Multi-Category Mesh Reconstruction From Image Collections

Alessandro Simoni, Stefano Pini, Roberto Vezzani, Rita Cucchiara
University of Modena and Reggio Emilia – *{name.surname, s.pini}@unimore.it*

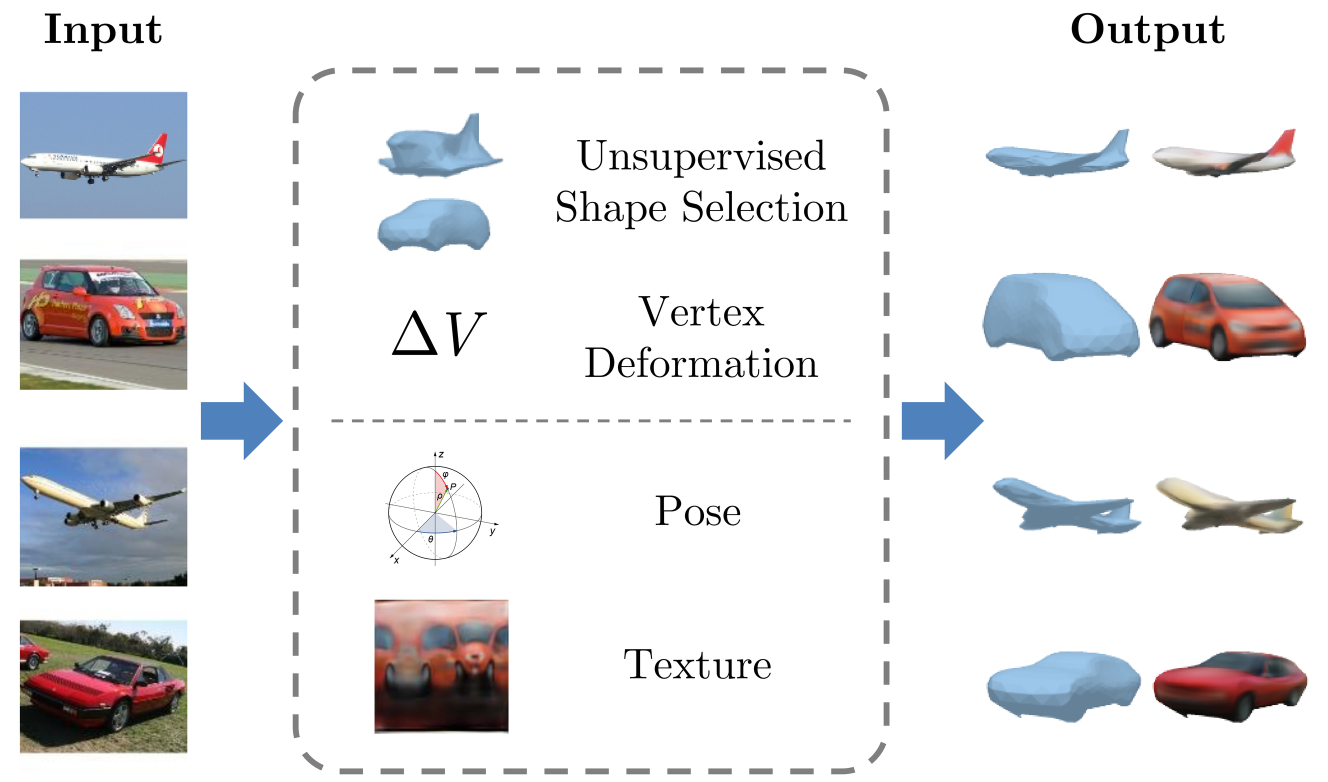
Motivations

Addressing the task as an **inverse graphics problem**, 3D mesh reconstruction from 2D single-view images has shown astonishing progress in the computer vision community. However, current literature approaches have some limitations:

- They learn **category-specific** priors training and evaluating on image collections of a **single object category**;
- They initialize the category meanshape with a **3D representative template model**.

Goals & Contributions

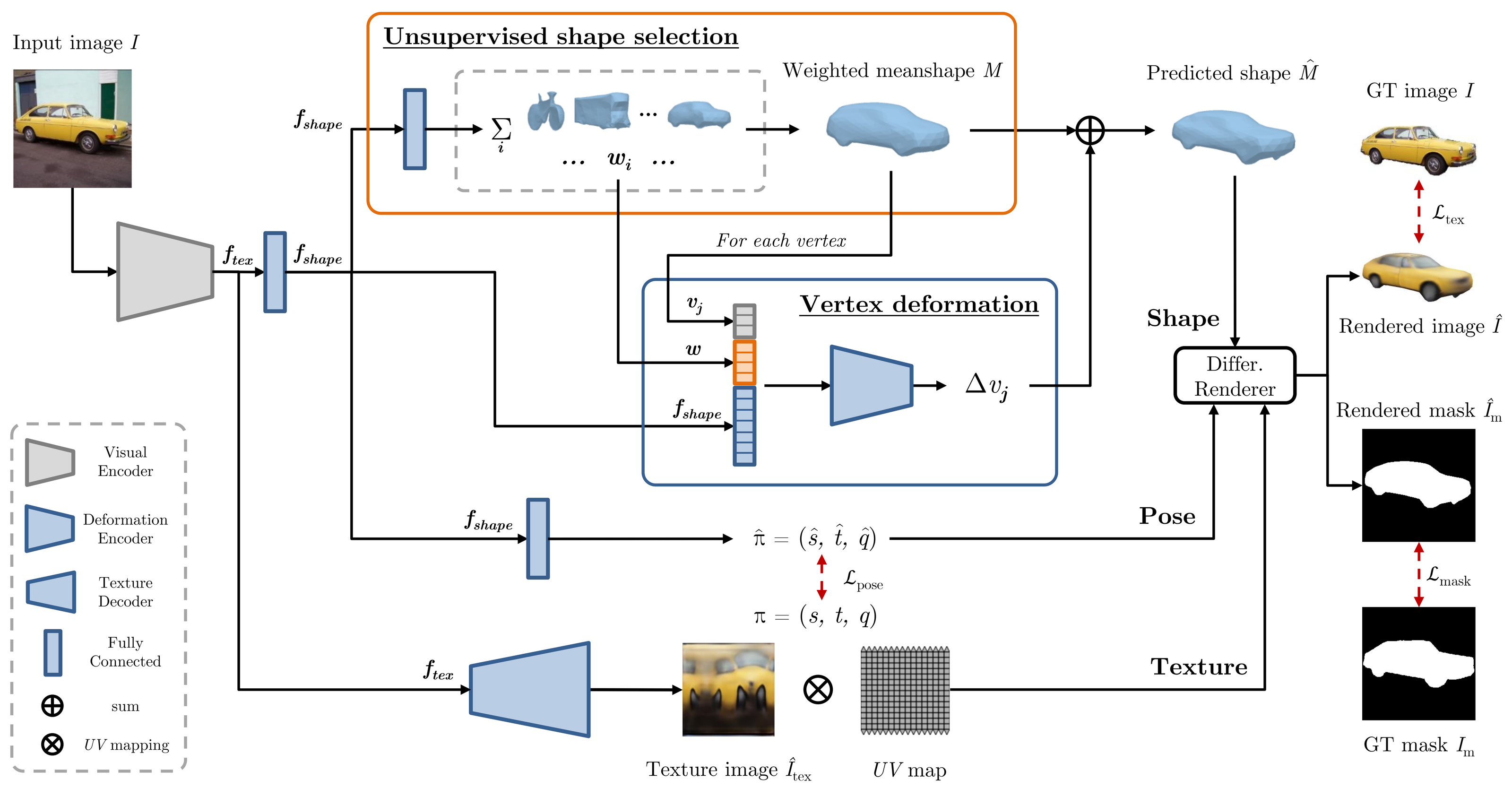
In this work, we propose an **end-to-end method** that takes single-view RGB images of **multiple object categories** and predicts their 3D textured shape. The method learns in an unsupervised manner a series of deformable 3D models, called **meanshapes**, representing each object category. To produce the final object shape, it infers **instance-specific deformations** alongside its pose and texture.



Our main contributions are as follows:

- The proposed method is trained end-to-end on single-view image collections of multiple object categories using just foreground masks and camera poses as supervision, but without **any explicit category nor 3D supervision**;
- The **unsupervised shape selection** module predicts meaningful meanshapes of different object categories, starting from a set of 3D spheres;
- The **vertex deformation** module is independent from the number of mesh vertices. Thus, it produces smooth 3D deformations and supports the **dynamic mesh subdivision** during training.

Proposed Method



The method starts by extracting two sets of visual features from an RGB image I with a **ResNet-18** network: (i) f_{shape} is used to learn shape priors and deformations, and (ii) f_{tex} is used to learn the object texture. These features are then processed by 4 different modules:

- The **unsupervised shape selection** module takes the features f_{shape} and predicts a set of weights w_i. The aim of this module is to learn a weighted meanshape M, approximating a hard shape selection over N meanshapes representing the object categories.
- The **vertex deformation** module predicts a vertex displacement Δv_j, taking as input the features f_{shape}, the set of weights w_i and a vertex v_j of the meanshape M. This operation is done in parallel for each vertex of M producing the final shape $\hat{M}$. In this way, the architecture is **independent** from the number of vertices and can process meshes of different sizes, enabling a **dynamic mesh subdivision**.
- The **3D pose predictor** module regresses a weak-perspective object pose $\hat{\pi} = (\hat{s}, \hat{t}, \hat{q})$ directly from f_{shape}.
- The **texture predictor** module takes f_{tex} and outputs an RGB texture image $\hat{I}_{tex}$, which is then mapped into the UV-space of $\hat{M}$ that is homeomorphic to a sphere.

As final step, the **SoftRas differentiable renderer** renders the 3D textured shape of the object combining $\hat{M}$, $\hat{I}_{tex}$ and $\hat{\pi}$.

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{def}} + \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{pose\_reg}} + \mathcal{L}_{\text{tex}}$$

The method is **trained end-to-end** using foreground masks and 3D pose as supervision along with some regularization terms for the shape smoothness/deformation and the quaternion rotation.

Results

Our method achieves on par or better results compared to the literature on Pascal3D+ and CUB datasets.

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

Conclusion

Our method recovers 3D textured meshes of objects from multiple categories, using RGB images as input and only foreground masks and coarse camera poses as supervision.
Code available at: `https://github.com/aimagelab/mcmr`