Weak Multi View Supervision
For Surface Mapping Estimation

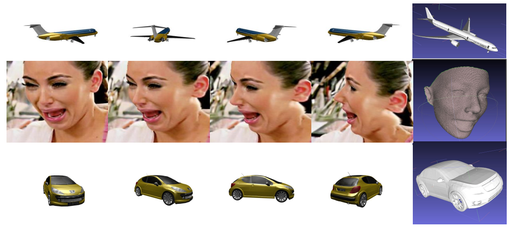
*Nishant Rai[1,2], *Aidas Liaudanskas[1], Srinivas Rao[1], Rodrigo Ortiz Cayon[1], Matteo Munaro[1], Stefan Holzer[1]
[1] Fyusion Inc [2] Stanford University

Introduction

Shape and structure estimation from images is a challenging task needing to deal with variance in shapes, varying perspectives and underlying geometry

Earlier approaches have seen success using dense supervision to learn this

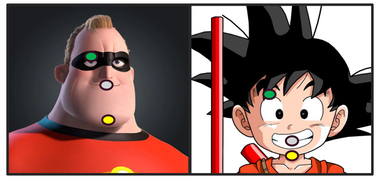
We introduce a **weakly supervised multi view approach**, along with a **richly annotated dataset** which allows us to train and evaluate comparable works



Example of multiple view samples from different categories in our dataset. We provide three categories; Faces, Planes and Cars along with additional annotations

Motivation

We look at the task of **surface mapping prediction from 2D images** which has traditionally been tackled using **heavy annotations** to learn models.



Dense Correspondence (left): Example of correspondence of semantic points across two different facial images. Dense correspondence refers to inferring this association for all points. We utilize a top-down map (below) which serves as a universal representation of such points mapping to their respective semantics

Data labelling for such a task takes a long time and makes extending to new classes very expensive

We propose a weakly supervised approach which learns to predict both the underlying geometry and surface mapping without using heavy data

Top-Down Map: Used to define a universal reference point to map points on a surface

Our task is defined as **Input:** 2D Image, **Predict:** Surface Mapping, 3D shape

Approach

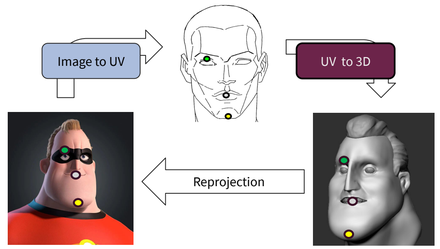
Instead of using ground truth label for supervision, we use **reprojection** and **multiview consistency**

We use camera pose **(known)**, top down **(predicted)** and 3D shape **(predicted)** to compute these losses

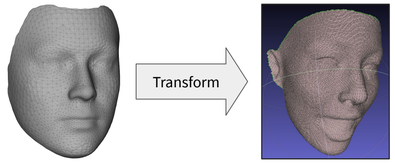
Modules

Our approach consists of three major components, each of them described below

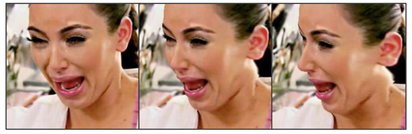
Although we use weak supervision during training, we require no additional information during inference, just a 2D image



Reprojection Cycle: We utilize reprojection as a form of supervision

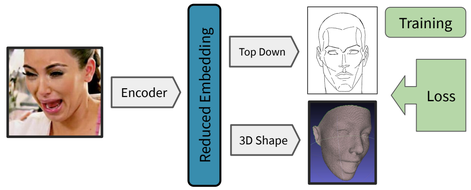


3D Shape Deformation: To handle variance in the underlying shape, we model deformations on top of a category average mesh



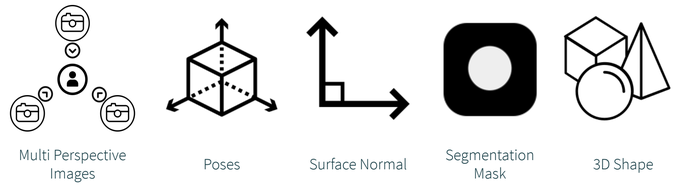
Multi View Info: We utilize predictions across different multiple views and utilize intra-view reprojection to enforce consistency

Overall Model (right): Our approach takes an image as input and uses two heads to predicts the top down prediction (UV) and 3D shape for the instance. We then utilize the predictions to compute our loss which is used to improve model predictions

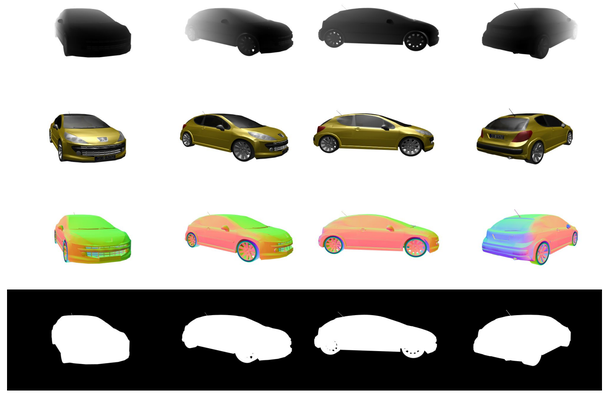


Dataset

Given the lack of datasets providing dense annotations for such tasks, we release our dataset with numerous annotations allowing training as well as evaluations



We hope this dataset encourages research by allowing better training and evaluation, through precise ground truth, of such approaches in the future



Our dataset contains numerous annotations for each category. Along with providing multi perspective images, we also provide values for surface normal, segmentation masks, camera poses, 3D shapes which allows to recompute all the above information as well.

Experiments

We evaluate our approach based on surface mapping and shape prediction

UV-Pck: Evaluates the surface mapping quality; higher is better

PosMap-Pck: Evaluates the 3D shape prediction quality; higher is better

We also analyze how our deformed 3D shape help vs using a fixed 3D shape

Deformed 3D shapes lead to **significantly higher surface mapping quality**

Experiments here are performed on the **multi-view face dataset** we release

Approach	UV-Pck@ 0.01	0.03	0.1	AUC
Single-view Reprojection with Fixed Mesh	5.3	32.2	90.6	94.0
Multi-view Reprojection with Fixed Mesh	**5.8**	**34.0**	**90.8**	**94.2**
Deformed Single-view Reprojection	**13.5**	57.8	96.0	95.7
Deformed Multi-view Reprojection	13.4	**58.4**	**96.2**	**95.9**

Benefits of using multi-view training as opposed to single view; We notice massive increase in UV performance when using deformed 3D shapes vs fixed shape

Approach	UV-Pck@ 0.03	0.1	PosMap-Pck@ 0.03	0.1
CSM*	32.2	90.6	71.7	**98.6**
Our Approach	**58.4**	**96.2**	**72.7**	**98.6**
Fully Supervised	94.9	99.5	82.0	99.8

Our approach vs the closest unsupervised approach (CSM - Kulkarni et al.) for this task. We notice a significant boost in surface mapping quality through our approach

Future Applications

Our approach looks at simultaneous prediction of surface mapping and 3D shapes using weak supervision allowing quick extension to new categories

We also release a dataset of Cars, Faces and Planes with rich annotations for such tasks, and hope it encourages research in this direction

Dataset and code available on Github!
Paper: https://arxiv.org/abs/2105.01388
Github: https://github.com/Fyusion/WMVS