TRANSMUTING A 2-D IMAGE INTO AN EXPRESSIVE 3-D MODEL

A Project Work

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE

ENGINEERING IN ARTIFICIAL

INTELLIGENCE AND MACHINE

LEARNING.

Submitted by:

SATWIK RAJ (19BCS6030)
LOVEDEEP SINGH KALSI (19BCS6007)
HARSHIT GUPTA (19BCS6013)
SANYAM SINGLA (19BCS6017)

Under the Supervision of:

Prof. BHANU PRIYANKA VALLURI

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
APEX INSTITUE OF TECHNOLOGY
CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,
PUNJAB

APRIL, 2021

DECLARATION

I, **'Satwik Raj’, 'Lovedeep Singh Kalsi’, 'Harshit Gupta’, 'Sanyam Singla,'** student of **'Bachelor of Engineering in Artificial Intelligence and Machine Learning,' session: 2020-21**, Department of Computer Science and Engineering, Apex Institute of Technology, Chandigarh University, Punjab, with this declare that the work presented in this Project Work entitled **'Transmuting a 2-D image into an expressive 3-D model'** is the outcome of our bona fide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. It contains no material previously published or written by another person or material accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Date:

April 28, 2021

Place:

Chandigarh University

SATWIK RAJ

(19BCS6030)

LOVEDEEP SINGH KALSI (19BCS6007)

HARSHIT GUPTA (19BCS6013)

SANYAM SINGLA (19BCS6017)

ABSTRACT

We hope to achieve a **high-accuracy 3D model** generated from a 2D image. Robots are mechanically capable of doing many tasks, carrying loads, precisely manipulating objects, picking, and packing, or collaborating with humans. However, they require the accurate 3D perception of things and the surrounding environment to do these tasks autonomously. Traditional methods build a 3D representation of the scene using structure from motion techniques or depth sensors. In contrast, more recent approaches use statistical models to learn geometry and the appearance of 3D objects and scenes. This thesis investigates processes to represent, understand and analyze 3D Models in natural images. We first propose two new methods for 3D model recognition and pose Estimation in single 2D images. Second, we try to compare the two pictures and use different datasets for scientific use.

We propose two novel approaches for recognizing 3D models: Detecting a human body from a 2D image after reading the idea and, with the help of our trained model, try to distinguish it from the rest of the picture. (2) Secondly, we use **multilevel pixel alignment** for reconstructing or repairing the lost body parts while visualizing. We show the state-of-the-art performance for detection and **pose Estimation** on challenging 3D model recognition datasets **(Render People dataset).**

3D object recognition methods focus on modeling the 3D Shape of the objects. However, many things may have similar 3D Shapes (washing machines, cabinets and microwave are all cuboidal); thus, recognizing them requires reasoning about appearance and geometry at the same time we are not using for the time being. The natural approach for recognition might extract pose-normalized appearance features.

ACKNOWLEDGEMENT

We feel extremely lucky for the opportunities we had that allowed us to write this thesis. We thank Mrs. Bhanu Velluri for being a great adviser and mentor. She fueled and guided by curiosity and supported us along the journey. We thank her for her insight, feedback, and support. We would like to thank all the team members, faculties that are included in this project for being awesome friends, lab mates, and faculties. This thesis would not be possible without all of your endless love and support. We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. We are highly indebted to Mrs. Bhanu Velluri for her guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We would like to express our gratitude towards the department for their kind co-operation and encouragement which helped us in the completion of this project.

TABLE OF CONTENTS

List of Symbols

Symbol	Description
I_L	*lower resolution input*
F_L	*predicted normal maps at the exact resolution*
B_L	*predicted normal maps at the exact resolution*
x_l	*the projected 2D location of X in the image space of* I_L*.*
I_H	*input image*
F_H	*frontal normal map*
B_H	*backside normal map*
x_H	*2D projection location at high resolution*
Φ^L	*the low-resolution feature extract*
Φ^H	*the image features from the high-resolution input*
$\Omega(X)$	*a 3D embedding extracted from the coarse level network*
S	*the set of samples at which the loss is evaluated*
λ	*the ratio of points outside surface in S*
$f * (\cdot)$	*the ground truth occupancy at that location*
$f^{\{L, H\}}$	*The pixel-aligned implicit functions*
L_{VGG}	*the perceptual loss*
$\lambda_{l1}L_{l1}$	*Product of relative weight and distance between the prediction and ground truth normal*

List of Figures

1. INTRODUCTION

1.1Problem definition

We are transmuting a 2-D image into an expressive 3-D model. Our goal was to estimate a 3D body from a single picture of a person. We have used the Multilevel Pixel Alignment to train our model, which is stacking the pixels from low resolution to high resolution using Multilevel Perceptron. So, we obtained high-fidelity 3D reconstruction of clothed humans from a single image at a resolution sufficient to recover detailed information such as fingers, facial features, and clothing folds. We estimated the pose using 2D Person Pose Estimation, which detects. It detects a skeleton (which consists of key points and connections between them) to identify human poses for every person inside the image. We have used a high-resolution image for making the model. We have used a high-quality image which gives more information as compared to low-resolution images. Our model can work with a 1K resolution image very perfectly. It provides a 3D object file as output which can further be processed to use that model.

Some applications of these models are: -

i. 3D model printing and making toys.
ii. Making gaming characters.
iii. Making 3D animations.
iv. AR/VR application.

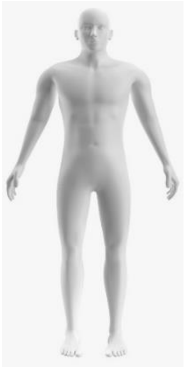

Fig. 1.1: 2D Image

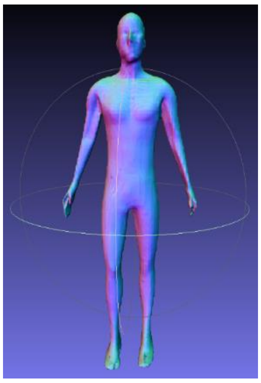

Fig. 1.2: 3D Model of image

1.2 Project Overview/Specification: -

To achieve the high-fidelity 3D reconstruction of clothed humans from a single image at a resolution sufficient to recover detailed information such as fingers, facial features, and clothing folds. As we have seen that in the existing approaches do not make full use of the high resolution (e.g., 1k or larger) imagery of humans that is now easily acquired using commodity sensors on mobile phones. We concluded that if we train our model in an end-to-end multilevel framework that infers 3D geometry of clothed humans at an unprecedentedly high 1k image resolution in a pixel-aligned manner, retaining the details in the original inputs without any post-processing. As we have trained our model on 2-3 pose only.

In this, we have used 3D human digitization that takes 1024×1024 resolution images as input.

Our method is composed of two levels: -

A coarse level like PIFu, focusing on integrating global geometric information by taking the downsampled 512 × 512 images as input and producing backbone image features of 128 × 128 resolution.

$$\mathbf{f^{L}(X) = g^{L}\,(\Phi^{L}\,(x_{L}\,, I_{L}\,, F_{L}, B_{L}),Z)}$$

where I_L = Lower resolution input, F_L and B_L are predicted normal maps at the exact resolution. $x_L \in R^2$ is the projected 2D location of X in the image space of I_L.

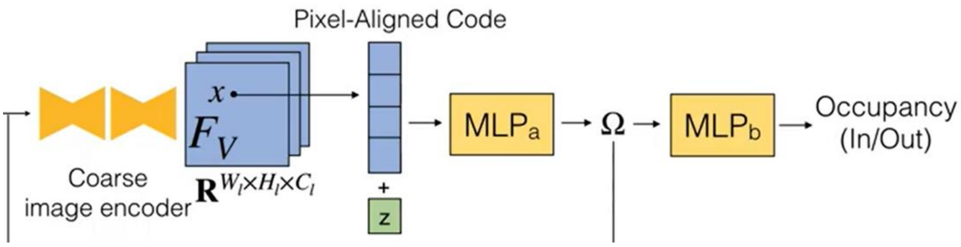

Fig. 1.3

A fine level focuses on adding more subtle details by taking the original 1024×1024 resolution image as input and producing backbone image features of 512×512 resolution (four times higher resolution than the implementation.

$$\mathbf{f^{H}(X) = g^{H}((\Phi^{H}(x_{H}, I_{H}, F_{H}, B_{H}),\Omega(X))}$$

where I_H, F_H, B_H are the input image, normal frontal map, and normal backside map, respectively, at a resolution of 1024×1024. $x_H \in R\ 2$ is the 2d projection location at high resolution, and thus in our case $x_H = 2x_L$.

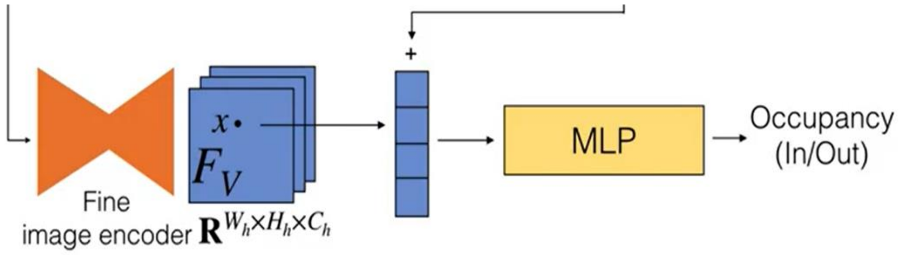

Fig. 1.4

Our method takes predicted frontside and backside normal maps (means our model is trained on both the front and backside of the image). Therefore, the backside must be inferred entirely by the MLP prediction network. Due to this problem's ambiguous and multimodal nature, the 3D reconstruction tends to be smooth and featureless.

We predict the backside and frontal normal in image space using a pix2pixHD network, mapping from RGB color to normal maps.

Dataset: -

To obtain high-fidelity 3D geometry and corresponding images, we use Render People data which consists of commercially available 100 high-resolution photogrammetry scans. The image in the dataset consists of a human full-body pose of a single person with all the angles. We split the dataset into a training set of 70 subjects and a test set of 30 subjects and render the meshes with precomputed radiance transfer using 163 second-order spherical harmonics from HDRI. Each subject is rendered from every other degree in the yaw axis with an elevation fixed with 0.

We sample points using a mixture of uniform volume samples and importance sampling around the surface using Gaussian perturbation around uniformly sampled surface points.

Implementation: -

In the out model, we have used 2D Person Pose estimation to estimate the pose of a single person that detects a skeleton (which consists of key points and connections between them) to identify human poses for every person inside the image. The pose may contain up to 18 key points: ears, eyes, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles.

In this model, we have used image encoders for both the low-resolution and high-resolution levels use a stacked hourglass network with 4 and 1 stacks, respectively, using the modification suggested by and batch normalization replaced with group normalization. The fine image encoder removes one downsampling operation to achieve a large feature embedding resolution. The feature dimensions are 128 × 128 × 256 in the coarse level and 512 × 512 × 16 in the fine level. The MLP for the coarse-level image encoder has the number of neurons of (257, 1024, 512, 256, 128, 1) with skip connections at the third, fourth, fifth layers. The MLP for the fine-level image encoder has the number of neurons of (272, 512, 256, 128, 1) with skip connections at the second and third layers.

The second MLP takes the output of the fourth layer in the first MLP as 3D embedding $\Omega \in R^{256}$ instead of absolute depth value together with high-resolution image features Φ H $(x_H, I_H, F_H, B_H) \in R^{16}$, resulting in the input channel size of 272 in total. The coarse PIFu module is pre-trained with the input image resized to 512 × 512 and a batch size of 8. The fine PIFu is trained with a batch size of 8 and a random window crop of size 512 × 512. We use RMSProp with weight decay by a factor of 0.1 every 10 epochs.

We train two networks that predict frontside and backside normal individually with the following objective functions:

$L_N = L_{VGG} + \lambda_{l1} L_{l1}$

where L_{VGG} is the perceptual loss and L_{l1} is the L_1 distance between the prediction and ground truth normal. The relative weight λ_{l1} is set to 5.0 in our experiments.

In this experiment, we implement: -

1) a pixel-aligned implicit function using only our fine-level image encoder by processing the full resolution as input during training, conditioning 1 with the jointly learned global feature using ResNet34 as a global feature encoder in spirit to a single PIFu (i.e., our coarse-level image encoder) by resizing input to 512 × 512, our proposed multilevel PIFu (two levels) by training all networks jointly (ML-PIFu, end-to-end), and ours with alternate training of the coarse and fine modules (ML-PIFu, alternate).

Loss Function: -

For loss function, we have used extended binary cross-entropy.

$$\mathcal{L}_o = \sum_{\mathbf{X}\in\mathcal{S}} \lambda f^*(\mathbf{X}) \log f^{\{L,H\}}(\mathbf{X}) \ + (1-\lambda)\left(1-f^*(\mathbf{X})\right)\log\left(1-f^{\{L,H\}}(\mathbf{X})\right)$$

where S denotes the set of samples at which the loss is evaluated, λ is the ratio of points outside surface in S, $f * (\cdot)$ denotes the ground truth occupancy at that location, and $f^{\{L, H\}}$.

Limitations: -

a. An image that is to be converted into the 3D model is limit to a single person only.

b. We have trained our model for 2-3 Pose only so other complex poses can result in the distorted 3D model as output.

c. Front-facing with standing works best (A simple exercise pose works best).

d. Our input image processing is limited to jpg format only other image format cannot be processed.

e. Image resolution less than 512 x 512 will result in a distorted 3D model.

f. Make sure the input image is well lit. Extremely dark or bright image and strong shadow often create artefacts.

g. It is trained with human-only; anime characters may not work well with this model.

Challenges: -

1. As this Model needs high-end hardware configuration so we cannot run this model on a local machine.
2. It takes a lot of time to train the model for at least 6-7 hours.
3. To overcome this model training time constraint we have used 100 images only to train our model in which 70 is used for training our model and 30 is used for testing purpose.

1.3 Minimum Hardware Specification

i. CPU – Intel i5 9th Gen or AMD Ryzen 7 processor and above.

ii. Ram – 16GB and above.

iii. GPU – 8GB NVIDIA minimum recommended and above.

1.4 Software Specification

i. Operating System: Linux

ii. Python 2.7 or higher with suitable IDE.

iii. Libraries: Pytorch, trimesh, os, argparse, NumPy, pandas, OpenCV, PIL, etc.

iv. CUDA 7.3

v. CUDNN 8.1

2. LITERATURE REVIEW

2.1 Existing System

The goal of the existing system was to make 3D human digitization that can be achieved by estimating the occupancy of a dense 3D volume, which determines whether a point in 3D space is inside the human body or not. In contrast to previous approaches, where the target 3D space is discretized, and algorithms focus on estimating each voxel's occupancy explicitly. Since no explicit 3D volume is stored in memory during training, this approach is memory efficient, and more importantly, no discretization is needed for the target 3D volume, which is important in obtaining high-fidelity 3D geometry for the target human subjects. The input size and the image feature resolution of other existing work are limited to at most 512×512 and 128 × 128 in resolution respectively, due to memory limitations in existing graphics hardware. Importantly, the network should be designed such that its receptive field covers the entire image so that it can employ holistic reasoning for consistent depth inference, thus, a repeated bottom-up and top-down architecture with intermediate supervision plays an important role to achieve robust 3D reconstruction with generalization ability. This prevents the method from taking higher resolution images as input and keeping the resolution in the feature embeddings, even though this would potentially allow the network to leverage cues about detail present only at those higher resolutions.

2.2 Proposed System

We found that while in theory, the continuous representation of the existing model can represent 3D geometry at an arbitrary resolution, the expressiveness of the representation is bounded by the feature resolution in practice. Thus, we need an effective way of balancing robustness stemming from long-range holistic reasoning and expressiveness by higher feature embedding resolutions.

We present a multilevel approach towards higher fidelity 3D human digitization that takes 1024×1024 resolution images as input. Our method is composed of two levels of existing system modules:

(1) A coarse level like Pixel Alignment Function, focusing on integrating global geometric information by taking the downsampled 512 × 512 images as input, and producing backbone image features of 128 × 128 resolution.

(2) A fine level that focuses on adding more subtle details by taking the original 1024×1024 resolution image as input and producing backbone image features of 512×512 resolution.

Flow Chart: -

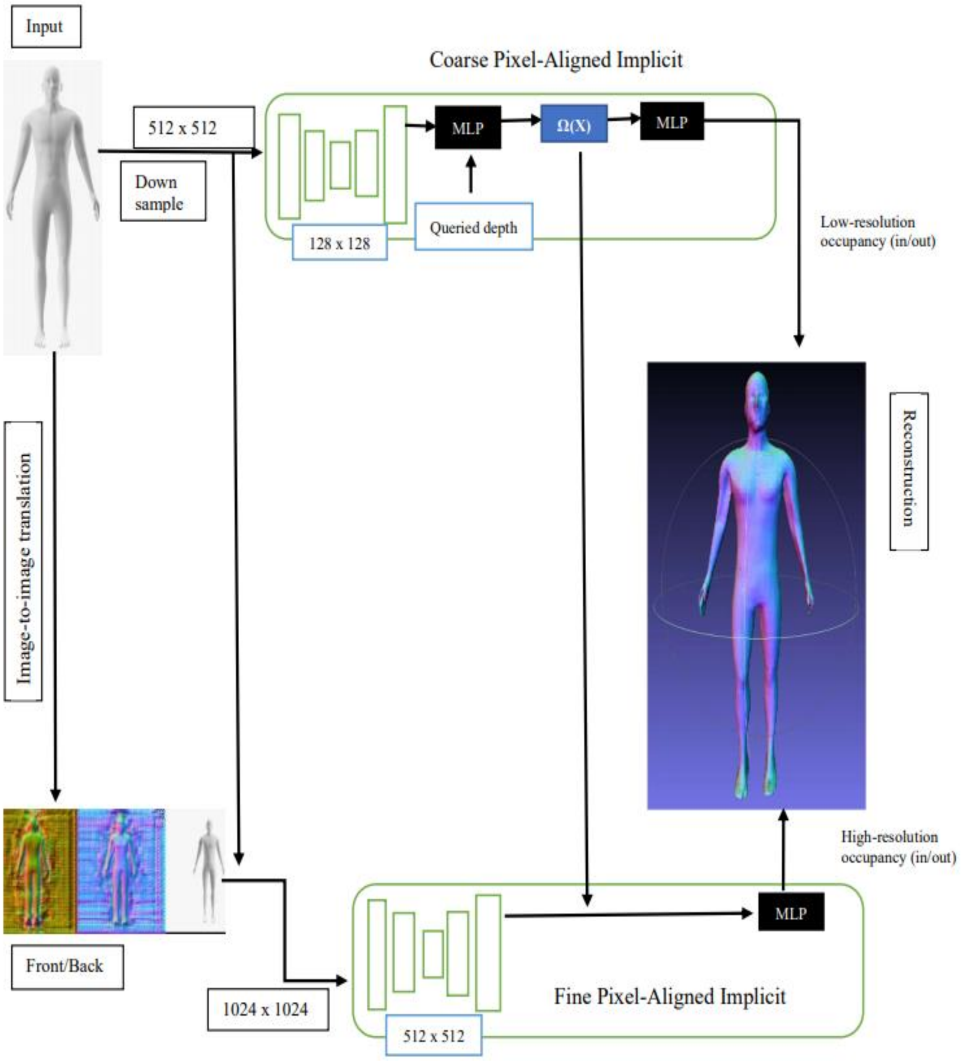
Input
512 x 512
Down sample
Coarse Pixel-Aligned Implicit
MLP
Ω(X)
MLP
128 x 128
Queried depth
Low-resolution occupancy (in/out)
Image-to-image translation
Reconstruction
High-resolution occupancy (in/out)
MLP
Front/Back
1024 x 1024
512 x 512
Fine Pixel-Aligned Implicit

Fig. 2.0

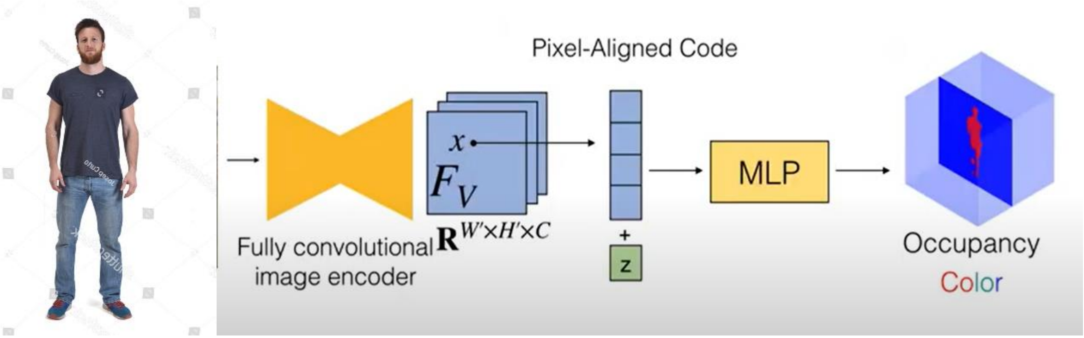
Pixel-Aligned Code
x
F_V
$\mathbf{R}^{W'\times H'\times C}$
+
z
MLP
Fully convolutional image encoder
Occupancy
Color

Fig. 2.1

2.3 Literature Review Summary

Table 2.1: Literature review summary

Year and citation	*Purpose of the study*	*Granularity Level*	*Type of vulnerabilities*	*Data set*	*Evaluation parameters*
2019	To know about the pixel alignment function for 3D image construction	PIFu aligns individual local features at the pixel level to the global context of the entire object in a fully convolutional manner, and does not require high memory usage, as in voxel-based representations.	Input image resolution is to limited to at most 512×512 and 128 × 128 in resolution respectively,	RenderPeople and BUFF, which has ground truth measurements, as well as DeepFashion which contains a diverse variety of complex clothing	Evaluations of method using ground truth 3D scan datasets obtained using high-end photogrammetry.
2011	To know more about the FLAME (Faces Learned with an Articulated Model and Expressions) model that we are using.	Vertex-based mode with a relatively low polygon count, articulation, and blend skinning.	Registering 4D Scans.	D3DFACS (Facial Action Units) (33,000 head scans)	Comparing FLAME Output Models with static 3D Scans. The FLAME model is significantly more accurate and compact.
1999	First generic 3D face model learned from scanned data.	Defines a linear subspace to define shape and texture using principal component analysis (PCA).	Requires user- specific calibration or training procedure.	Head scan of 200 young Caucasian Adults.	Perfectly fits a model to data, this research became an inspiration for Basel Face Model (2016-2017)
2016	Study about Human Mesh Recovery (HMR)	In contrast to most current methods that compute 2D or 3D joint locations, HMR produces a richer and more useful mesh representation that is parameterized by shape and 3D joint	Reprojection loss leaves the model constrained.	Large database of 3D Human meshes (we are not directly using this).	This paper proves that there is no need for 2D-to-3D supervision, which supports the notion of our project.

		angles.			
2019	Neural Network for Detailed Human Depth Estimation	To achieve minute geometric details, it separates the depth map into a smooth base shape and design a network with two branches to regress them respectively.	Have to use some off-the-shelf tools like MaskRCNN	Thousands of 256 x 256 segmented cropped images to evaluate the depth of the image	Quantitative comparison with fused `ground truth' captured by real depth cameras and qualitative examples on unconstrained Internet images demonstrate the strength of the proposed method
2017	Learning Shape, Reflectance and Illuminance of Faces in the Wild using SfSNet as the learning framework	SfSNet is a designed to reflect a physical Lambertian rendering midel. SfSNet learns from a mixture of labelled synthetic and unlabelled real-world images to produce an accurate decomposition of an image into albedo and lighting.	Supervised learning can generalize poorly if real test data comes from a different distribution than the synthetic training data.	CelebA Dataset(Contains thousands of pictures of celebrities wearing furry clothes).	SfSNet produces significantly better quantitative and qualitative results than state-of-the-art methods for inverse rendering and independent normal and illumination estimation.
2020	Multi-Level Pixel Alignment Implicit Function for high resolution 3D human Digitization	A coarse level observes the whole image at lower resolution and fine level estimates highly detailed geometry by observing higher-resolution images	Memory and computing power limitation for accurate 3-D predictions.	Thousands of synthetically generated 3-D human mesh. Model are used to train function f from in an ene-to-end fashion	Fully leverages 1K resolution using a multi-level architecture, produces more accurate 3-D prediction which we needed.
2016	DeepFashion : Powering Robust Clothes Recognition and Retrieval with Rich Annotations	Determines clothing features by jointly predicting clothing attributes and landmarks. The estimated landmarks are then employed to pool or gate the learned features.	Limited in the amount of annotations and are difficult to cope with the various challenges in real-world applications	Pre-train FashionNet on a subset of 300, 000 images of DeepFashion, another subset of 50, 000 images is used as validation data.	The performance of different methods on category classification and attribute prediction.

3. PROBLEM FORMULATION

From the literature review, it is observed that studies highlight the need for an efficient and scalable approach for detecting 2d image and converting it into 3d model. The existing techniques are not able to detect all types of vulnerable 2d images and thus there is a requirement for some advanced algorithms such as ***multilevel pixel alignment.*** Different approaches suffer from high false-negative rate and not scalable to large software systems due to high time complexity. Firstly, there is a need to transmute a 2D image into its 3D model, which requires prior knowledge of the 3D Shape. It is Essential to obtain ***3D model-from-2D visualization*** which is the act of recovering the ***"lost" third dimension***. When an observer views a cross-section of a fully 3D model, he or she directly perceives a 2D image. However, the information about the third dimension is implied if the image is known to be a cross-section obtained by cutting through the object at a specific location. Thus, a given cross-section can be linked to its 3D origin, and the entire 3D structure can be reconstructed from a series of cross-sections by locating their origins in space and assembling them into a ***coherent structure***. While this seems straightforward in principle the reconstruction of a 3D model from 2D images ***is challenging and strongly depends on the observer's spatial abilities***. In particular, the failure to understand the correspondence between 2D and 3D representations has been highlighted as an important impediment to this form of visualization. The present report aims to investigate the factors that may impede or facilitate such spatial understanding. Hence, the visualization process and then constructing a 3d model from the 2D image using some advanced featured algorithms.

We focus here on integrating ***spatial information*** within and across a sequence of images to visualize an object posed in external space. An applied motivation for the present work is identifying two human 3d models and thus them for scientific purposes. In scientific terms, wherever we are required to compare two images, we can use a similar model for the same for the identification or use a ***particular colour dataset*** for advanced comparisons. ***Orientation*** is also important in identification so that a 3d model can be compared through every possible orientation. However, as was noted, this requires a difficult visualization process. This report is directed at understanding whether and how 3D object representations, particularly their spatial properties such as slant, can be reconstructed from 2D aperture views, or slices. We concentrate on contrasting two conditions that are expected to facilitate vs. impede the visualization of 3D structure from a set of 2D slices.

4. RESEARCH OBJECTIVES

The proposed research aims to carry out work leading to developing a 3d model from a 2D image. The proposed aim will be achieved by dividing the work into the following objectives and addressing some questions:

1. To understand and explore various level of performance that can be achieved with a seemingly simple task: visualizing a 2d image and detecting a human body from the 2d image given
2. To study and analyze whether the human body detected from the 2d image has all the requirements and the ***facial expressions*** required to convert it into 3d model. To address these issues, one must train the algorithm (using multilevel Pixel alignment in our case) to detect the proper human body.
3. To design and develop the technique to the extent that visualization is achieved, i.e., particularly with 3d viewing (***building up lost parts of the body****)*, performance be impaired when two dimensions of object orientation must be considered simultaneously. To address these issues, we have used *(***multilevel Pixel alignment*** in our case) which uses an advanced algorithm and formulation to the ***obtained high-fidelity*** 3D reconstruction of clothed humans from a single image at a resolution sufficient to recover detailed information such as fingers, facial features, and clothing folds
4. To verify and validate the proposed system, the magnitude of the error and systematic error trends were used to make inferences about the underlying processes with each type of display. Finally, after constructing the 3d model to two images, one can compare them so that they can be used for scientific purposes and to fulfill plenty of applications.

The convenience of 3D-model today, such as TVs, Blu-Ray players, gaming consoles, and smartphones, is not yet matched by 3D content production. Today there exists an urgent need to convert the existing 2D content to 3D. Converting 2D to 3D models will not be scientifically true but yes, it will make an illusion from 2D to 3D on which things can far better be analyzed. So, in this research, our main objective is to convert a 2D image to a 3D model using ***multilevel pixel alignment*** and some of the valuable algorithms using python libraries playing an important role or will be playing an important role in the future in the following sectors: -

For the Industrial sector: -

With the increase of films released in 3D, 2D to 3D conversion has become more common. Most non-CGI stereo 3D blockbusters are converted fully or at least partially from 2D footage. Even Avatar contains several scenes shot in 2D and converted to 3D. The reasons for shooting in 2D instead of 3D are financial, technical, and sometimes artistic. Today on social media platforms we see several emojis as well as many 3d effects that are a small example of 2D to 3Dconversion.

For the educational sector: -

It is well known that pictures, diagrams, animated texts if displayed in 2D do not have a wide vision, and sometimes students do not find the things suitable to understand while being displayed in 2D, not only have a wide vision but also it helps students to understand the concepts efficiently. Therefore, in this research, we created new 3D content for education by applying 2D to 3D conversion. And for art education also it is very helpful to display the designs in 3D rather than 2D.

For the medical sector: -

For surgical education, the Conversion of 2D to 3D plays a vital role. Today every aspect of the medical field as well as machines used in the medical field use the concept of 2D to 3D conversion. For example: - XRAY, ECG etc. A 2D image is converted to a 3D model for scientific evaluations and then is analyzed on different parameters. Based on the colour-coding dataset given to us, we can also find the different types of acids or nutrients being present in our body and hence analyze them for scientific use.

5. METHODOLOGY

The following methodology will be followed to achieve the objectives defined for the proposed research work:

1. We have studied the 3D construction using Pixel alignment. We concluded that we can improve the 3D Shape from the image by using high-resolution images captured from the mobile phone. So, we used the Multilevel Pixel Alignment function. We used an image encoder for both the low-resolution and high-resolution levels to use a stacked hourglass network with 4 and 1 stacks respectively, using the modification suggested by and batch normalization replaced with group normalization.
2. We are using Python programming language and on that, we will be implementing the pose estimation and pixel alignment and we will be using different libraries like NumPy, scikit images, OpenCV, Pytorch, os, matplotlib, PIL, argparse, trimesh.
3. We have estimated the pose of a person using 2D People poses Estimation which detects a skeleton to identify human poses for every person inside the image. The pose may contain up to 18 key points: ears, eyes, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles. First, we assess the importance of 3D embedding that accounts for holistic context for high-resolution inference. To support larger input resolution at inference time, we train the network with random cropping of 512 × 512 from 1024 × 1024 images, like 2D computer vision tasks.
4. Second, we evaluate our design choice from both a robustness and fidelity perspective. To achieve high-resolution reconstruction, it is important to keep feature resolution large enough while maintaining the ability to reason holistically.

   We found that if our fine level module is conditioned on the absolute depth value instead of the learned 3D embedding, training with a sliding window significantly degrades both training and test accuracy.
5. Since the multilevel approach relies on the success of previous stages in extracting 3D embeddings, improving the robustness of our baseline model is expected to directly merit our overall reconstruction accuracy.

6. We qualitatively compare our method with state-of-the-art 3D human reconstruction methods with various shape representations on the publicly available People Snapshot dataset. The shape representations include a multi-scale voxel (DeepHuman), pixel-aligned implicit function (PIFu), and a human parametric model with texture mapping using displacements and surface normal (Tex2shape). While Tex2shape and DeepHuman adopt a coarse-to-fine strategy, the results show that the effect of refinement is marginal due to the limited representation power of the base shapes.

6. CONCLUSION AND DISCUSSION

So, in this research, our main objective is to convert a 2D image to a 3D model using Multilevel Pixel Alignment function and some of the valuable algorithms using python libraries playing an important role or will be playing an important role in the future in the following sectors:

1. Industrial Sector – For making 3D objects.
2. Educational Sector – For creating 3D animations.
3. Medical Sector – For visualizing the human body bone fractures.
4. IT Sector – For Game development.

Input Images: -

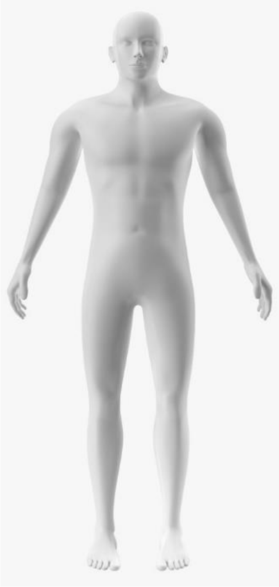

Fig. 3.0 Input Image without clothes on

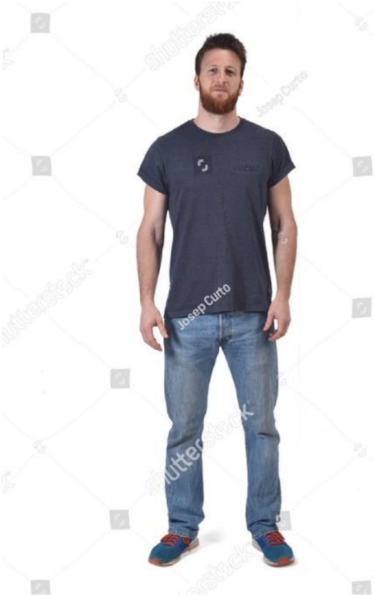

Fig. 3.1 Input Image With clothes

on

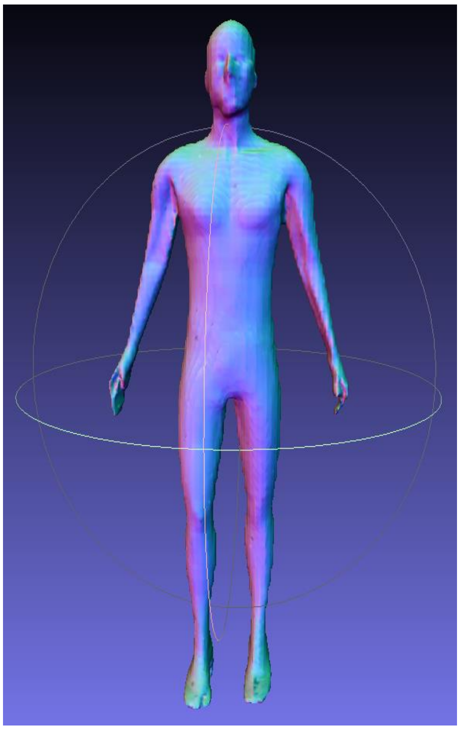

Figure 3.1.1 Front Portion

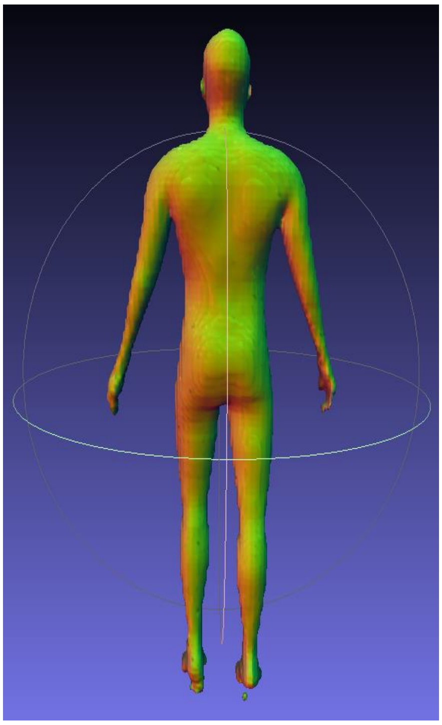

Figure 3.1.2 Back Portion

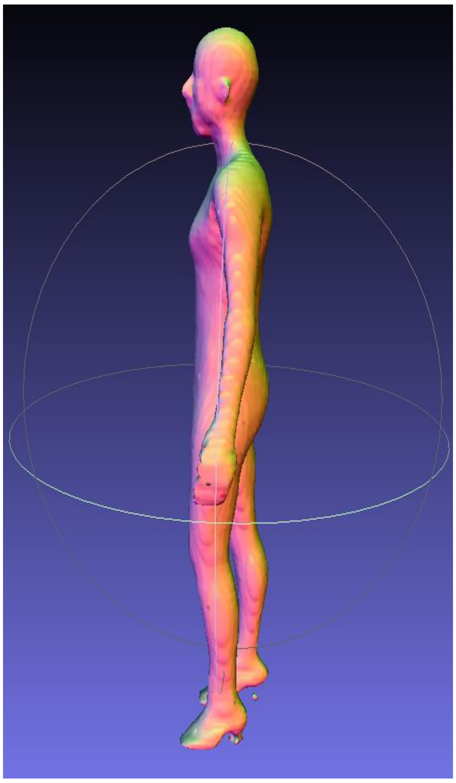

Figure 3.1.3 Left Portion

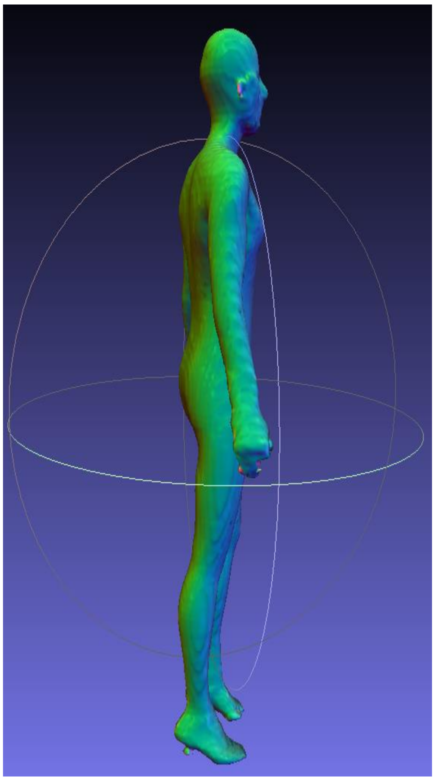

Figure 3.1.4 Right Portion

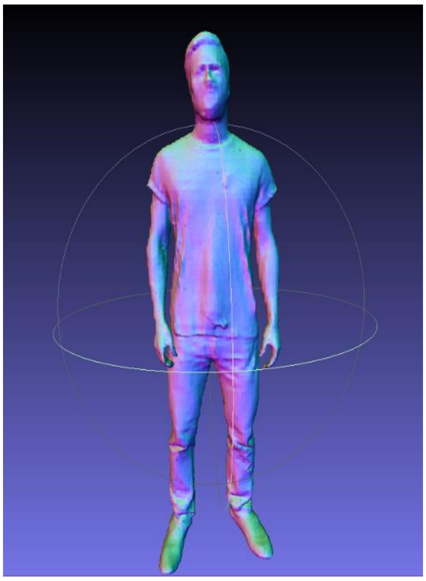

Figure 3.1.5 Front Portion

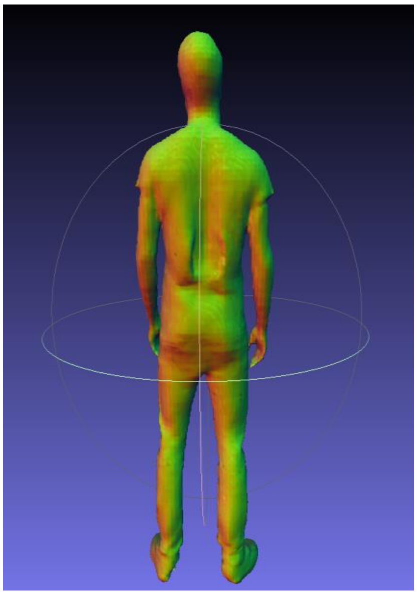

Figure 3.1.6 Back Portion

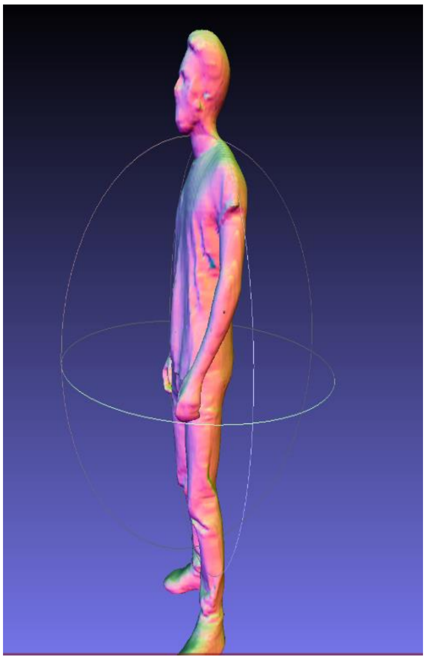

Figure 3.1.7 Left Portion

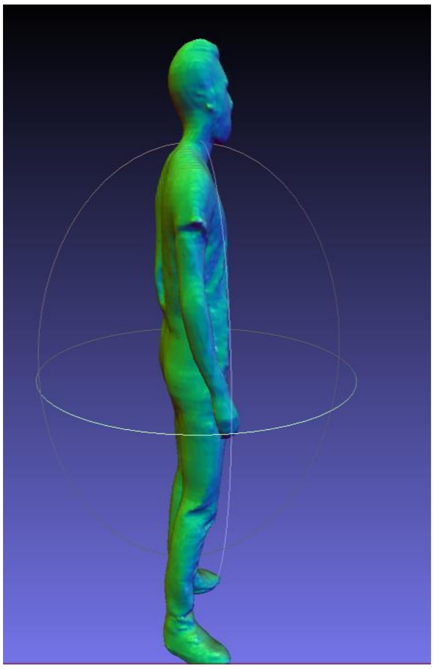

Figure 3.1.8 Right Portion

7. FUTURE SCOPE

Since the multilevel approach relies on previous stages in extracting 3D embeddings, improving our baseline model's robustness is expected to merit our overall reconstruction accuracy directly. Future work may include incorporating human-specific priors e.g., semantic segmentation, pose, and parametric 3D face models and adding 2D supervision of implicit surface to support in-the-wild inputs further. Furthermore, the model can be trained on more complex poses which include animal poses.

8. REFERENCES

[1] D. Rattan, R. Bhatia, and M. Singh, "Software clone detection: A systematic review,"
Information and Software Technology, vol. 55, no. 7, pp. 1165–1199, Jul. 2013.

[2] J. F. Islam, M. Mondal, and C. K. Roy, "Bug Replication in Code Clones: An Empirical Study," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2016, pp. 68–78.

[3] M. R. Islam and M. F. Zibran, "A Comparative Study on Vulnerabilities in Categories of Clones and Non-cloned Code," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2016, pp. 8–14.

[4] M. R. Islam, M. F. Zibran, and A. Nagpal, "Security Vulnerabilities in Categories of Clones and Non-Cloned Code: An Empirical Study," in *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2017, pp. 20–29.

[5] C. K. Roy, M. F. Zibran, and R. Koschke, "The vision of software clone management: Past, present, and future (Keynote Paper)," in *2014 Software Evolution Week - IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE)*, 2014, pp. 18–33.

[6] J. Krinke, "A Study of Consistent and Inconsistent Changes to Code Clones," in *14th Working Conference on Reverse Engineering (WCRE 2007)*, 2007, pp. 170–178.

[7] D. Chatterji, J. C. Carver, N. A. Kraft, and J. Harder, "Effects of cloned code on software maintainability: A replicated developer study," in *2013 20th Working Conference on Reverse Engineering (WCRE)*, 2013, pp. 112–121.

[8] D. Rattan, R. Bhatia, and M. Singh, "An Empirical Study of Clone Detection in MATLAB/ Simulink Models," *International Journal of Information and Communication Technology*.

[9] D. Rattan, R. Bhatia, and M. Singh, "Detecting High-Level Similarities in Source Code and Beyond," *International Journal of Energy, Information and Communications*, vol. 6, no. 2, pp. 1–16, 2015.

[10] D. Rattan, R. Bhatia, and M. Singh, "Detection and Analysis of Clones in UML Class Models," *International Journal of Software Engineering, IJSE*, vol. 8, no. 2, pp. 66– 99, 2015.

[11] D. Rattan, R. Bhatia, and M. Singh, "Model clone detection based on tree comparison," in *2012 Annual IEEE India Conference (INDICON)*, 2012, pp. 1041– 1046.

[12] C. K. Roy and J. R. Cordy, "NICAD: Accurate Detection of Near-Miss Intentional Clones Using Flexible Pretty-Printing and Code Normalization," in *2008 16th IEEE International Conference on Program Comprehension*, 2008, pp. 172–181.

[13] M. Mondal, C. K. Roy, and K. A. Schneider, "SPCP-Miner: A tool for mining code clones that are important for refactoring or tracking," in *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2015, pp. 484–488.

[14] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4):107, 2017.

[15] A. S. Jackson, C. Manafas, and G. Tzimiropoulos. 3D Human Body Reconstruction from a Single Image via Volumetric Regression. In ECCV Workshop Proceedings, PeopleCap 2018, pages 0–0, 2018.

[16] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-toend recovery of human shape and pose. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7122– 7131, 2018.

[17] V. Lazova, E. Insafutdinov, and G. Pons-Moll. 360-degree textures of people in clothing from a single image. In International Conference on 3D

Vision (3DV), sep 2019.

[18] S. Liu, S. Saito, W. Chen, and H. Li. Learning to infer implicit surfaces without 3d supervision. arXiv preprint arXiv:1911.00767, 2019.

[19] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In ACM siggraph computer graphics, volume 21, pages 163–169. ACM, 1987.

[20] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision, pages 483–499, 2016.

[21] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10975–10985, 2019.

[22] Renderpeople, 2018. https://renderpeople.com/ 3d-people.

[23] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for highresolution clothed human digitization. In ICCV, 2019.

[24] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, pages 1576–1585, 2017.

[25] S. Tang, F. Tan, K. Cheng, Z. Li, S. Zhu, and P. Tan. A neural network for detailed human depth estimation from a single image. In Proceedings of the IEEE International Conference on Computer Vision, pages 7750–7759, 2019.

[26] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In IEEE Conference on Computer Vision and Pattern Recognition, pages 4191–4200, 2017.

[27]] V. Sitzmann, M. Zollhofer, and G. Wetzstein. Scene represen- ¨ tation

networks: Continuous 3d-structure-aware neural scene representations. arXiv preprint arXiv:1906.01618, 2019.

[28] P.-P. Sloan, J. Kautz, and J. Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In ACM Transactions on Graphics, volume 21, pages 527–536, 2002.

[29] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero. Facsimile: Fast and accurate scans from an image in less than a second. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[30] S. Tang, F. Tan, K. Cheng, Z. Li, S. Zhu, and P. Tan. A neural network for detailed human depth estimation from a single image. In Proceedings of the IEEE International Conference on Computer Vision, pages 7750–7759, 2019.

[31] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions.