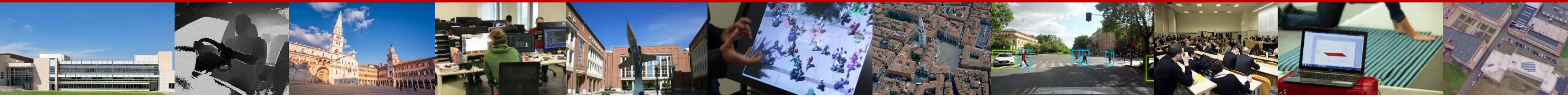
3DV 2021
International Conference on 3D Vision
London, UK / Online
December 1-3, 2021

Multi-Category Mesh Reconstruction From Image Collections



Alessandro Simoni, Stefano Pini, Roberto Vezzani, Rita Cucchiara

{alessandro.simoni, s.pini, roberto.vezzani, rita.cucchiara}@unimore.it

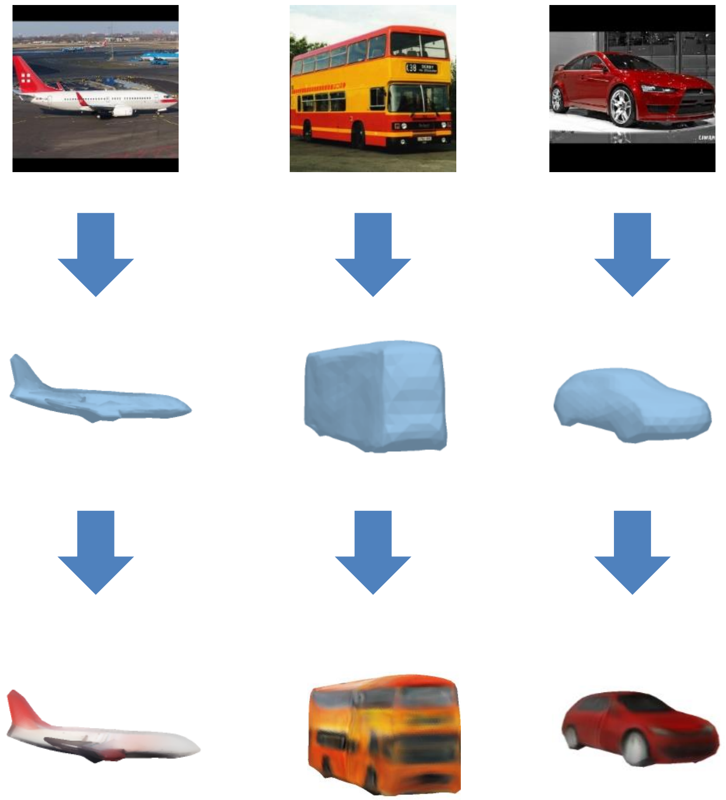
University of Modena and Reggio Emilia, Italy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

AIRI
AI Research & Innovation Center

Task definition



3D mesh reconstruction from 2D images is constantly progressing in the computer vision community.

Recent deep learning approaches [1,2,3] can restore shape, pose and texture from single-view RGB images as an **inverse graphics problem**.

All these methods share the following approach:

- They learn a mean 3D shape, called **meanshape**, representing a single object category
- They infer instance-specific deformation, texture and 3D pose that are applied to the learned meanshape

1. *Kanazawa, Angjoo et al. “Learning category-specific mesh reconstruction from image collections”. In ECCV. 2018.*
2. *Goel, Shubham et al. “Shape and viewpoint without keypoints”. In ECCV. 2020.*
3. *Li, Xueting et al. “Self-supervised single-view 3d reconstruction via semantic consistency”. In ECCV. 2020.*

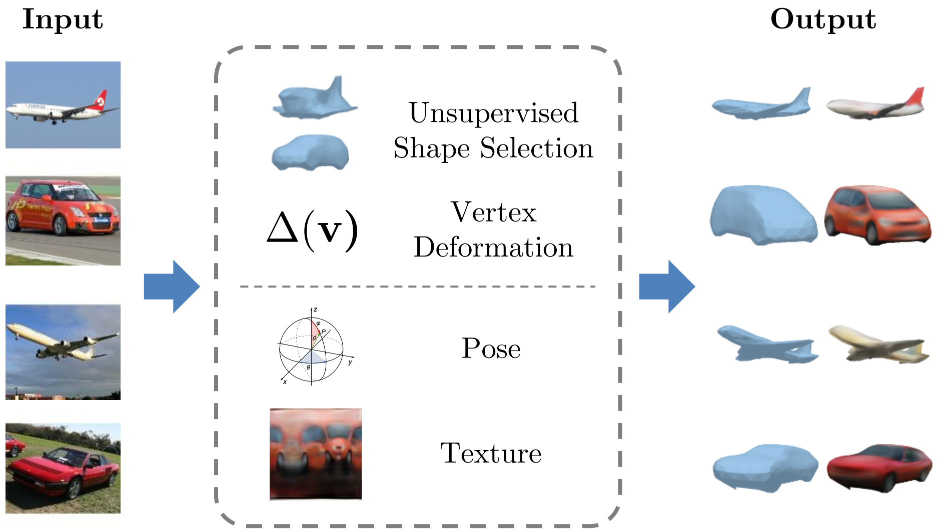
Motivations & Proposal

Limitations of current literature approaches:

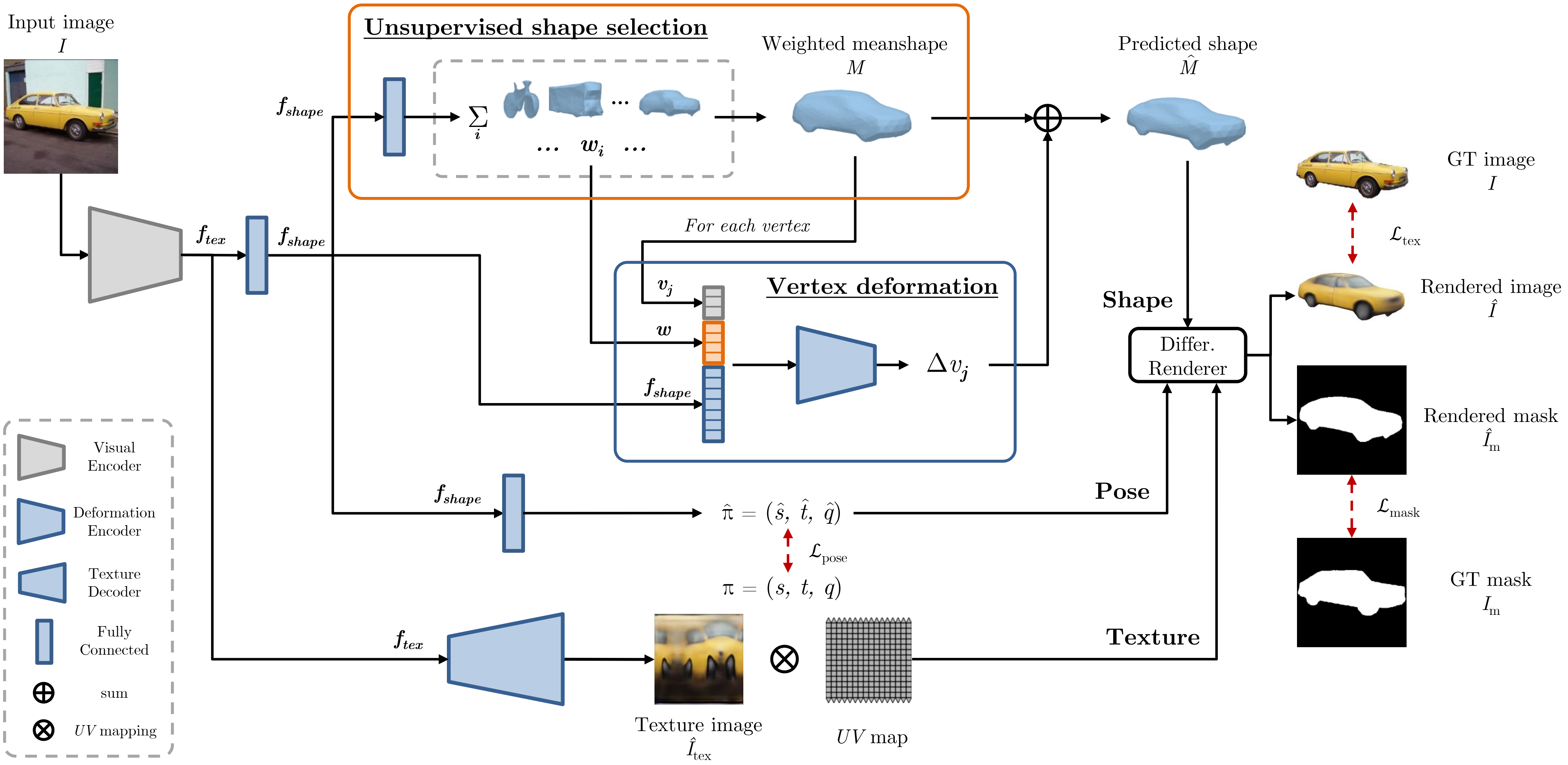
- They are **category-specific**, so trained and evaluated on image collections of a single object category
- They initialize the learnable meanshape with a category-specific **3D template model**

Our **proposal**:

- **End-to-end method** trained on image collections of multiple object categories
- **Multiple-meanshape unsupervised learning**
- Learning shapes directly from **spherical initialization**
- **No explicit category nor 3D supervision**, but only foreground masks and camera poses



Proposed method



Unsupervised Shape Selection

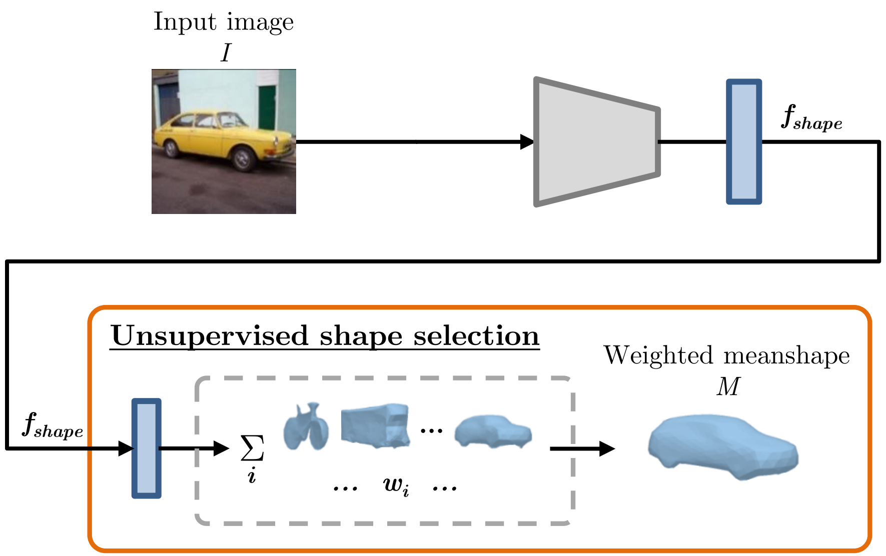
- Input → visual features f_{shape}
- Model → 2 fc layers + softmax + N deformable meanshapes (N = object categories)
- Output → set of weights w

Scores w are used to compute a weighted sum of the N meanshapes’ vertices producing a **weighted meanshape**:

$$M = (V, F) = (\sum_{i=1}^{N} w_i V_i \,,\, F)$$

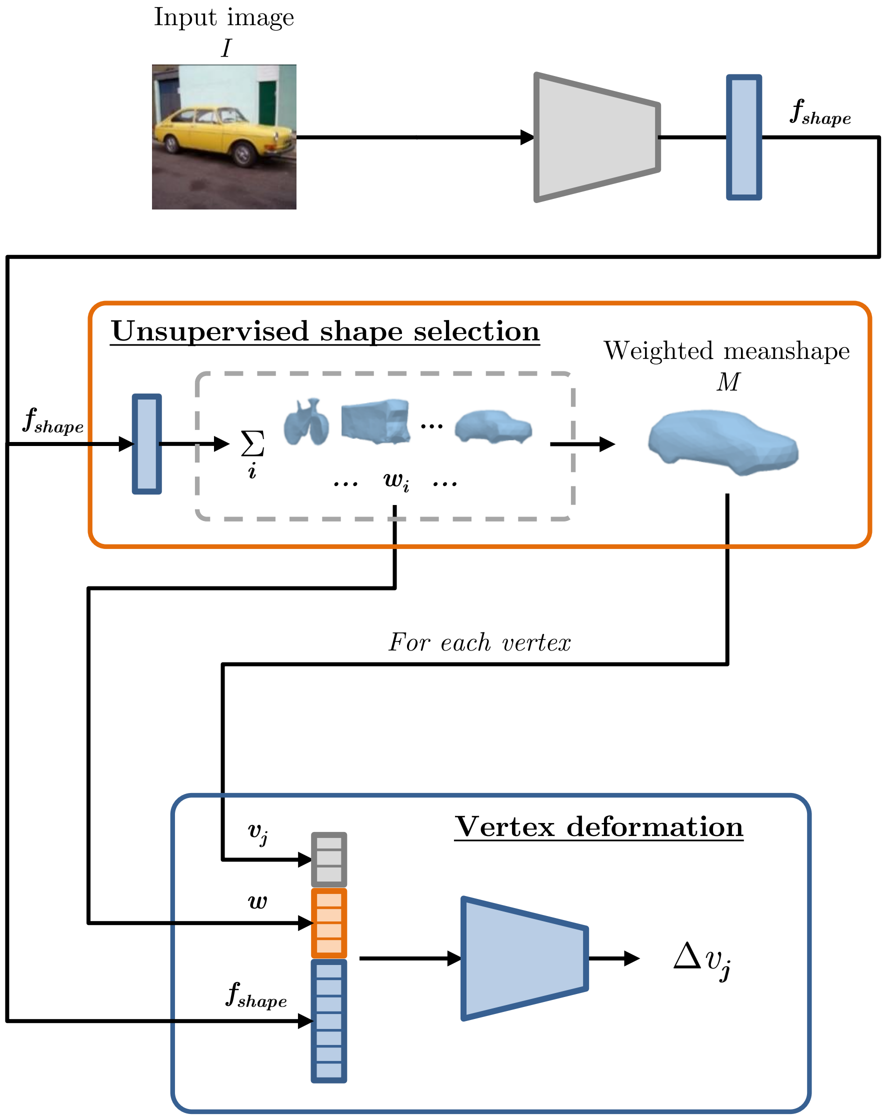
This operation results in a **smooth and differentiable approximation** of a hard shape selection.

Meanshapes are initialized as **spheres** and progressively **updated** and **specialized** in different object categories.



Vertex Deformation

- Input → visual features f_{shape} + weigths w + vertex v_j
- Model → Lightweight mlp network [1,2]
- Output → single vertex deformation Δv_j



The weighting scores w create a **connection** between the weighted meanshape M and the predicted deformation ΔV that are summed together producing the final shape:

$$\hat{M} = M + \Delta V = (V + \Delta V, F)$$

This network configuration is **independent** from the number of mesh vertices enabling:

- **dynamic mesh subdivision** during training
- **robustness** towards different mesh dimension

1. *Groueix, Thibault et al. “A papier-mâché approach to learning 3d surface generation”. In CVPR. 2018.*
2. *Park, Jeong Joon et al. “Deepsdf: Learning continuous signed distance functions for shape representation”. In CVPR. 2019.*

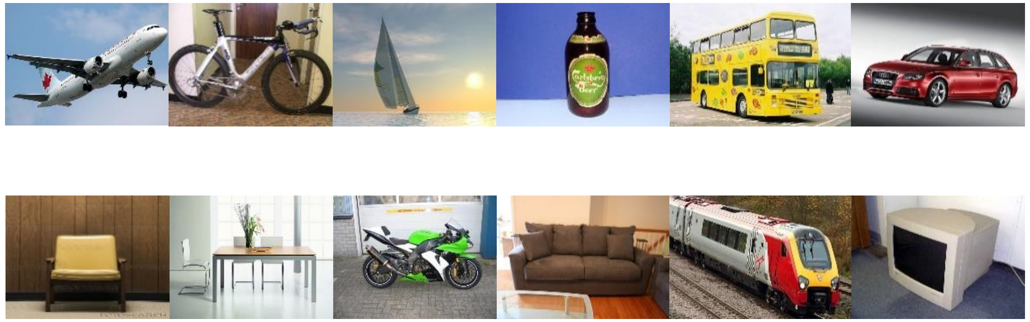
Datasets

Pascal3D+ [1]

12 rigid object classes

Annotations:

- 2D keypoints
- 3D model class
- 3D pose



CUB [2]

“Bird” class with 200 bird species

Annotations:

- Bounding box
- Rough segmentation
- Attributes (size, shape, color, …)
- 3D pose computed with SfM [3]

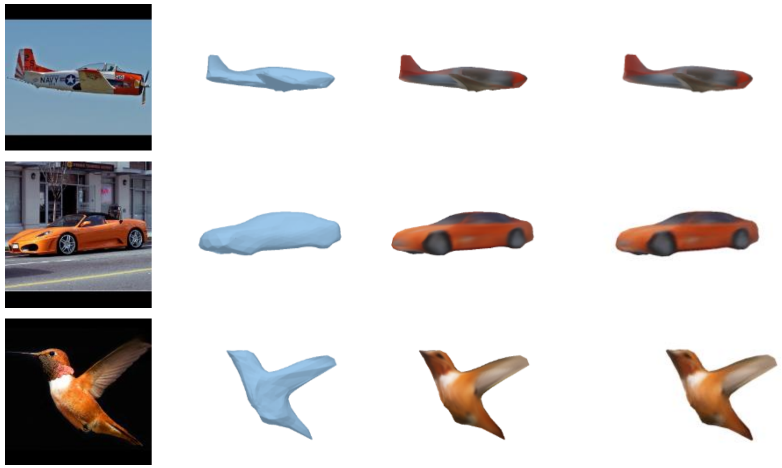
1. Xiang, Yu et al., “Beyond pascal: A benchmark for 3d object detection in the wild”. In WACV, 2014.
2. Wah, Catherine et al., “The Caltech-UCSD Birds-200-2011 Dataset”. In Technical Report CNS-TR-2011-001 (California Institute of Technology). 2011.
3. Kanazawa, Angjoo et al. “Learning category-specific mesh reconstruction from image collections”. In ECCV. 2018.

Results

Our method achieves **on par or better results** compared to category-specific approaches on Pascal3D+ and CUB.

Evaluation metrics are:

- **3D IoU** [1] for Pascal3D+
- **Mask IoU** for CUB (no 3D models available)



Pascal3D+

Approach	Training	Aeroplane	Car	Avg
CSDM [17]	indep.	0.400	0.600	0.500
DRC [48]	indep.	0.420	0.670	0.545
CMR [16]	indep.	**0.460**	0.640	0.550
IMR [47]	indep.	0.440	0.660	0.550
U-CMR [7]	indep.	-	0.646	-
Ours (N meanshapes)	indep.	**0.460**	**0.684**	**0.572**
Ours (2 meanshapes)	**joint**	**0.448**	**0.686**	**0.567**

CUB

Approach	Mask IoU ↑		Texture metrics		
	Pred cam	GT cam	SSIM ↑	L1 ↓	FID ↓
CMR [16]	**0.706**	0.734	**0.718**	**0.063**	290.32
DIB-R [2]	-	**0.757**	-	-	-
U-CMR [7]	0.637	-	0.689	0.077	**190.35**
Ours (1 meanshape)	0.658	0.721	0.717	0.064	227.24
Ours (14 meanshapes)	0.642	0.723	0.715	0.065	231.95

1. *Tulsiani, Shubham et al., "Multi-view supervision for single-view reconstruction via differentiable ray consistency". In CVPR, 2017.*

Ablation study

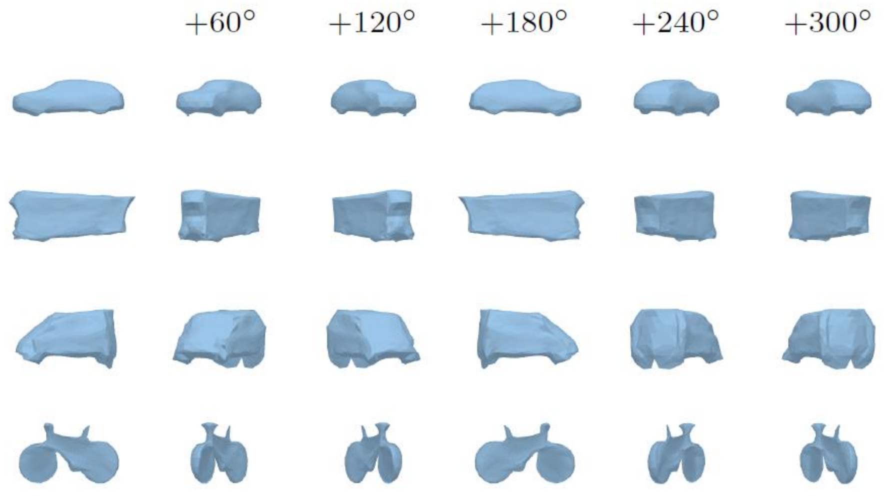
Some ablation studies on Pascal3D+.

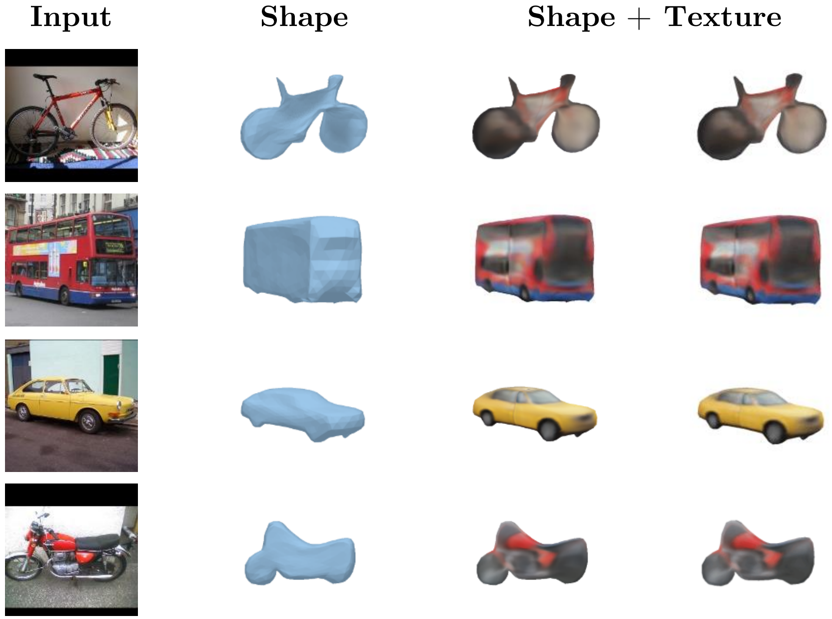
Multiple-meanshape learning achieves **better** results w.r.t. single meanshape approach.

Training classes	Number of meanshapes	3D IoU ↑	Mask IoU ↑ Pred cam	Mask IoU ↑ GT cam	Texture metrics SSIM ↑	Texture metrics L1 ↓	Texture metrics FID ↓
aeroplane, car	1	0.532	0.592	0.689	0.736	0.066	365.01
aeroplane, car	2	**0.552**	**0.671**	**0.702**	**0.737**	**0.062**	**344.80**
bicycle, bus, car, motorbike	1	0.517	0.665	0.751	0.601	0.100	390.41
bicycle, bus, car, motorbike	4	**0.543**	**0.711**	**0.759**	**0.607**	**0.094**	**380.15**
12 Pascal3D+ classes	1	0.409	0.602	0.670	0.660	0.088	357.51
12 Pascal3D+ classes	12	**0.425**	**0.620**	**0.685**	**0.665**	**0.086**	**345.90**

Dynamic mesh subdivision during training has also a **positive impact** on results.

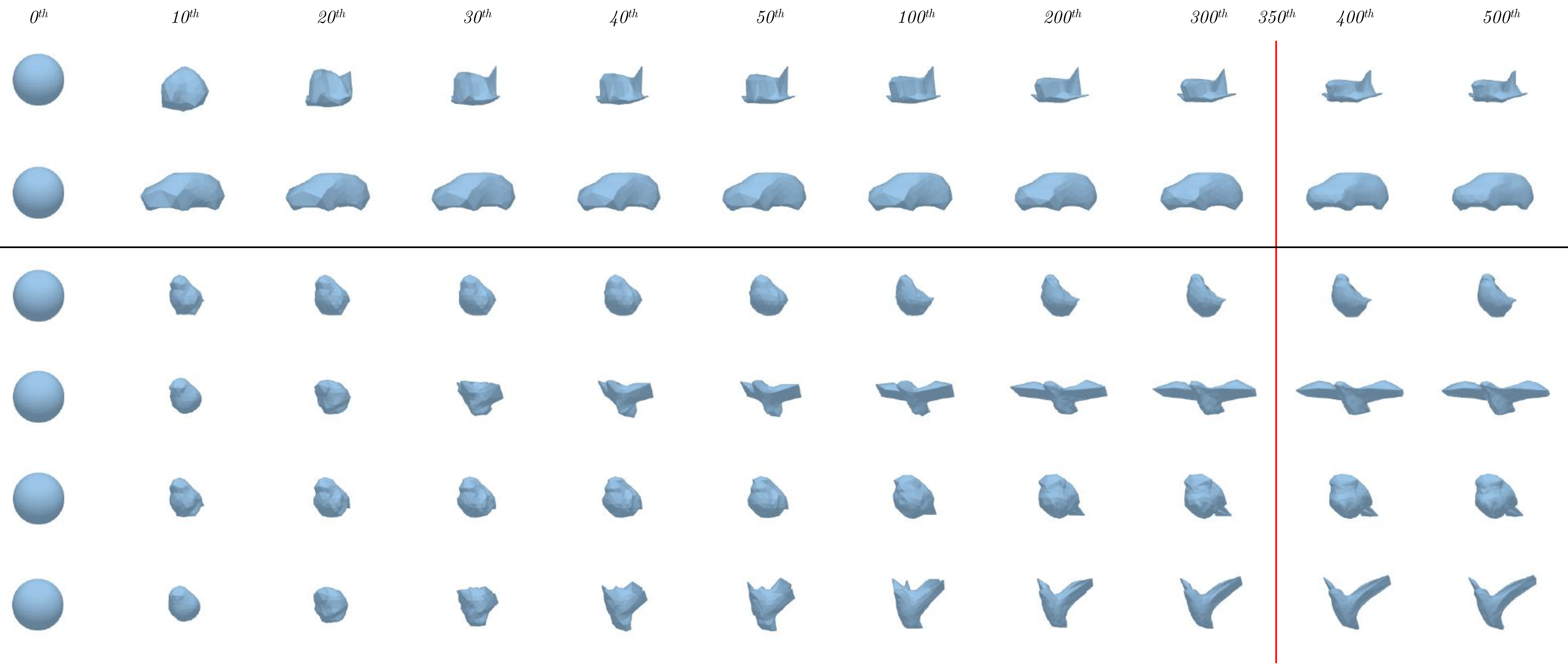
Subdivision level	Mask IoU ↑ Pred cam	Mask IoU ↑ GT cam	Texture metrics SSIM ↑	Texture metrics L1 ↓	Texture metrics FID ↓
3	0.701	**0.759**	0.600	0.096	395.96
4	0.685	0.756	0.593	0.101	385.68
$3 \rightarrow 4$	**0.711**	**0.759**	**0.607**	**0.094**	**380.15**





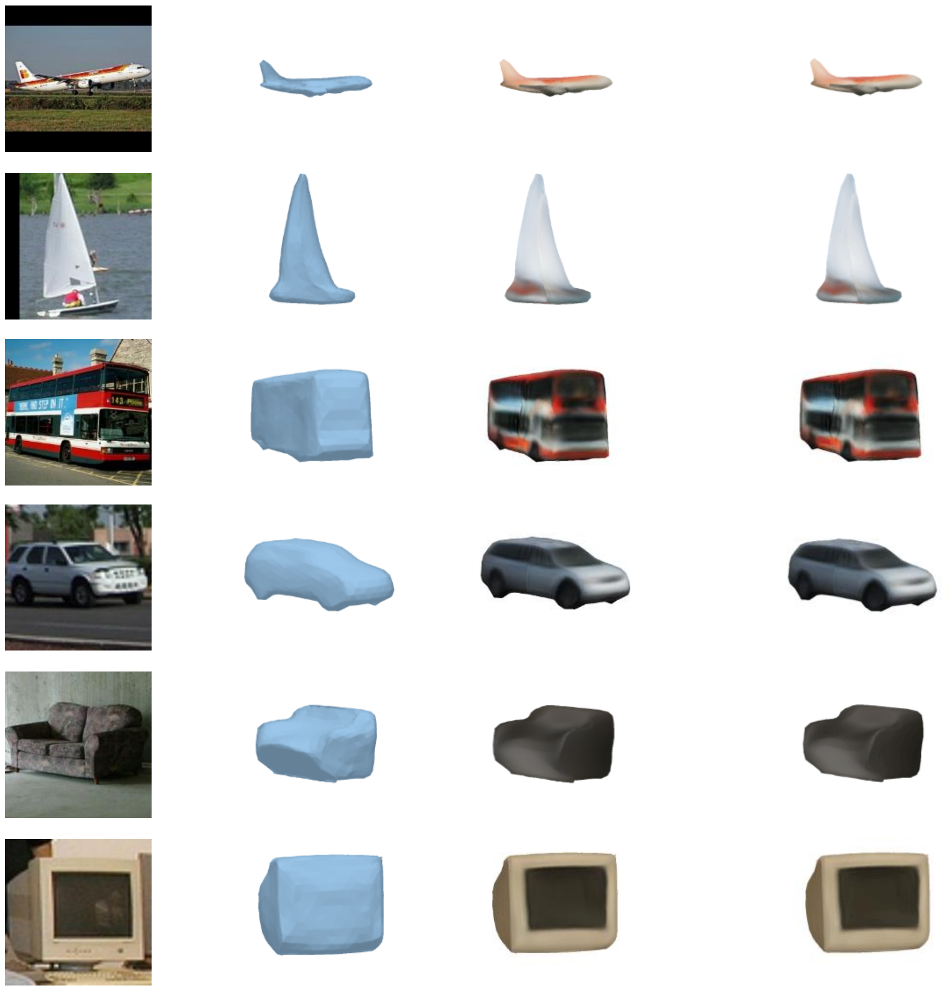
Multiple-meanshape learning

Category **meanshapes** are learned during training, evolving from **spherical initialization**.

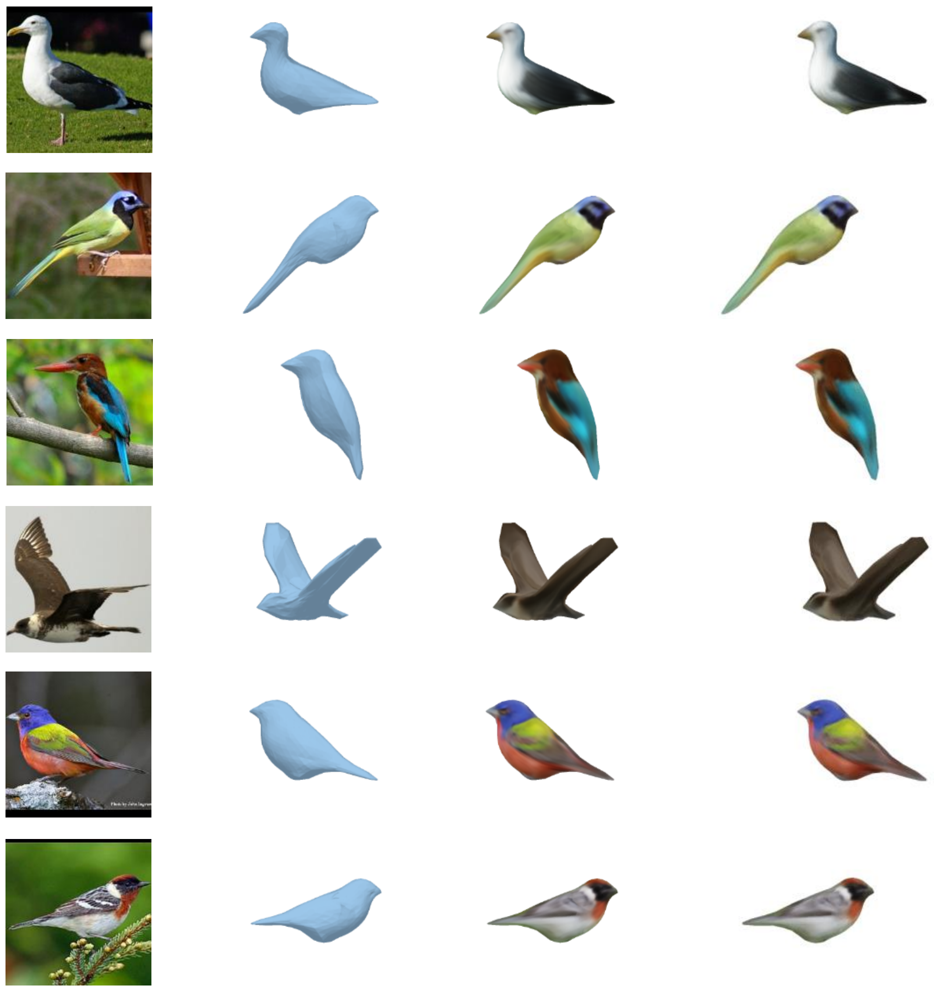


Qualitative results

Pascal3D+



CUB



Conclusion

We propose a **multi-category end-to-end method** to reconstruct a 3D object shape with only foreground masks and rough camera pose as supervision.

An **unsupervised shape selection module** (USS) is introduced in order to learn category meanshapes starting from spherical initialization.

A **vertex deformation module** predicts single vertex displacement conditioned on the output of the USS module enabling dynamic mesh subdivision during training.

The proposed method achieves **on par or better results on Pascal3D+ and CUB** datasets compared to category-specific literature approaches, while being able to predict shapes of different categories at the same time.

You can find the code on the GitHub repo:

https://github.com/aimagelab/mcmr

3DV 2021

International Conference on 3D Vision
London, UK / Online
December 1-3, 2021

Thank you for your attention!

Multi-Category Mesh Reconstruction From Image Collections

Alessandro Simoni, Stefano Pini, Roberto Vezzani, Rita Cucchiara

{alessandro.simoni, s.pini, roberto.vezzani, rita.cucchiara}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImageLab

AIRI AI Research & Innovation Center