	35.15/3.12	46.45/2.05	50.07/2.44	48.39/2.03	**52.55/1.86**

IMPROVEMENT IN CvxNet and ONet



Category-wise improvement in CvxNet and ONet. Executing the approaches on the masked images produced a much better performance in all cases.

REFERENCES

[1] Deng, Boyang et al., "Cvxnet: Learnable convex decomposition." **CVPR**, **2020**.
[2] Mescheder, Lars et al., "Occupancy networks: Learning 3d reconstruction in function space." **CVPR**, **2019**.
[3] Gkioxari, Georgia et al., "Mesh r-cnn." **ICCV, 2019**.