INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

SYSTEMATIC REVIEW

3D Human Reconstruction from A Single Image Using Parametric Model-Conditioned Implicit Representation

E Shanmuga Priya[1]*, P Velvizhy[1], Arul Deepa[2]

1 Department of CSE, College of Engineering, Anna University, Guindy, Chennai, India
2 Department of IST, College of Engineering, Anna University, Guindy, Chennai, India

OPEN ACCESS

Received: 30-06-2023
Accepted: 13-10-2023
Published: 26-11-2023

Citation: Priya ES, Velvizhy P, Deepa A (2023) 3D Human Reconstruction from A Single Image Using Parametric Model-Conditioned Implicit Representation. Indian Journal of Science and Technology 16(44): 4054-4062. https://doi.org/10.17485/IJST/v16i44.1618

* **Corresponding author**.

vijaylaya2000@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: PaMIR is a novel approach for image-based human reconstruction that utilizes a parametric model-conditioned implicit representation. This method enables the generation of a complete 3D mesh of a human body from a single input image. It uses a neural network that is conditioned on a parametric model of the human body to produce an implicit representation of the 3D surface. **Objectives:** To develop a novel approach for image based human reconstruction by training neural network and to generate high quality images. **Method:** In our PaMIR-based reconstruction framework, a novel deep neural network is proposed to regularize the free-form deep implicit function using the semantic features of the parametric model, which improves the generalization ability under the scenarios of challenging poses and various clothing topologies. **Findings:** The quantitative comparison shows that PaMIR method outperforms the state-of-the-art methods in terms of surface reconstruction accuracy. The errors are also provided when ground-truth SMPL annotations are available to present the upper limit of the reconstruction accuracy if the SMPL estimation is perfect. Overall, this method is more general, more robust and more accurate than HMD, Molding Humans, Deep Human and PIFu. **Novelty:** A novel depth-ambiguity-aware training loss is further integrated to resolve depth ambiguities and enable successful surface detail reconstruction with imperfect body reference. Finally, we propose a body reference optimization method to improve the parametric model estimation accuracy and to enhance the consistency between the parametric model and the implicit function. With the PaMIR representation, our framework can be easily extended to multi-image input scenarios without the need of multi-camera calibration and pose synchronization.

Keywords: Parametric; framework; NonParametric; HMD; SMPL; Tex2Shape

1 Introduction

Image based parsing of human bodies is a popular topic in computer vision and computer graphics. Among all the tasks of image-based human parsing, recovering 3D humans from a single RGB image attracts more and more interests given its wide applications in Virtual Reality/Augmented Reality (VR/AR) content creation, image and video editing, telepresence, and virtual dressing[1]. However, as an ill-posed problem, recovering 3D humans from a single RGB image is very challenging due to the lack of depth information and the variations of body shapes, poses, clothing types and lighting conditions.

The 3D representation methods can be roughly classified into two categories: parametric methods and non-parametric methods[2]. However, the low dimensional parametric model limits the performance when handling different clothing types such as long dresses and skirts[3]. To achieve high-quality geometric detail reconstruction while maintaining robustness to challenging pose and clothing styles[4]. In this paper, we propose Parametric Model-Conditioned Implicit Representation, dubbed PaMIR, to incorporate the parametric SMPL model and the free-form implicit surface function into a unified learning and optimization framework[5]. The implicit surface function overcomes the resolution limit of volumetric representation and enables detailed surface reconstruction capability[1]. Although being able to describe arbitrary clothes, current non-parametric methods still suffer from challenging poses and self-occlusions due to the lack of semantic information and depth ambiguities[6].

To fill the accuracy gap between the ground truth training data and the inaccurate testing input, we further propose a new depth ambiguity-aware training loss and a body reference optimization module in the PaMIR based framework and achieve surface detail reconstruction even under imperfect body reference initialization and depth ambiguity.

This work is to develop a method for generating a complete 3D mesh of a human body from a single input image[7]. This involves the use of a parametric model of the human body to provide prior knowledge about body shape, pose, and texture, which is then used by a neural network to produce an implicit representation[8] [9].

The main objectives of PaMIR are:

- To develop a novel approach for image-based human reconstruction that uses a parametric model-conditioned implicit representation.
- To train a neural network that can generate accurate and realistic 3D models of human bodies from 2D images.
- To incorporate a shape refinement module that allows for the generation of high-quality, detailed meshes with sharp edges and fine features.
- To evaluate the performance of PaMIR on several benchmark datasets and demonstrate that it outperforms existing state-of-the-art methods in terms of accuracy and visual quality.

In contrast, we proposed a deep learning-based framework to combine the parametric SMPL model and the non-parametric deep implicit function for 3D human model reconstruction from a single RGB image[10]. Benefiting from the proposed PaMIR representation, the depth-ambiguity-aware reconstruction loss and the reference body optimization algorithm, our method outperforms state-of-the-art methods in terms of both robustness and surface details.

2 Methodology

The proposed system architecture is shown in Figure 1. It has the following modules.

2.1 Non Parametric Shape Estimation

The first part of the pipeline consists of a typical image-based CNN following the ResNet-50 architecture as showed in Figure 2. From the original design ignore the final fully connected layer, keeping only the feature vector after the average pooling layer. This CNN is used as a generic feature extractor from the input representation[11]. Having the feature vector extracted by the generic image-based network, attach these features to the 3D coordinates of each vertex in the template mesh[12]. Graph CNN uses as input the 3D coordinates of each vertex along with the input features and has the goal of estimating the regressed 3D vertex coordinates. This processing is performed by a series of Graph Convolution layers.

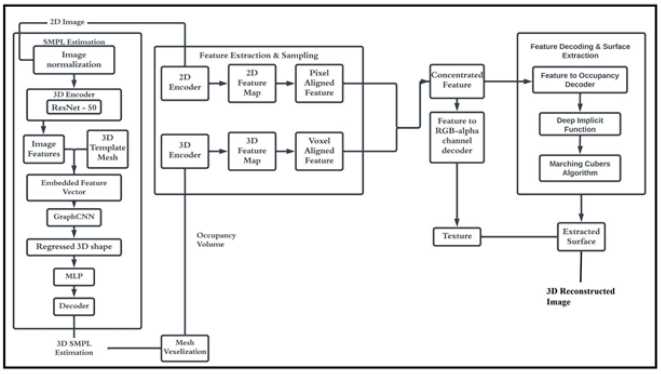

Fig 1. Proposed System Architecture

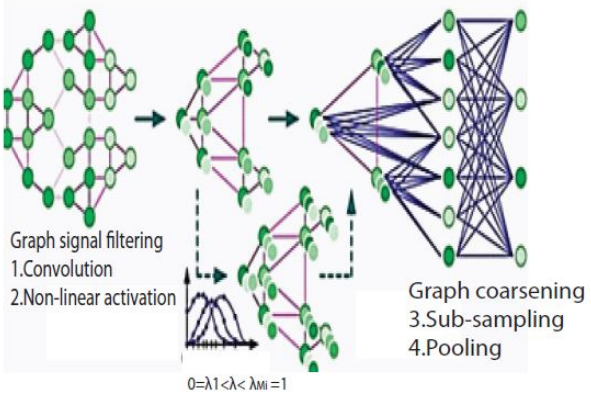

Fig 2. Graph convolutional Neural Network Algorithm

Algorithm 1: Graph Convolution Neural Network

Input: 2D Image
Output: Regressed Shape
Function def graph_CNN(self, image):
batch_size = image.shape[0]
ref_vertices = self.ref_vertices[None, :, :].expand(batch_size, -1, -1)
image_resnet = self.resnet(image)
image_enc=image_resnet.view(batch_size, 2048, 1).expand(-1,ref_vertices.shape[-1])
x=torch.cat([ref_vertices,image_enc], dim=1)
x = self.gc(x)
shape = self.shape(x)
camera=self.camera_fc(x).view(batch_size, 3)
return shape, camera

2.2 Parametric SMPL Estimation

Multi-Layer Perceptron network is trained to regresses pose (θ) and shape (β)parameters of the Skinned Multi-Person Linear parametric model given the regressed 3D vertex coordinate as input. Decoder is used to convert the output of Multi-Layer Perceptron to SMPL parametric shape.

Algorithm 2: MLP Algorithm

Input: Regressed 3D Shape
Output: SMPL Mesh
Function def SMPL(self, x):
batch_size = x.shape[0]
x = x.view(batch_size, -1)
x = self.layers(x)
rotmat = rotmat * det
rotmat = rotmat.view(batch_size, 24, 3, 3)
rotmat = rotmat.to(orig_device)
return rotmat, betas

2.3 Feature Extraction and Sampling

In the feature extraction step, the input image is encoded into a feature map by a 2D convolution network, while the occupancy volume is encoded into a feature volume by a 3D convolution network[13]. The occupancy volume is obtained by converting SMPL model through voxelization. For each point in the 3D space, pixel-aligned image feature and voxel-aligned volume feature are sampled in the feature map and the feature volume, respectively. Then the two feature vectors are then concatenated.

The quality of the occupancy volume will depend on the voxel size, with smaller voxels resulting in higher resolution but larger files. Additionally, some voxelization algorithms may introduce errors or artifacts in the voxelized mesh, which can affect the accuracy of the occupancy volume. Therefore, it is important to choose an appropriate voxelization algorithm and voxel size based on your specific application requirements.

Algorithm 3: Feature Extraction and sampling

Input: 2D image & 3D SMPL mesh
Output: Concatenated Features
Function def concat_feature(self, img, vol, pts, pts_proj,attention_net=None):
pt_sdf = pt_output.view([1, point_num, 1])
pt_sdf_list.append(pt_sdf)
return pt_sdf_list

2.4 Surface Extraction

The concatenated feature is translated to an occupancy probability value. Feature-to-occupancy decoder is used to find occupancy probability value. Learn an implicit function F(C(p)) that can classify whether p is inside or outside the surface, by sampling the occupancy probability field over the 3D space and extract the iso- surface of the probability field at threshold 0.5 using the Marching Cubes algorithm.

Algorithm 4: Surface Extraction
Input: Concatenated feature.
Output: : Extracted surface of an image
Function def test_pifu(self, img, vol_res, betas, pose, scale, trans):
mesh = dict()
mesh['v'] = vertices / vol_res - 0.5
mesh['f'] = simplices[:, (1, 0, 2)]
mesh['vn'] = normals
return mesh

- **Step1: Determination of case index**

The eight voxels of the volume cell have an assigned state, indicating whether their respective voxel value is greater, equal, or smaller than the specified iso-value. The state of each voxel is interpreted as a binary digit (1 for inside, and 0 for outside) and composed into an eight-bit number index. Figure 3 shows which voxel state goes to which position.

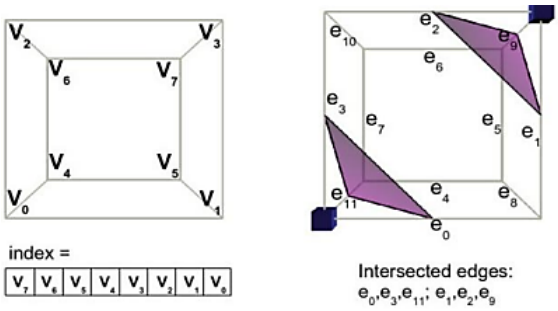

Fig 3. Determining case Index

- **Step 2: Determination of intersected edges**

After the case index for the cell is determined, the cell edges that are intersected by the iso-surface can be looked up in the case table. In (right), six cell edges are intersected by the iso-surface. As an example, in case 9 (9=1*20+1*23), the states of voxel 0 and voxel 3 are set. Now have 256 possible configurations of voxel states, and hence there will be 256 possible triangulations of a volume cell. However, not all these configurations generate different triangulations. Most cases can be sorted into 15 equivalence classes, considering rotation or mirroring.

- **Step 3: Computation of Intersections**

Once the intersected cell edges are determined, the intersection itself is calculated by linear interpolation as shown in Figure 4.

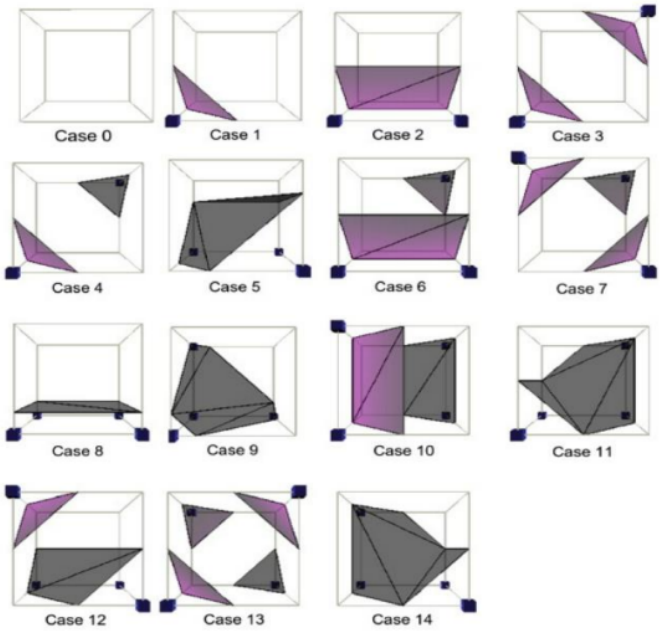

Fig 4. Determining Intersected edged

- **Step 4: Triangulation of the Intersections**

Then by combining the interpolated intersections with the information, how the edge intersections are composed into triangles will be known, which is also stored in the case table. With that, the triangles can be determined.

- **Step 5: Computation of Outward-pointing Surface Normal**

Usually, surface normal are determined based on the orientation of the triangle. However, in the context of surface extraction, the gradients of the scalar field might be approximated by computing intensity differences. The use of the original data of the scalar field leads to a more precise definition of surface normal. This computation is performed for the three vertices of a triangle and as a simple solution it is averaged. The resulting gradient needs to be normalized for the use in an illumination model. The very same strategy is used for the integration of illumination in direct volume rendering.

2.5 Texture Reconstruction

To perform texture inference, make some simple modification to the network as in geometry inference. Then define the output of the decoder as an RGBα vector field instead of a scalar field. In texture inference, the output of the decoder is taken as RGBα vector field. The RGB value is the network prediction of the color of a specific point on the mesh surface, while the alpha channel is used to blend the predicted value with the observed one [(14)]. The texture network first recovers the color on the whole surface as well as an alpha channel which is used to blend the predicted color with the observation. The extracted surface combines with texture of an image to produce an output of 3D reconstructed image.

- **Algorithm 5: Texture Reconstruction**

Input: Concatenated Feature
Output: Photorealistic 3D Reconstructed Image.
Function
defforward_infer_color_value_group(self, img, vol, pts, pts_proj, group_size):
pts_clr = np.concatenate(pts_clr)
pts_clr = np.array(pts_clr)
return pts_clr

3 Results and Discussion

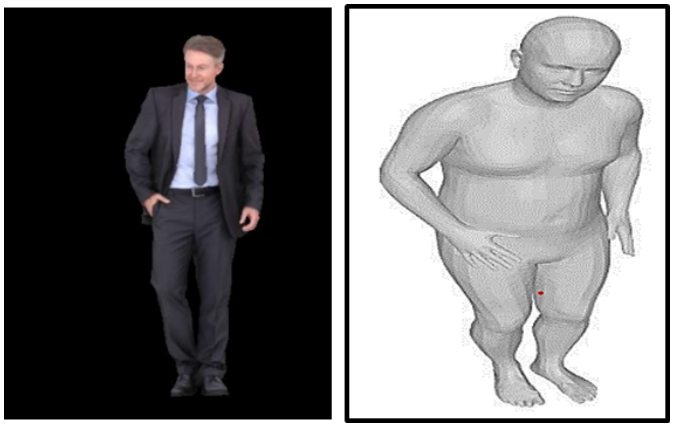

Fig 5. 3D model of the human body

This method normalizes images according to the mean and standard deviation values specified for the ResNet model. Then, it normalizes the images. Finally, it returns the normalized images. The ResNet architecture is composed of multiple ”blocks” that use the Bottleneck block [(15)]. It defines a Mesh class used for handling certain graph operations. The method takes an input image and a set of reference vertices and encodes them using a ResNet architecture. The encoded image is then concatenated

with the reference vertices. The reference vertices and these feature vectors are then concatenated and passed through a series of GraphResBlock layers, which are used to update the feature vectors at each vertex based on the features of its neighbors.

The output of the GCNN is then passed through a sequence of fully connected layers to predict the 3D shape and camera parameters of the object in the input image[16]. This is a neural network model that takes in some input and outputs the parameters of the SMPL (Skinned Multi-Person Linear) body model. The SMPL model is a 3D model of the human body that represents body shape and pose using a linear combination of learned basis shapes and joint rotations as shown in Figure 5 .

This method implements a forward pass of a neural network that predicts the rotation matrix and shape parameters of the SMPL model given an input feature vector. Then other function performs an optimization to refine the SMPL parameters by minimizing the difference between the projected 3D keypoints and the detected 2D keypoints, as well as the difference between the predicted SDF and the ground truth SDF.

An instance of HourglassNet, with input channel count of 4, output channel count of 128, and number of stages of 4.This is the definition of an MLP (Multi-Layer Perceptron) network, which is a type of feedforward neural network that consists of multiple layers of fully connected linear units (called neurons) followed by nonlinear activation functions.

This is similar to shape reconstruction, but only the output of MLP network changes to produce scalar field instead of vector field. The 2D and 3D features are concatenated and passed through a multi-layer perceptron (MLP) to predict the texture of each point in the point cloud. The predicted texture is a 3-channel RGB value, while the attention weights are a single scalar value.

3.1 Performance Metrics

- **Body Fitting Loss**

Table 1. Body Fitting Loss values at specific steps

Epochs	Body fitting loss
0	0.297105
15	0.270218
25	0.256969
49	0.236855

- **Depth-ambiguity-aware Reconstruction Loss**

Reconstruction loss is defined as the mean Square error between the occupancy probability of the point samples and the ground-truth ones. Predicted models may not be well aligned with the ground-truth along the z-axis. To deal with this issue, here novel depth ambiguity aware reconstruction loss is proposed. This can be calculated using (1). Δpi = (0, 0, Δzi) is the compensating translation along z-axis.

$$L_R = \frac{1}{n_p}\sum_{i=1}^{n_p} |F(C(p_i + \Delta p_i)) - F*(p_i)|^2 \quad (1)$$

- **Mean Per Joint Position Error**

Mean Per Joint Position Error is the average Euclidean distance between the location of real-life joints on human bodies and the location of predicted joints on 3D pose model. For a frame f and skeleton S, MPJPE can be calculated using (2).

$$E_{MPJPE}(f, S) = \frac{1}{N_S}\sum_{i=1}^{N_S} \left\| m_{f,S}^{(f)}(i) - m_{gt,S}^{(f)}(i) \right\|_2 \quad (2)$$

- **Comparative Analysis**

By qualitatively comparing PaMIR method with several state-of-the-art methods including HMD, Tex2Shape, Moulding Humans, DeepHuman and PIFu. Among them, HMD and Tex2Shape are parametric methods based on SMPL model deformation, PIFu uses a deep implicit function as geometry representation, Moulding Humans uses the combination of a

front depth map and a back depth map for representation, and Deep Human combines volumetric representation with the SMPL model.

In this experiments, DeepHuman, PIFu and Moulding Humans are all retrained on the dataset, while parametric methods like HMD and Tex2shape are not because the dataset contains loose garments like dresses which will deteriorate the performance of those methods.

By quantitatively comparing PaMIR method with the state-of-the-art methods using both Twindom testing dataset and BUFF rendering dataset to evaluate the geometry reconstruction accuracy. Similar to the experiments in PIFu, point-to-surface error as well as the Chamfer distance is used as error metric.

This method needs high-quality human scans for training. However, it is highly costly and time consuming to obtain a large-scale dataset of high-quality human scans. Moreover, the currently available scanners for human bodies require the subject to keep static poses in a sophisticated capturing environment, which makes them incapable to capture real-world human motions in the wild and consequently the training data is biased towards simple static poses like standing [(17)] [(18)]. Although the proposed method already makes a step forward in terms of generalization capability, it still fails in the cases of extremely challenging poses. One important future direction is to alleviate the reliance on ground-truth by exploring large-scale image and video dataset for unsupervised training.

The quantitative comparison shows that PaMIR method outperforms the state-of the-art methods in terms of surface reconstruction accuracy. The errors are also provided when ground-truth SMPL annotations are available to present the upper limit of the reconstruction accuracy if the SMPL estimation is perfect. Overall, this method is more general, more robust and more accurate than HMD, Moulding Humans, DeepHuman and PIFu [(6)].

Table 2. Numerical comparison results

Dataset	**BUFF**		**TWINDOM**	
Method	P2S (Point to Surface) (cm)	Chamfer (cm)	P2S (Point to Surface) (cm)	Chamfer (cm)
HMD	2.48	3.92	2.50	4.01
Moulding Humans	2.25	2.68	2.84	3.35
DeepHuman	2.15	2.80	2.35	2.97
PIFu	1.93	2.22	2.34	2.65
PaMIR Method	1.52	1.92	1.80	2.12
PaMIR Method using groundtruth SMPL	0.709	0.936	0.744	1.00

4 Conclusion

A deep learning-based framework for combining the parametric SMPL model and the non-parametric deep implicit function for 3D human model reconstruction from a single RGB image is proposed in this work. A temporally consistent reconstruction results by applying PaMIR method to video frames individually is obtained. This method will enable many applications and further stimulate the research in this area. From a more general perspective of view, this approach made a step forward towards integrating semantic information into the free-form implicit fields which have attracted more and more attention from the research community for its flexibility, representation power and compact nature. Moreover, the semantic implicit functions can also be adopted in research on 3D perception and parsing. This method needs high-quality human scans for training. However, it is highly costly and time consuming to obtain a large-scale dataset of high-quality human scans. Moreover, the currently available scanners for human bodies require the subject to keep static poses in a sophisticated capturing environment, which makes them incapable to capture real-world human motions in the wild and consequently the training data is biased towards simple static poses like standing. Although the proposed method already makes a step forward in terms of generalization capability, it still fails in the cases of extremely challenging poses. One important future direction is to alleviate the reliance on ground-truth by exploring large-scale image and video dataset for unsupervised training.

References

1) Zheng Z, Yu T, Liu Y, Dai Q. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(6):3170–3184. Available from: https://doi.org/10.1109/TPAMI.2021.3050505.

2) Chen L, Peng S, Zhou X. Towards efficient and photorealistic 3D human reconstruction: A brief survey. *Visual Informatics*. 2021;5(4):11–19. Available from: https://doi.org/10.1016/j.visinf.2021.10.003.
3) Saxena A, Sun M, Ng AY. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*. 2021;31:824–840. Available from: https://doi.org/10.1109/TPAMI.2008.132.
4) Priya ES, Velvizhy P, Deepa KA. Neural Style Transfer with distortion handling for Audio and Image. *International Conference on Data Science*. 2022;p. 1–10. Available from: https://doi.org/10.1109/ICDSAAI55433.2022.10028972.
5) Chen Z, Zhang H. Learning Implicit Fields for Generative Shape Modeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019;p. 5932–5941. Available from: https://doi.org/10.1109/CVPR.2019.00609.
6) Zheng Z, Yu T, Wei Y, Dai Q, Liu Y. DeepHuman: 3D human reconstruction from a single image. *Proceedings of the IEEE International Conference on Computer Vision*. 2021;p. 7738–7748. Available from: https://doi.org/10.48550/arXiv.1903.06473.
7) Chang Z, Koulieris GA, Shum HPH. 3D Reconstruction of Sculptures from Single Images via Unsupervised Domain Adaptation on Implicit Models. In: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology;vol. 26. ACM. 2022;p. 1–10. Available from: https://doi.org/10.1145/3562939.3565632.
8) Jiang C, Sud A, Makadia A, Huang J, Niebner M, Funkhouser T. Local Implicit Grid Representations for 3D Scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020;p. 6000–6009. Available from: https://doi.org/10.1109/CVPR42600.2020.00604.
9) Guler RA, Kokkinos I. HoloPose: Holistic 3D Human Reconstruction In-The-Wild. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019;p. 10876–10886. Available from: https://doi.org/10.1109/CVPR.2019.01114.
10) Chibane J, Alldieck T, Pons-Moll G. Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020;p. 6968–6979. Available from: https://doi.org/10.1109/CVPR42600.2020.00700.
11) Gabeur V, Franco JS, Martin X, Schmid C, Rogez G. Moulding humans: Non-parametric 3D human shape estimation from single images. *arxiv*. 2021;p. 2232–2241. Available from: https://arxiv.org/pdf/1908.00439.pdf.
12) Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. Occupancy Networks: Learning 3D Reconstruction in Function Space. . Available from: https://doi.org/10.1109/CVPR.2019.00459.
13) Kolotouros N, Pavlakos G, Black M, Daniilidis K. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019;p. 2252–2261. Available from: https://doi.org/10.1109/ICCV.2019.00234.
14) Velvizhy P, Pravi A, Selvi M, Ganapathy S, Kannan A. Fuzzy-based review rating prediction in e-commerce. 2020. Available from: https://doi.org/10.1504/IJBIDM.2020.108034.
15) Gwak J, Christopher B, Choy D, Xu K, Chen S, Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. 2020. Available from: https://doi.org/10.1007/978-3-319-46484-8_38.
16) Deepa KA, Priya S, Velvizhy P. KNN Based Under Sampling: A Cognitive Centred Solution for Imbalanced Dataset Problem in Anaphora Resolution. . Available from: https://doi.org/10.17485/IJST/v16i30.1523.
17) Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. *Computer Vision and Pattern Recognition*. 2020. Available from: https://doi.org/10.48550/arXiv.1607.08128.
18) Kocabas M, Athanasiou N, Black MJ. VIBE: Video inference for human body pose and shape estimation. 2020. Available from: https://doi.org/10.48550/arXiv.1912.05656.