3D Magic Mirror: Clothing Reconstruction from a Single Image via a Causal Perspective

Zhedong Zheng[1,2] Jiayin Zhu[1,2] Wei Ji[1,2] Yi Yang[3] Tat-Seng Chua[1,2]
[1]Sea-NExT Joint Lab, Singapore [2] School of Computing, National University of Singapore [3]Zhejiang University
{zdzheng,jiwei,dcscts}@nus.edu.sg,zhujiayin@u.nus.edu,yiyangcs@zju.edu.cn

ABSTRACT

This research aims to study a self-supervised 3D clothing reconstruction method, which recovers the geometry shape, and texture of human clothing from a single 2D image. Compared with existing methods, we observe that three primary challenges remain: (1) the conventional template-based methods are limited to modeling non-rigid clothing objects, *e.g.*, handbags and dresses, which are common in fashion images; (2) 3D ground-truth meshes of clothing are usually inaccessible due to annotation difficulties and time costs. (3) It remains challenging to simultaneously optimize four reconstruction factors, *i.e.*, camera viewpoint, shape, texture, and illumination. The inherent ambiguity compromises the model training, such as the dilemma between a large shape with a remote camera or a small shape with a close camera.

In an attempt to address the above limitations, we propose a causality-aware self-supervised learning method to adaptively reconstruct 3D non-rigid objects from 2D images without 3D annotations. In particular, to solve the inherent ambiguity among four implicit variables, *i.e.*, camera position, shape, texture, and illumination, we study existing works and introduce an explainable structural causal map (SCM) to build our model. The proposed model structure follows the spirit of the causal map, which explicitly considers the prior template in the camera estimation and shape prediction. When optimization, the causality intervention tool, *i.e.*, two expectation-maximization loops, is deeply embedded in our algorithm to (1) disentangle four encoders and (2) help the prior template update. Extensive experiments on two 2D fashion benchmarks, *e.g.*, ATR, and Market-HQ, show that the proposed method could yield high-fidelity 3D reconstruction. Furthermore, we also verify the scalability of the proposed method on a fine-grained bird dataset, *i.e.*, CUB.

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**.

KEYWORDS

3D Mesh Reconstruction, Fashion Generation, Expectation Maximization, Causality, Self-supervised Learning.

1 INTRODUCTION

Nowadays, online shopping become popular [29]. People can purchase clothing via online shopping sites, *e.g.*, Amazon and eBay. However, there remains a gap between the display images and the real product quality [6]. In an attempt to minimize such a visualization gap, we study the 3D clothing reconstruction from a single image in this paper. Given a 2D clothing image and the foreground mask, we intend to reconstruct a 3D mesh, which recovers the geometry shape, and texture of the target clothing. Besides, the clothing reconstruction can also be applied to many computer vision applications, including virtual reality [43], interactive system [39, 42] and 3D printing [3].

Figure 1: Motivation. Here we compare the proposed approach with prevailing template-based methods, *i.e.*, HMR [17] and ROMP [44] on a fashion dataset ATR [24]. We re-implement and visualize results with the color mapping according to the projected location. The first row is the front view, and the second row is the 3D mesh rotated with 45°. The template-based model can capture the human poses but miss non-rigid objectives, such as hairs, handbags and dresses.

However, there remain three challenges. First, existing works [17, 26, 27] typically focus on human pose estimation and body reconstruction via parametric models, *e.g.*, a morphable body template [31]. However, pre-defined body parameters usually are not scalable from the human body to the non-rigid clothing, *e.g.*, dresses and handbags, largely losing fine-grained clothing details. As shown in Figure 1, we re-implement two prevailing methods, *i.e.*, HMR [17] and ROMP [44], which both successfully capture the human pose but miss the cloth shape. In contrast, our methods leverage a deformable model to further facilitate learning non-rigid objects.

Second, 3D annotations are difficult to obtain due to the annotation difficulty and time costs. There are no public large-scale 3D clothing mesh datasets for supervised learning. In contrast, the availability of the large-scale 2D fashion datasets, such as ATR [24] and Market-1501 [56], makes training data-hungry deep-learned approaches become feasible. The success has been proved in the 2D computer vision tasks, including human parsing [16, 25, 30, 41], person re-ID [14, 50, 51], attribute recognition [9, 28, 47] and pedestrian image generation [10, 32, 38, 58]. With the recent development in self-supervised learning and deep-learned models, one straightforward idea is raised whether to leverage the 2D data for 3D reconstruction, even without the 3D annotations. This idea inspires us to explore the feasibility of self-supervised learning with causality design. The causality-aware structure helps the prior knowledge learning (*e.g.*, learning the human prototype) during training.

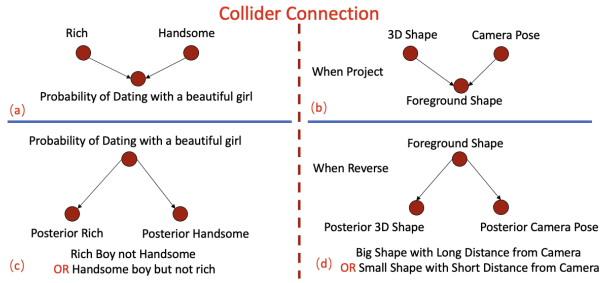

Figure 2: Explanation of the Collider Connection. Here we also show the common example "Dating with Beautiful Girl" from The Book of Why [36] in (a), which is similar to our simplified dilemma on two variants, *i.e.*, shape and camera pose in (b). Since we study an inverse problem, we also draw (c) and (d). We note that there is a compensation effect when we estimate the posterior probability. As shown in (d), given the observed foreground shape, the model needs to estimate the posterior 3D shape and posterior camera pose. There are two possible alternatives for the network to learn, *i.e.*, a big shape from a long distance or a small shape from a short distance. Therefore, it makes the model difficult to converge an answer.

Third, one scientific question still remains in the single image 3D reconstruction. The primary implicit variables for reconstruction are camera viewpoint, shape, texture, and illumination. Ideally, these four factors are independent of each other. However, it remains challenging to disentangle the implicit variables in practice. During implementation, one typical dilemma is the ambiguity between the camera and shape [23]. For example, given one 2D image, it is hard to decide the size of the target. There are two possible answers. One large object is far from the camera or one small object is close to the camera. Despite different original sizes, the two objects have the same projection size in the photos. In this paper, we introduce the structure causal map [35], which is widely used in causal inference. It helps us to identify the problem as a "collider" connection and we show a simplified sample of two variants in Figure 2. Following the spirit of the causal map (Figure 3), we discuss the failure reason and leverage the “intervention” tool, *i.e.*, two expectation-maximization loops, to avoid this "collider" dilemma. (1) **Encoder Loop.** A key underpinning the factor disentanglement is using independent encoders for different implicit variables. The weights of encoders are independent of each other. Despite the disentangle design in the network forwarding, we observe that, if the reconstruction fails, all four encoders are usually penalized equally. It largely impacts the single encoder training. The phenomenon that the loss of the "result" impacts the "cause" encoder is the compensation effect in "collider". For instance, even if the texture encoder is satisfied, the reconstruction failure in the 3D shape also leads to a large punishment on the texture encoder. Therefore, in this work, we propose an expectation-maximization loop to “free” one factor while "intervening" other factors. We first obtain the three out of four factor expectation, and then maximize the likelihood of the rest one factor. In this way, we find that the loss propagation can effectively penalize the worst encoder. It also helps the encoder disentanglement from the loss back-propagation aspect. (2) **Prototype Loop.** We introduce a fixed-size prototype shape as previous works [13, 21, 23] and update the prototype shape during training. The general assumption is that the shapes of different instances are close to the prototype shape. Therefore, the shape changes are limited around the prototype. By the prototype design, we physically free the shape encoder from estimating the vertex movement due to object size changes. Therefore, the camera encoder is fully responsible to estimate the scale changes by adjusting the camera distance prediction. We can regard the prototype template as a latent variable. During training, the prototype is generated from the average output shape in the whole dataset, which can be viewed as an expectation step (E-step). Then updating both the shape encoder and camera encoder is a maximization step (M-step), which maximizes the parameter likelihood based on the prototype expectation. This loop further disentangles the shape encoder and camera encoder and makes the shape encoder focus on learning intra-class variants, *e.g.*, instance deformation, rather than the global camera distance changes.

In an attempt to overcome the above-mentioned challenges, this paper proposes Causality-aware Self-supervised Learning (CASL) to adaptively reconstruct the input 2D clothing image without the 3D mesh annotation. We study the existing works (see Figure 3) and introduce an explainable structural causal map (SCM) to build our model and guide the optimization strategy. (1) Following the spirit of the causal map, we deploy four independent 3D attribute encoders and a differentiable render for reconstruction. The encoders are to extract 3D attributes from 2D images and foreground masks, including camera viewpoint, geometric shape, texture, and illumination. Then the attributes are fed to the differentiable render to reconstruct the 3D mesh. Different from existing works [13, 23], we explicitly introduce the prior template to help the camera estimation and shape offsets estimation, which is aligned with human observation. If we foreknow the human or bird prototype, it helps us to predict the camera position as well as the intra-class variant (such as leg movement). (2) We leverage the causality intervention tool, *i.e.*, two Expectation-Maximization Loops, to help the prototype learn and disentangle encoders from the confusing loss punishment. Extensive experiments verify the effectiveness of the proposed method on clothing reconstruction. Besides, we observe that the proposed method also has the potential to improve the reconstruction performance of general objects, like birds in the fine-grained CUB dataset [46].

To summarize, our contributions are three-folds:

- We identify the three challenging problems in the 3D clothing reconstruction: 1) Non-rigid objects; 2) No 3D annotations; 3) Reconstruction ambiguity.
- In an attempt to solve these challenges, we propose a self-supervised learning method with causality design to reconstruct the 3D clothing mesh from large-scale 2D image datasets. Following the spirit of the causal map, we re-design the encoder structure and leverage the "intervention" tool, *i.e.*, two expectation-maximization loops, to facilitate the 3D attribute encoder learning.
- Extensive experiments on two fashion datasets verify the effectiveness of the proposed method both quantitatively and qualitatively. Furthermore, the experiments on the fine-grained bird dataset also show that the proposed method has good scalability to other reconstruction tasks.

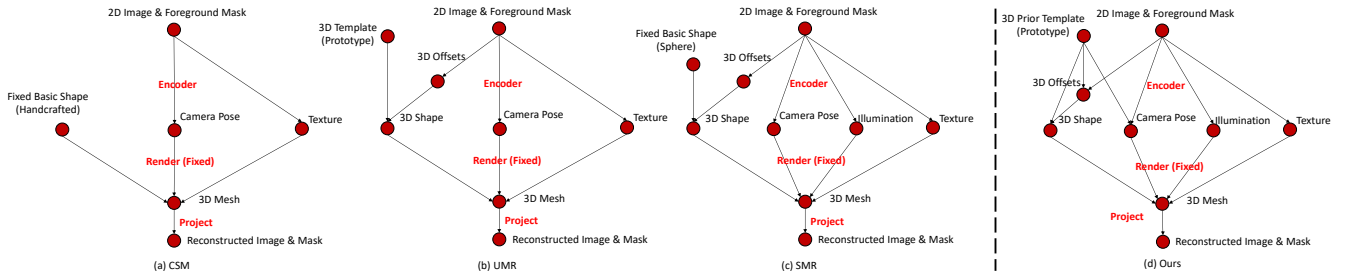

Figure 3: The Structural Causal Map (SCM). We compare the proposed method with three typical 3D reconstruction works, including CSM [21], UMR [23] and SMR [13]. Here we show the image reconstruction loop, *i.e.*, "2D → 3D → 2D". (a) Given one 2D image and the foreground mask, CSM only applies two encoders for Camera and Texture, since it assumes that the basic shape is shared, ignoring intra-class changes. (b,c) UMR and SMR further introduce the shape encoder, which is to predict the shape offset. The final shape is obtained by adding the shape offset to the shape template. (d) In this work, we argue that the predicted shape offsets are also conditioned on the prior template. Besides, the prior template also impacts the camera prediction. Therefore, we explicitly introduce the dependency with the prior prototype. Besides, it is worth noting that our prototype is initialized from a sphere and iteratively updated during training, which does not require any hand-crafted initialization. (Qualitative comparison with more other methods is in Table 1.)

Methods	Viewpoint Annotations	Semantic Keypoint	Manual Template	Part Seg.	Foreground Mask
VPL [19]	✓				✓
CMR [18]† *	(✓)	✓	(✓)		✓
CSM [21]			✓		✓
DIB-R [5]	✓				✓
IMR [45]*			(✓)		✓
ACMR-vid [22]		✓			✓
UMR [23]				✓	✓
WLDO [2]		✓	✓		✓
Texformer [52]			✓	✓	✓
SMR [13]					✓
Ours					✓

Table 1: Comparison with existing methods in terms of supervisions. The proposed method harnesses relatively weak supervision for 3D reconstruction from a single image. *: the method deploys the manual template for initialization. †: the viewpoint annotation is optional. It is also worth noting that some works take 2D input images with blank white background as inputs, and we also view this line of works deploys the foreground mask.

2 RELATED WORK

2.1 3D Reconstruction from Single Image

Human can estimate the 3D structure from one single image. Many works deploy a parameter-based template [17, 27, 44], which is robust but also limits the representative ability to non-rigid objects. To enable more degrees of freedom, Deephuman [59] adopts a U-Net voxel prediction method to reconstruct the human body and clothing, yielding excellent performance, but relatively dense depth and ground-truth 3D annotations are needed. To reduce the dependency on the 3D annotations, PrGANs [8] train a generative adversarial network (GAN) to generate the 3D voxel model (from scratch) from lots of 2D images with different viewpoints. To further avoid the dense prediction in the voxel format, Pixel2Mesh [48] trains a graph neural work with a convolutional neural network encoder to reconstruct 3D mesh by deforming a sphere. We note that Pixel2Mesh is a fully supervised learning work. The ground-truth mesh from ShapeNet [4] is used. Lots of regularization losses, *e.g.*, laplacian regularization and flatten regularization, have been proposed, which are widely used in following works. With the spirit in Pixel2Mesh, lots of researchers have explored weaker supervision, *i.e.*, self-supervised learning. VPL [19] leverages the viewpoint annotation to ensure the reconstruction consistency between different views. However, the annotated viewpoint-image pairs are hard to acquire or be simulated for living objects, such as birds or humans. To invade the viewpoint annotation, one of the early works is Canonical Surface Mapping (CSM) [21], which treats 3D reconstruction as a dense key-point estimation problem and applies the cycle consistency to regularize the model learning. However, CSM deploys a fixed pre-defined shape template, which largely limits intra-class shape changes for different instances. To address the shape limitation, CMR [18] first proposes to use a learnable shape template and Li *et al.* [23] further propose UMR adopt a two-stage training strategy for template updating. The model is first trained to obtain one category-aware template and then the second stage fine-tunes the learned mesh for every instance. The segmentation parsing [15] is also used in UMR for better alignment. The contemporary work, IMR [45], first introduces the mapping function instead of the vertex location regression, saving computation costs due to a large number of vertices. Taking a step further, SMR [13] is proposed to better align the 3D mesh by mix-up and conduct the auxiliary vertex classification task.

The proposed method is mainly different from existing work in three aspects: (1) **Weaker supervision.** As shown in Table 1, the proposed method demands limited supervision and mainly leverages the large-scale multi-view images to learn the prior knowledge. (2) **Model design.** As shown in Figure 4, the network design follows the causal map. In particular, the model contains four independent encoders for four implicit factors, *i.e.*, shape, camera pose, illumination, and texture. Besides, we explicitly take more dependencies, *e.g.*, 3D prior template, into consideration. For instance, both shape encoder and camera encoder harness the integration module (IM) to introduce the 3D prior knowledge. (3) **Optimization strategy.** To deal with the compensation effect in the loss punishment, we deploy

the "intervention" tool, *i.e.*, two expectation-maximization loops, to facilitate the 3D attribute encoder learning.

2.2 Causal Learning

Causal learning is to identify causalities from a set of empirical factors, which can be either pure observations or outcomes, *e.g.*, changes and interventions [37]. To represent the causalities, a causal model is usually defined via structural equations and graphs. For instance, the structural causal model proposed by Pearl *et al.* [35] is a combination of directed acyclic graphs (DAGs). According to the structural causal model, manipulations can be conducted to optimize the estimated relations between variables, *e.g.*, "Do" operation is to cut certain directed edges and control the target variable [35]. Different from data-driven deep learning approaches, the causality methods intend to discover and leverage the causal relation between entities. For example, given data on the chocolate preferences of Nobel Prize winners, deeply-learned models are prone to conclude that the more chocolate a person eats, the more likely he or she is to win a Nobel Prize. Data-driven methods only see the correlation between phenomenons, while causal learning identifies the shared cause, *i.e.*, richer people both eat more chocolates and win more Nobel Prizes. Several existing works also explore implicit causal learning to discover the causal factors during training. CausalGAN [20] trains a generator, which is consistent with an implicit causal graph, and is able to sample from either conditional labels or interventional distributions. Similarly, CausalVAE [54] is a VAE-based causal framework to learn disentangled representations. CausalVAE introduces a Structural Causal Model layer, which discovers latent causal factors in data with graph constraints. Both methods implicitly harness causal mapping by learning latent code or adding one graph constraint loss. However, the causality learned from data is not always accurate and explainable, limiting the causality effect.

Different from CausalGAN and CausalVAE, our work explicitly builds the model according to the structural causal map. In particular, our model follows the spirit of causality between semantic entities to (1) explicitly consider the causality relation between entities, *e.g.*, leveraging the prior template to help both camera encoder and shape encoder learning; and (2) explicitly apply the "intervention" tools to solve the ambiguity of learning multiple variables, *e.g.*, camera, texture, shape, and illumination.

2.3 Expectation Maximization

Expectation Maximization (EM) is an iterative method to find the parameters with maximum likelihood in statistical models [33]. The EM algorithm iteratively conducts two kinds of steps: an expectation step (E-step) to obtain the expectation of latent variables and a maximization step (M-step) which computes parameters to maximize the expected likelihood based on the latent variables. Since M-step updates parameters, it affects the E-step in the next round. In this way, the EM algorithm can keep updating until the convergence. The EM algorithm is usually applied to scenarios that miss the observation of implicit variables, such as Gaussian mixture model [53].

For 3D reconstruction, we also meet a similar problem, when we need to estimate four reconstruction factors simultaneously, *i.e.*, camera position, shape, texture, and illumination from a single 2D image. It is worth noting that if we have three out of four factors, we can easily optimize the rest by efficient gradient backward. This is an M step. Inspired by the expectation-maximization algorithm, we can fix the encoders (E-step) to conduct factor predictions, *e.g.*, camera position, texture, and illumination, and then maximize the likelihood of the rest variable, *e.g.*, shape (M-step). If readers are familiar with the causality, this process actually is a "Do" operation in the causal map [36]. In the E step, we remove the arrows that target camera pose, illumination, and texture, but leave one arrow for 3D Shape Offsets. In the M-step, we back-propagate various losses from the reconstructed image, and only the parameters of the shape encoder are updated (maximize the likelihood). In practice, we iteratively fix three out of four encoders and train the rest one encoder.

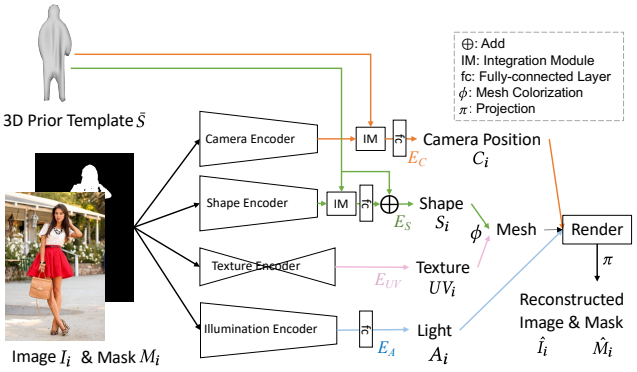

Figure 4: Overview. Here we show a "2D→3D→2D" loop. We follow the causal map in Figure 3 (d) to design the pipeline. Given one pair of clothing image I_i and the mask M_i, we deploy four independent encoders E_c, E_s, E_{UV}, and E_A for camera position, shape, texture and light estimation. We introduce the integration module (IM) to fuse the local feature from 3D prior template $\bar{S}$. Then we apply the colorize function ϕ to obtain the mesh, and utilize the render to re-project the mesh to 2D space via the project function π. Finally, we obtain the reconstructed $\hat{I}_i$ and $\hat{M}_i$. During inference, we could manipulate intermediate camera attributes C_i to generate novel-view images of the target person.

3 METHOD

3.1 Overview

Given a clothing image I_i and the foreground mask M_i, we aim to infer the corresponding 3D mesh with texture (see Figure 4). $i \in [1, N]$ and N is the number of the samples in the dataset. We do not require any extra 3D annotations. In this work, different from CausalGAN [20] and CausalVAE [54], we explicitly follow the structural causal map in Figure 3(d) to build the whole pipeline. Generally, the spirit of causality helps us to (1) disentangle the 3D representation from inherent correlations, such as the ambiguity between the shape and camera encoder, (2) re-consider the causal relation between entities, such as the 3D prior template (prototype) and camera position estimation. Specifically, we deploys four independent encoders, *i.e.*, shape encoder E_S, camera pose encoder E_C, illumination encoder E_A and texture encoder E_{UV}. The design

follows the structural causal map. The decoder is based on the differentiable render [7], which does not contain any learnable parameters. Therefore, we can also regard the render as a fixed decoder. Following existing works [21, 23], we also introduce a 3D prototype $\bar{S} \in \mathbb{R}^{|\bar{S}|\times 3}$, which explicitly involves the prior body structure to the network learning. The 3D prior template can be initialized with an arbitrary mesh. Without loss of generality, we apply the sphere shape (contains 642 vertices and 1280 faces) to initialize the 3D prior template. Here we set 642 vertices as the default setting to illustrate the proposed approach if not specified. During training, we keep updating the prior template $\bar{S}$. When inference, the model can deploy the latest 3D prior template (prototype) for testing.

3.2 Model Structure

Shape Encoder. We follow the causal map to explicitly introduce the 3D prior into the encoder learning. Given the input image-mask pair I_i, M_i and the 3D prior template $\bar{S}$, the shape encoder estimates the offsets $\Delta S_i \in \mathbb{R}^{642\times 3}$ for every vertex:

$$\Delta S_i = E_S(I_i, M_i, \bar{S}), \tag{1}$$

$$S_i = \bar{S} + \Delta S_i. \tag{2}$$

The final 3D shape S_i is the sum of the 3D prior template and the 3D per-vertex offsets, and $S_i \in \mathbb{R}^{642\times 3}$. Different from most existing works [13, 23], which independently estimates the ΔS_i from the input image I_i and the mask M_i, our shape encoder explicitly takes the prior template into the deformation prediction as $E_S(I_i, M_i, \bar{S})$. The main idea is straightforward, since predicting shape offsets depends on foreknowing the shape prior. We explicitly provide the shape template to help the training. In particular, the shape encoder contains one convolutional neural network (CNN) as backbone, one integration module (IM) and a fully connected layer (fc).

Integration Module. As shown in Figure 5, we fuse the visual feature from both the input image/mask and the 3D prior template. We harness the integration module (IM) to extract the local visual feature according to the 2D location by projecting the 3D template to the X-Y plane. On the other hand, the global feature is generated by averaging the input visual feature (by global average pooling) and then we repeat the aggregated feature as the original size. The final ΔS_i is predicted by a fully-connected layer on the concatenated feature of both global features and local features.

Camera Pose Encoder. Similarly, the camera pose estimation also depends on the shape prior and the image. However, the previous works usually ignore the causal dependency on the shape prior $\bar{S}$. When people infer the object position, *e.g.*, distance, azimuth, and elevation, it is necessary to foreknow the general object shape (general size, general shape, symmetry to which axis). Therefore, we also deploy one basic convolutional neural network (CNN) followed by an integration module (IM) and fully-connected layers as the camera position encoder. The camera pose encoder can be formulated as:

$$C_i = E_C(I_i, M_i, \bar{S}), \tag{3}$$

where C_i contains four factors, *i.e.*, distance (1-dim), azimuth (1-dim), elevation (1-dim), and X-Y position offset (2-dim). "dim" is dimension. The distance is also formulated as object scale in other works [23], while the azimuth is called as the rotation degree. We notice that several existing works [13] do not include X-Y position,

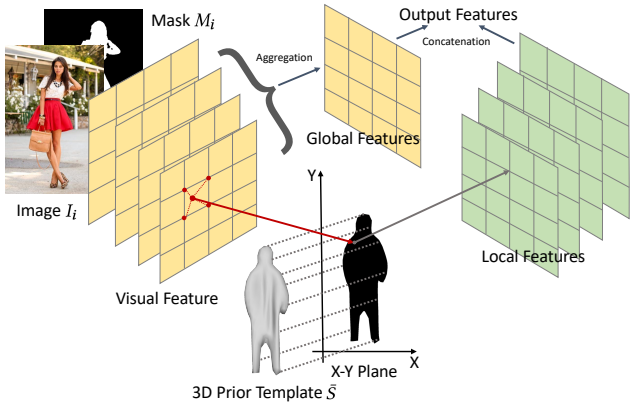

Figure 5: Integration Module (IM). We harness IM to explicitly fuse the prior spatial information from template $\bar{S}$ and the visual feature from (I_i, M_i). The visual feature is the output of the backbone, *e.g.*, HR-Net [55]. As the arrow direction, we leverage the grid sampler to obtain the local feature from the corresponding X-Y of the visual feature. We concatenate global features and local features as final outputs.

since they assume that the object is in the middle of 2D images. However, we find that including X-Y position offset also improves the robustness of the network during both training and testing time. In the experiment, we further visualize the distribution of the learned four camera factors in Figure 9.

Illumination Encoder. The illumination encoder is to regress the illumination direction and strength, which can be simply formulated as a 9-channel Spherical Harmonics coefficient [5]. Therefore, we adopt a basic convolutional neural network followed by a 9-channel fully-connected layer to predict the illumination vector from the input image-mask pair:

$$A_i = E_A(I_i, M_i). \tag{4}$$

Texture Encoder. In this work, we do not predict the color for every vertex. We follow existing works [18, 21] to learn a texture flow as the UV map by a U-Net structure [40]. Given the input image I_i and the corresponding foreground mask M_i, we first predict the texture flow and then map the color according to the spatial location.

$$UV_i = E_{UV}(I_i, M_i). \tag{5}$$

Decoder (Render). Finally, the decoder, *i.e.*, render, can reconstruct the 3D mesh with color by simply combining the shape S_i and UV_i. If we want to re-project the mesh to the 2D space, we further need the camera pose C_i and the illumination direction A_i. Therefore, the reconstructed image $\hat{I}_i$ can be written as:

$$\hat{I}_i = \pi(\phi(S_i, UV_i), C_i, A_i), \tag{6}$$

where ϕ is the function to colorize the 3D mesh S_i with the UV map UV_i. π denotes the projection function mapping the mesh through camera parameters C_i with the illumination A_i. As a side product, we can also obtain the reconstructed foreground mask $\hat{M}_i$ during projection. We note that both ϕ and π are based on the physical mapping, so there do not contain any learnable parameters.

3.3 Optimization Objectives

Image Reconstruction Loss. As shown in the right part of Figure 6, we calculate the pixel level l_1 loss of the foreground area between the reconstructed image and the input 2D image:

$$\mathcal{L}_{img} = \mathbb{E}[||I_i \odot M_i - \hat{I}_i \odot \hat{M}_i||_1], \tag{7}$$

where $\odot$ denotes element-wise multiplication, and $\mathbb{E}$ denotes the expectation. $\hat{I}_i$ and $\hat{M}_i$ are the reconstructed image and mask projected from the 3D mesh. We note that the $\mathcal{L}_{image}$ focuses on the low-level input. Sometimes the generation quality is good but with small position shifts. To further ensure the generation quality from high-level activations, we also introduce the adversarial loss:

$$\mathcal{L}_{adv} = \mathbb{E}[logD(I_i \oplus M_i) + log(1 - \hat{I}_i \oplus \hat{M}_i)], \tag{8}$$

where $\oplus$ means concatenation. For instance, $I_i \oplus M_i$ is a 4 channel input. D denotes a multi-layer discriminator to classify whether the input is real or generated from our render. In practice, we adopt a basic WGAN structure [1] as the discriminator.

Otherwise, we introduce the IoU loss to compare the overlapping area between the generated mask with the ground-truth input mask.

$$\mathcal{L}_{IoU} = \mathbb{E}[1 - \frac{M_i \cap \hat{M}_i}{M_i \cup \hat{M}_i}]. \tag{9}$$

Attribute Reconstruction Loss. As shown in the left part of Figure 6, we also conduct the 3D attribute reconstruction to ensure that the encoder and the decoder is self-consistent:

$$\begin{aligned}\mathcal{L}_{attri} = &\mathbb{E}[||S_i - E_S(\hat{I}_i, \hat{M}_i, \bar{S})||_1] + \mathbb{E}[||C_i - E_C(\hat{I}_i, \hat{M}_i, \bar{S})||_1] \\ &+\mathbb{E}[||A_i - E_A(\hat{I}_i, \hat{M}_i)||_1] + \mathbb{E}[||UV_i - E_{UV}(\hat{I}_i, \hat{M}_i)||_1].\end{aligned} \tag{10}$$

The 3D attributes predicted from the reconstructed image $\hat{I}_i$ should be the same as the predicted attribute from I_i.

Mesh Regularization. (1) Laplacian loss [48] is a regularization to prevent self-intersection of mesh faces. It encourages adjacent vertices to move in the same direction, consequently, avoiding the local part of the mesh producing outrageous deformation. For each vertex position p in the mesh shape S_i, the laplacian coordinate is $\delta_p = p - \sum_{k \in K(p)} \frac{k}{||K(p)||}$, where $K(p)$ are the neighbor vertices of p with connected edges. Specifically, the laplacian loss can be defined as $\mathcal{L}_{lpl} = \mathbb{E}[||\hat{\delta_p} - \delta_p||_2^2]$, where $\hat{\delta_p}$ and δ_p are laplacian coordinates of a vertex before and after the updation respectively; (2) Flatten loss is another regularization for keeping faces from intersecting [48]. The cosine of the angle between two adjacent faces is calculated. The flatten loss is defined as $\mathcal{L}_{flat} = \mathbb{E}[(cos(\Delta\theta_i) + 1)^2]$, where $\Delta\theta_i$ is the angle between two adjacent faces. The angle around 180° implies a smooth mesh surface; (3) Symmetry loss constrains mesh deformations to be reflectional symmetric in the depth [45]. It can be expressed as $\mathcal{L}_{sym} = \mathbb{E}[||Z(p) + Z(\tilde{p})||_1]$, where Z denotes the depth of the vertex and $\tilde{p}$ is the reflected vertex of p; (4) Deformation loss [18, 23] is a regularization to discourage the mesh to deform excessively and to facilitate learning a meaningful average shape. It is defined as $\mathcal{L}_{deform} = \mathbb{E}[||\Delta S||_2]$.

Total Loss. We train four encoders and discriminator to optimize the total objective, which is a weighted sum of above-mentioned losses:

$$\begin{aligned}\mathcal{L}_{total} = &\lambda_{rec}(\mathcal{L}_{img} + \mathcal{L}_{IoU}) + \lambda_{attri}\mathcal{L}_{attri} + \lambda_{adv}\mathcal{L}_{adv} \\ &+\lambda_{reg}(\mathcal{L}_{sym} + \mathcal{L}_{deform} + \lambda_{lpl}\mathcal{L}_{lpl} + \lambda_{flat}\mathcal{L}_{flat}).\end{aligned} \tag{11}$$

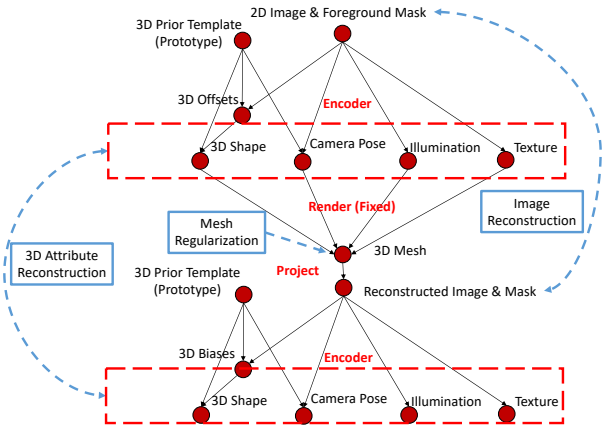

Figure 6: Losses. Here we show three kinds of losses on the causal map, which are the image reconstruction loss, the attribute reconstruction loss and the mesh regularization.

In practice, we refer to existing works [13, 23, 48], and empirically set $\lambda_{rec} = 2, \lambda_{attri} = 1, \lambda_{adv} = 1 \times 10^{-5}, \lambda_{reg} = 0.1, \lambda_{lpl} = 0.1, \lambda_{flat} = 0.01$.

3.4 Optimization Strategy

Encoder Loop. During the implementation, we notice one main challenge is simultaneously optimizing the four encoders. The problem mainly lies in the image reconstruction loss. For instance, even if three out of four encoders provide correct prediction, the rest provides the wrong attribute, such as incorrect shape offsets. All four encoders are penalized equally. This is one typical "collider" case in causality. Therefore, one straightforward idea is to train one encoder (predict one cause) while fixing the other three encoders (control other causes). In this way, we can effectively penalize the target encoder. In particular, we adopt one expectation-maximization loop, which also is an "invention" tool in causal learning. For example, we fix the three encoders, *e.g.*, E_c, E_A, E_{UV}, which cuts the arrows from inputs to three attributes, *i.e.*, C_i, A_i and UV_i. Only the shape attribute S_i still keeps the dependency from the input image-mask pair. Hence, when the loss is back-propagating, only the shape encoder is penalized. In this way, we disentangle four encoders not only in the forward passing design (independent encoder weights) but also in the loss of back-propagation.

Prototype Loop. We also observe a common ambiguity between prototype updating and shape estimation. The problem is mostly due to Eq. 2. Since it is an addition equation, during gradient back-propagation, $\bar{S}$ and ΔS_i receive the punishment equally. It is hard to distinguish $\bar{S}$ from ΔS_i. Therefore, we adopt the causal invention tool, *i.e.*, Expectation-Maximization. During training, we fix the prototype (control one cause) and maximize the shape offsets likelihood (predict another cause). After every training epoch, we leverage the mean shape offsets to update the $\bar{S} = \bar{S} + \mathbb{E}[\Delta S]$. In practice, different from existing works (*e.g.*, two-stage training [23]), we adopt a linear warming-up strategy [11] to update prototype slowly in the early epochs and harness the exception handling by clipping extreme deformations. In this way, we disentangle the prototype updating from the shape offsets estimation and learn the model in one go.

Methods	HMR [17]	ROMP [44]	Ours
MaskIoU (%) ↑	69.7	70.3	82.0

Table 2: Comparison with two off-the-shelf template-based methods on the human clothing ATR dataset. Since no texture mapping is contained in template-based methods, we only compare MaskIoU (%), which reflects the "2D → 3D → 2D" reconstruction quality on the unseen test set.

4 EXPERIMENT

We evaluate the proposed approach on two fashion datasets, ATR [24] and Market-HQ [56], and one widely-used bird dataset CUB [46].

Evaluation Metric. Since there are no ground-truth 3D meshes, we follow existing works [13] and adopt 2D metric, *e.g.*, FID [12], SSIM [49] and MaskIoU, to evaluate projected images. FID compares the distribution of two sets of images. We denote the 3D reconstruction results as FID_{recon}, the generated image with different viewpoints as FID_{novel} and the side-view image as FID_{90}. We follow [13] to rotate the camera viewpoint to save images for FID_{novel}. SSIM provides a low-level similarity comparison between reconstructed images and input images. It reflects the effectiveness of the process "2D → 3D → 2D". Besides, we also apply the MaskIoU to compare overlapping regions between the ground-truth mask and the foreground area projected from reconstructed 3D meshes.

Reproducibility. Our code is based on Pytorch [34]. More training details and model structures are provided in Supplementary Material.

4.1 Quantitative Experiments

Comparison with Template-based Methods. We first compare with the off-the-shelf template-based methods [17, 44] in Table 2. This line of methods is based on the body template with great robustness, but is not well scalable to the non-rigid clothing. Since no texture mapping function is built in template-based methods, we focus on comparing MaskIoU (%), which reflects the "2D → 3D → 2D" shape reconstruction quality. We observe that the proposed method achieves a higher MaskIoU score of 82.0% on the test set. The result is also consistent with the visualization in Figure 1. For clothing reconstruction, the proposed method is more scalable than template-based methods, covering more regions of interest.

Comparison with Single-image Reconstruction Methods. As shown in Table 3, we compare the proposed method with other state-of-the-art approaches [5, 13, 18, 22, 23] on the CUB dataset. Among existing works, SMR [13] has achieved the high-fidelity reconstruction and novel-view generation performance. In contrast, the proposed method yields a competitive reconstruction performance (84.0% MaskIoU and 82.5% SSIM) with SMR. At meantime, for novel-view generation, ours surpasses SMR with 76.2 FID_{novel}.

4.2 Qualitative Experiments

Reconstruction and Novel-view Results . As shown in Figure 7, we reconstruct the person with non-rigid clothing. We could observe that the model not only successfully learns the legs and arms, but also captures non-rigid objects, including hair, dress and handbag. Since we disentangle four encoders, the proposed method could easily manipulate 3D attributes, including distance, elevation, shape and texture, for customization.

Methods	MaskIoU (%) ↑	SSIM (%) ↑	FID_{novel} ↓
CMR [18]	73.8	44.6	115.1
DIB-R [5]	75.7	-	-
ACMR-vid [22]	77.3	-	-
UMR [23]	73.4	71.3	83.6
SMR [13]	80.6	**83.2**	79.2
Ours	**83.0**	82.5	**76.2**

Table 3: Comparison with other single-image reconstruction methods on the CUB bird dataset. MaskIoU (%) and SSIM (%) reflects the reconstruction quality on the unseen test set, while FID_{novel} compares the distribution difference between generated images from novel views and the original dataset.

Exchanging Clothing. Inspired by 2D GAN-based work [58], we also show the result of changing the texture of any two persons but with a 3D mesh manner (see Figure 8). In particular, we apply the shape encoder and the texture encoder to extract the shape S_i and the UV texture map UV_j. Then we deploy the render to generate the new mesh based on S_i and UV_j. The first row and the first column are the input RGB images. The rest is the projected results of the new 3D meshes. We rotate the mesh for better 3D visualization. It verifies the robustness of the method. The learned UV map could successfully be aligned on different human meshes, even though we have not introduced any part annotations during the training process.

4.3 Ablation Study and Further Discussion

Camera Attribute Distribution. We observe that the camera encoder successfully captures the camera distribution in the Market-HQ test set. As shown in Figure 9, most samples are $2 \sim 4$ unit distances from the camera, and most persons are in the center of the figure with 0 X-Offsets and 0 Y-Offsets. Most samples appear in 0°, 180° or -180°, which means that most people are facing towards or backward to the camera. It is aligned with the dataset setup since cameras are set up in front of the supermarket entrance or exit. The elevation of most samples is from -10 to 10, which is also aligned with the data collection setup, *i.e.*, six horizontal-view cameras. Besides, we also show the distribution of mean shape offset ΔS. We observe that most deformations are relatively small since we have introduced the 3D prior template. In a summary, the learned attribute statistics also verify that we disentangle the camera hyper-parameter, *e.g.*, scale changes and position offsets, from the shape encoder.

Does the two expectation-maximization loops help the encoder learning? Yes. As shown in Table 4, we conduct two ablation studies on Market-HQ. (1) One is to stop the prototype updating, *i.e.*, No Prototype Loop, and we deploy the fixed sphere as the basic shape. It directly limits the shape deformation, compromising the reconstruction performance. (2) Besides, we also explore training all encoders simultaneously without encoder loop as "No Encoder Loop". We observe that the model can easily over-fit the front-view reconstruction quality but it does not perform well in the novel view, especially we look at the 3D mesh from the side view, *i.e.*, 90°.

Does the integration module work? Yes. As one minor contribution, we design the integration module to explicitly fuse the prior prototype as local features for learning shape. As shown in Table 4,

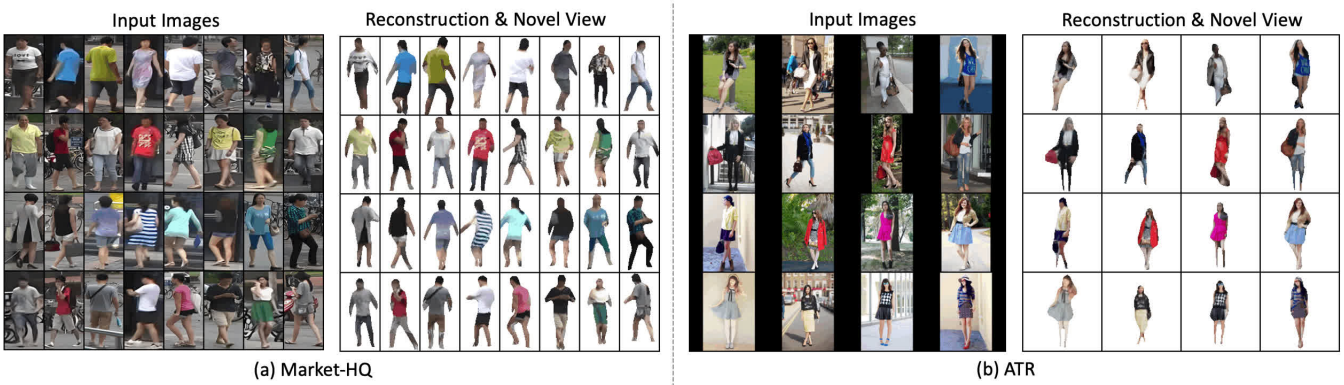

Figure 7: Novel-view 3D clothing generation from single images on the unseen test set of Market-HQ and ATR. (Please open the paper in Adobe Reader to see the mesh rotation.) Here we gradually "Do" / change the camera azimuth degree to render the human.

Figure 8: 3D clothing changing by exchanging the 3D mesh shape and texture. (Please open Adobe Reader to see the movement.)

removing integration module, *i.e.*, No IM, leads to a performance drop in both reconstruction and novel-view generation.

Limitation. There are two faces or two backs of heads on one reconstructed mesh. It is because our work is still based on a single image, and the learned model only "sees" one single view of the human. Especially on the ATR dataset (most photos are frontal faces), it can not learn the 3D prior, *i.e.*, one person only has one face and one back of the head. Therefore, even if we introduce the discriminator from WGAN [1] on the 2D shape and texture, it can not provide 3D-aware adversarial loss. The model still largely relies on the symmetric structure to generate the back view. Hence, we think that, in the future, the large-scale multi-view image datasets may help to further solve this limitation upon our work.

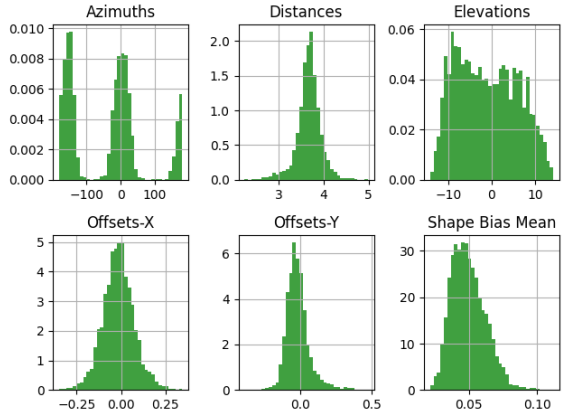

Figure 9: Histogram of 3D Camera Attributes C on Market-HQ. Here we show the distribution of azimuths, distances, elevations, Offsets-X and Offsets-Y. Besides, we also provide the distribution of the mean shape offset ΔS over the test set.

Methods	MaskIoU (%) ↑	SSIM (%) ↑	FID_{recon} ↓	FID_{novel} ↓	FID_{90} ↓
No IM	71.5	55.8	49.1	83.1	123.8
No Prototype Loop	77.7	59.6	38.1	66.1	116.0
No Encoder Loop	**85.3**	**70.6**	**19.5**	65.4	190.1
Ours	81.1	61.8	29.9	**58.2**	**109.0**

Table 4: Ablation Study on Market-HQ. "No IM" denotes that we remove the integration module. We observe that although "No Encoder Loop" leads the model to over-fitting the front-view reconstruction, the side-view performance FID_{90} is extremely poor. In contrast, the full model takes a balance point between reconstruction and novel-view generation.

5 CONCLUSION

In this work, we study the 3D clothing reconstruction problem to build a "3D Magic Mirror". We notice that there are three primary challenges, *i.e.*, non-rigid clothing, no 3D ground-truth annotations, and the inherent reconstruction ambiguity. In an attempt to solve the three problems, we study the self-supervised learning method with causality design to reconstruct 3D clothing mesh from large-scale 2D image datasets. Specifically, we follow the spirit of the structural causal map to re-design the output dependency, and leverage two expectation-maximization loops to facilitate the training

process. In the experiment, we observe that, despite using relatively weak supervision, the proposed method is still competitive with other existing works, and shows great scalability to different datasets. In the future, we will continue to explore the application to 3D object re-id [60] and building reconstruction [57].

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[2] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. 2020. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*.

[3] Michael P Chae, Warren M Rozen, Paul G McMenamin, Michael W Findlay, Robert T Spychal, and David J Hunter-Smith. 2015. Emerging applications of bedside 3D printing in plastic surgery. *Frontiers in surgery* 2 (2015), 25.

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[5] Wenzheng Chen, Jun Gao, Huan Ling, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. 2019. Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer. In *Advances In Neural Information Processing Systems*.

[6] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3001–3011.

[7] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. 2022. Kaolin: A Pytorch Library for Accelerating 3D Deep Learning Research. https://github.com/NVIDIAGameWorks/kaolin.

[8] Matheus Gadelha, Subhransu Maji, and Rui Wang. 2017. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*. IEEE, 402–411.

[9] Lian Gao, Di Huang, Yuanfang Guo, and Yunhong Wang. 2019. Pedestrian attribute recognition via hierarchical multi-task learning and relationship attention. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1340–1348.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in neural information processing systems* 31 (2018).

[11] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[13] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. 2021. Self-Supervised 3D Mesh Reconstruction from Single Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6002–6011.

[14] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. 2019. Illumination-invariant person re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*. 365–373.

[15] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. 2019. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 869–878.

[16] Wei Ji, Xi Li, Yueting Zhuang, Omar El Farouk Bourahla, Yixin Ji, Shihao Li, and Jiabao Cui. 2018. Semantic Locality-Aware Deformable Network for Clothing Segmentation.. In *IJCAI*. 764–770.

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *Computer Vision and Pattern Regognition (CVPR)*.

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 371–386.

[19] Hiroharu Kato and Tatsuya Harada. 2019. Learning view priors for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9778–9787.

[20] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. 2017. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023* (2017).

[21] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. 2019. Canonical surface mapping via geometric cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2202–2211.

[22] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. 2020. Online adaptation for consistent mesh reconstruction in the wild. *Advances in Neural Information Processing Systems* 33 (2020), 15009–15019.

[23] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. 2020. Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision*. Springer, 677–693.

[24] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. 2015. Deep Human Parsing with Active Template Regression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37, 12 (Dec 2015), 2402–2414. https://doi.org/10.1109/TPAMI.2015.2408360

[25] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. 2015. ICCV. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[26] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1954–1963.

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. Mesh Graphormer. *ICCV* (2021).

[28] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition* 95 (2019), 151–161.

[29] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3330–3337.

[30] Xinchen Liu, Meng Zhang, Wu Liu, Jingkuan Song, and Tao Mei. 2019. Braidnet: Braiding semantics and details for accurate human parsing. In *Proceedings of the 27th ACM International Conference on Multimedia*. 338–346.

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. *Advances in neural information processing systems* 30 (2017).

[33] Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine* 13, 6 (1996), 47–60.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[35] Judea Pearl. 2009. *Causality*. Cambridge university press.

[36] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

[37] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

[38] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*. 650–667.

[39] Haolin Ren, Zheng Wang, Zhixiang Wang, Lixiong Chen, Shin'ichi Satoh, and Daning Hu. 2020. An Interactive Design for Visualizable Person Re-Identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4536–4538.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[41] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4814–4821.

[42] Zijun Sha, Zelong Zeng, Zheng Wang, Yoichi Natori, Yasuhiro Taniguchi, and Shin'ichi Satoh. 2020. Progressive domain adaptation for robot vision person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4488–4490.

[43] Keng Hua Sing and Wei Xie. 2016. Garden: A mixed reality experience combining virtual reality and 3D reconstruction. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 180–183.

[44] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. 2021. Monocular, One-stage, Regression of Multiple 3D People. In *ICCV*.

[45] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. 2020. Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504* (2020).

[46] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.

[47] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2275–2284.

[48] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In *Proceedings of the European conference on computer vision (ECCV)*. 52–67.

[49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[50] Zheng Wang, Wu Liu, Yusuke Matsui, and Shin'ichi Satoh. 2020. Effective and efficient: Toward open-world instance re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4789–4790.

[51] Longhui Wei, Xiaobin Liu, Jianing Li, and Shiliang Zhang. 2018. VP-ReID: Vehicle and person re-identification system. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 501–504.

[52] Xiangyu Xu and Chen Change Loy. 2021. 3D human texture estimation from a single image with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13849–13858.

[53] Guorong Xuan, Wei Zhang, and Peiqi Chai. 2001. EM algorithms of Gaussian mixture model and hidden Markov model. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, Vol. 1. IEEE, 145–148.

[54] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9593–9602.

[55] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. 2021. Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10440–10450.

[56] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[57] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*. 1395–1403.

[58] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2138–2147.

[59] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7739–7749.

[60] Zhedong Zheng, Nenggan Zheng, and Yi Yang. 2020. Parameter-efficient person re-identification in the 3d space. *arXiv preprint arXiv:2006.04569* (2020).