Abstract

The function of constructing the hierarchy of objects is important to the visual process of the human brain. Previous studies have successfully adopted capsule networks to decompose the digits and faces into parts in an unsupervised manner to investigate the similar perception mechanism of neural networks. However, their descriptions are restricted to the 2D space, limiting their capacities to imitate the intrinsic 3D perception ability of humans. In this paper, we propose an Inverse Graphics Capsule Network (IGC-Net) to learn the hierarchical 3D face representations from large-scale unlabeled images. The core of IGC-Net is a new type of capsule, named graphics capsule, which represents 3D primitives with interpretable parameters in computer graphics (CG), including depth, albedo, and 3D pose. Specifically, IGC-Net first decomposes the objects into a set of semantic-consistent part-level descriptions and then assembles them into object-level descriptions to build the hierarchy. The learned graphics capsules reveal how the neural networks, oriented at visual perception, understand faces as a hierarchy of 3D models. Besides, the discovered parts can be deployed to the unsupervised face segmentation task to evaluate the semantic consistency of our method. Moreover, the part-level descriptions with explicit physical meanings provide insight into the face analysis that originally runs in a black box, such as the importance of shape and texture for face recognition. Experiments on CelebA, BP4D, and Multi-PIE demonstrate the characteristics of our IGC-Net.

1. Introduction

A path toward autonomous machine intelligence is to enable machines to have human-like perception and learning abilities [19]. As humans, by only observing the objects, we can easily decompose them into a set of part-level components and construct their hierarchy even though we have never seen these objects before. This phenomenon is supported by the psychological studies that the visual process of the human brain is related to the construction of the hierarchical structural descriptions [11, 22, 23, 29]. To investigate the similar perception mechanism of neural networks, previous studies [18, 35] incorporate the capsule networks, which are designed to present the hierarchy of objects and describe each entity with interpretable parameters. After observing a large-scale of unlabeled images, these methods successfully decompose the digits or faces into a set of parts, which provide insight into how the neural networks understand the objects. However, their representations are limited in the 2D space. Specifically, these methods follow the analysis-by-synthesis strategy in model training and try to reconstruct the image by the decomposed parts. Since the parts are represented by 2D templates, the reconstruction becomes estimating the affine transformations to warp the templates and put them in the right places, which is just like painting with stickers. This strategy performs well when the objects are intrinsically 2D, like handwritten digits and frontal faces, but has difficulty in interpreting 3D objects in the real world, especially when we want a view-independent representation like humans [2].

How to represent the perceived objects is the core research topic in computer vision [3, 25]. One of the most popular theories is the Marr’s theory [22, 23]. He believed that the purpose of the vision is to build the descriptions
of shapes and positions of things from the images and construct hierarchical 3D representations of objects for recognition. In this paper, we try to materialize Marr’s theory on human faces and propose an Inverse Graphics Capsule Network (IGC-Net), whose primitive is a new type of capsule (i.e., graphics capsule) that is defined by computer graphics (CG), to learn the hierarchical 3D representations from large-scale unlabeled images. Figure 1 shows an overview of the proposed method. Specifically, the hierarchy of the objects is described with the part capsules and the object capsules, where each capsule contains a set of interpretable parameters with explicit physical meanings, including depth, albedo, and pose. During training, the input image is first encoded to a global shape and albedo embeddings, which are sent to a decomposition module to get the spatially-decoupled part-level graphics capsules. Then, these capsules are decoded by a shared capsule decoder to get explicit 3D descriptions of parts. Afterward, the parts are assembled by their depth to generate the object capsules as the object-centered representations, naturally constructing the part-object hierarchy. Finally, the 3D objects embedded in the object capsules are illuminated, posed, and rendered to fit the input image, following the analysis-by-synthesis manner. When an IGC-Net is well trained, the learned graphics capsules naturally build hierarchical 3D representations.

We apply IGC-Net to human faces, which have been widely used to investigate human vision system [31] due to the similar topology structures and complicated appearances. Thanks to the capacity of the 3D descriptions, IGC-Net successfully builds the hierarchy of in-the-wild faces that are captured under various illuminations and poses. We evaluate the IGC-Net performance on the unsupervised face segmentation task, where the silhouettes of the discovered parts are regarded as segment maps. We also incorporate the IGC-Net into interpretable face analysis to uncover the mechanism of neural networks when recognizing faces.

The main contributions of this paper are summarized as:

- This paper proposes an Inverse Graphics Capsule Network (IGC-Net) to learn the hierarchical 3D face representations from unlabeled images. The learned graphics capsules in the network provide insight into how the neural networks, oriented at visual perception, understand faces as a hierarchy of 3D models.

- A Graphics Decomposition Module (GDM) is proposed for part-level decomposition, which incorporates shape and albedo information as cues to ensure that each part capsule represents a semantically consistent part of objects.

- We execute the interpretable face analysis based on the part-level 3D descriptions of graphics capsules. Besides, the silhouettes of 3D parts are deployed to the unsupervised face segmentation task. Experiments on CelebA, BP4D, and Multi-PIE show the effectiveness of our method.

2. Related Work

2.1. Capsule Network

The connections of the human brain are thought to be sparse and hierarchical [1, 4, 9, 15], which inspires the design of capsule networks to present the objects with dynamic parse trees. Given inputs, capsule networks [12, 13, 18, 26, 27, 35] will encode the images to a set of low-level capsules, which describe the local entities of the objects, and then assemble them into higher-level capsules to describe more complicated entities. The parameters of capsules are usually with explicit meanings, which enables the interpretability of neural networks. Recently, some capsule networks have been proposed to explore the hierarchy of objects. SCAE [18] proposes to describe the objects with a set of visualizable templates through unsupervised learning. However, SCAE can only handle simple 2D objects like digits. HP-Capsule [35] extends SCAE to tackle human faces, which proposes subpart-level capsules and uses the compositions of subparts to present the variance of pose and appearance. Due to the limitation of 2D representations, HP-Capsule can only tackle faces with small poses. Sabour et al. [27] propose to apply the capsule network to human bodies, but it needs optical flow as additional information to separate the parts. In this paper, we propose graphics capsules to learn the hierarchical 3D representations from unlabeled images.

2.2. Unsupervised Part Segmentation

We evaluate the graphics capsule performance on the unsupervised face segmentation task. Several methods have been proposed for this challenging task. DFF [7] proposes to use non-negative matrix factorization upon the CNN features to discover semantics, but it needs to optimize the whole dataset during inference. Choudhury et al. [6] follow a similar idea, which uses k-means to cluster the features obtained by a pre-trained network. SCOPS [14] and Liu et al. [20] propose to constrain the invariance of images between TPS transformation. However, their methods rely on the concentration loss to separate parts, leading to similar silhouettes of different parts. HP-Capsule [35] proposes a bottom-up schedule to aggregate parts from subparts. The parts of the HP-Capsule rely on the learning of subpart-part relations, which is unstable when tackling faces with large poses. Compared with these methods, our IGC-Net can provide interpretable 3D representations of the parts, which are with salient semantics and keep semantic consistency across the in-the-wild faces with various poses.
2.3. Unsupervised 3D Face Reconstruction

Learning to recover the 3D face from 2D monocular images has been studied for years. Following the analysis-by-synthesis strategy, many methods [5, 8, 32, 39] propose to estimate the parameters of the 3D Morphable Model [24], which describes the faces with a uniform topology predefined by humans. Recently, several works [34, 37, 38] have been proposed to only use the symmetric character of faces to learn 3D face reconstruction. Under the graphics decomposition, these methods achieve promising results. Inspired by them, we propose the graphics capsule to learn the hierarchical 3D face representations from images, which provides insight into how neural networks understand faces by learning to decompose them into a set of parts.

3. Method

Based on previous explorations in capsule networks [18, 35], our goal is to explore a system that can build hierarchical 3D representations of objects through browsing images. Specifically, we focus on the human faces and aim to learn the part-object hierarchy in an unsupervised manner, where each part is represented by a set of interpretable CG parameters, including shape, albedo, 3D poses, etc. In the following sections, we will introduce the graphics capsule and the overview of the network in Section 3.1, the graphics decomposition module that is used to build hierarchy in Section 3.2, and the loss functions that enable unsupervised learning in Section 3.3.

3.1. Overview

To learn a hierarchical 3D representation from unlabeled images, we propose an Inverse Graphics Capsule Network (IGC-Net), whose capsules are composed of interpretable CG descriptions, including a depth map $D \in \mathbb{R}^{H \times W}$, an albedo map $A \in \mathbb{R}^{C \times H \times W}$ and 3D pose parameters $p \in \mathbb{R}^{1 \times 6}$ (rotation angles and translations). Our IGC-Net is applied to human faces, which have been widely used to investigate the human vision system due to their similar topology structures and complicated appearances. The overview of IGC-Net is shown in Figure 1. Following a bottom-up schedule, a CNN-based image encoder first encodes the input image $I$ into the shape and the albedo embeddings $f_s$ and $f_a$:

$$f_s, f_a = \text{ImageEncoder}(I). \quad (1)$$

Then a Graphics Decomposition Module (GDM) is employed to decompose the global embeddings into a set of part-level embeddings, which can be further decoded into interpretable graphics capsules:

$$\{\hat{e}_s^1, ..., \hat{e}_s^M\}, \{\hat{e}_a^1, ..., \hat{e}_a^M\} = \text{GDM}(f_s, f_a),$$

$$\{D_p^m, A_p^m, p_p^m\} = \text{GraphicsDecoder}(\hat{e}_s^m, \hat{e}_a^m), \quad (2)$$

where $\hat{e}_s^m$ is the shape embedding of the $m$th part, $\hat{e}_a^m$ is the corresponding albedo embedding, $M$ is the number of part capsules, and $\Theta_p^m : \{D_p^m, A_p^m, p_p^m\}$ is a graphics capsule that describes a part with depth, albedo, and 3D pose. Afterward, the part capsules are assembled according to their

![Figure 1. Overview of the Inverse Graphics Capsule Network (IGC-Net). The input image is first encoded to global shape and albedo embeddings and then sent to a decomposition module to get the spatially-decoupled part-level graphics capsules. Afterward, these capsules are decoded to get the explicit 3D descriptions of parts, which are assembled by their depth to generate the object capsules as object-centered representations. Finally, the object capsules are illuminated, posed, and rendered to fit the input image. After training, the learned graphics capsules naturally build hierarchical 3D representations.](image-url)
Humans can decompose an object into a set of parts and construct a hierarchy by just observation. To realize this ability in neural networks, we propose the Graphics Decomposition Module (GDM) to decompose the global embedding of the image into a set of semantic-consistent part-level descriptions. The illustration of GDM is shown in Figure 2.

Taking shape decomposition as an example, GDM maintains $M$ shape basis $\{W^m_s\}$ as the implicit part templates. Given the global embeddings $f_o$ extracted in Eqn. 1, GDM performs cross attention between the global embedding and the basis to get $M$ disentangled $D$ dimensional embeddings:

$$e_s^m = f_o W^m_s, \quad m = 1, ..., M.$$  \tag{5}

To further reduce the entanglement between $\{e_s^m\}$ and generate independent part-level embeddings, an $M$-way one-hot attention vector is generated for each of the $D$ dimensions, by deploying that only one embedding can preserve its value and the others are set to 0 at each dimension. This dimension attention is formulated as:

$$\hat{e}_s^m = e_s^m \odot M_{[m,:]}$$

$$M_{[:,d]} = \text{hard softmax}(\{e_s^1(d), e_s^2(d), ..., e_s^M(d)\}),$$

$$\text{hard softmax}(\mathbf{e}) = \frac{\mathbf{e}}{\sum_i \mathbf{e}(i)} \odot \text{onehot}\left(\frac{\mathbf{e}}{\sum_i \mathbf{e}(i)}\right),$$  \tag{6}

where $M_{M \times D}$ is the attention matrix, whose $m$th row is $M_{[m,:]}$ and $d$th column is $M_{[ :,d]}$. $e_s^m(d)$ is the $d$th dimension of the embedding $e_s^m$, onehot($\cdot$) is the one-hot operation, and $\hat{e}_s^m$ is the final part-level shape embedding. The same pipeline is applied to the albedo embeddings, where the only difference is that the attention $M$ is copied from the shape embeddings, which ensures that the shape and the albedo information are decomposed synchronously.

By incorporating both shape and albedo information as cues, GDM successfully decomposes parts from objects under varied poses and appearances, ensuring that each part capsule represents a semantic-consistent part.

### 3.3. Loss and Regularization

When training IGC-Net with unlabelled images, we employ the following constraints to learn the hierarchical 3D representations effectively:

**Reconstruction.** We adopt the negative log-likelihood loss [34] to measure the distance between the original image $\mathbf{I}$ and the reconstructed image $\hat{\mathbf{I}}$:

$$\mathcal{L}_{\text{rec}} = -\frac{1}{|\Omega|} \sum \log \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{\sqrt{2}||\hat{\mathbf{I}} - \mathbf{I}||}{\sigma}$$

$$-\frac{1}{|\Omega|} \sum \log \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{\sqrt{2}||\hat{\mathbf{I}}_{flip} - \mathbf{I}||}{\sigma},$$  \tag{7}

where $\Omega$ is for normalization and $\sigma \in \mathbb{R}^{H \times W}$ is the confidence map estimated by a network to present the symmetric
probability of each position in $\mathbf{I}$, $\hat{\mathbf{I}}_{flip}$ is the image reconstructed with the flipped albedo and shape. Following unsup3d [34], we also incorporate the perceptual loss to improve the reconstruction results:

$$L_{rec} = -\frac{1}{|\Omega^{(k)}|} \sum_{i} \ln \frac{1}{\sqrt{2\sigma^{(k)}}} \exp - \frac{\sqrt{2} |f^{(k)}(\hat{\mathbf{I}}) - f^{(k)}(\mathbf{I})|}{\sigma^{(k)}}$$

$$- \frac{1}{|\Omega^{(k)}|} \sum_{i} \ln \frac{1}{\sqrt{2\sigma^{(k)}}} \exp - \frac{\sqrt{2} |f^{(k)}(\hat{\mathbf{I}}_{flip}) - f^{(k)}(\mathbf{I})|}{\sigma^{(k)}},$$

(8)

where $f^{(k)}(\cdot)$ is the $k$-th layer of a pre-trained image encoder (VGG [28] in this paper) and $\sigma^{(k)}$ is the corresponding confidence map.

**Semantic Consistency.** In GDM, shape embedding is used as the cue for part decomposition. To improve the semantic consistency across samples, we employ a contrastive loss on the shape embedding $\hat{e}^{m}_s$ of each capsule, which is formulated as:

$$L_{contra} = - \frac{B}{B} \sum_{b=1}^{M} \log \frac{\sum_{i \neq b} \exp(e^{m, (b)}_s, e^{m, (i)}_s / \tau)}{\sum_{i \neq b} \exp(e^{m, (b)}_s, e^{m, (i)}_s / \tau) + \sum_{j \neq m} \sum_{i \neq b} \exp(e^{m, (b)}_s, e^{m, (i)}_s / \tau),}$$

(9)

where $B$ is the batch size, $M$ is the number of part capsules, $\hat{e}^{(i)}_s$ is the shape embedding of the $j$th part that belongs to the $i$th sample. $L_{contra}$ maximizes the shape similarity between the same capsule across the samples and minimizes the similarity across different capsules. $\tau$ is the hyperparameter utilized to control the discrimination across the negative pairs.

**Sparsity.** To prevent the network from collapsing to use one capsule to describe the whole objects, we employ the spatial sparsity constraint on the visible regions $V^{m}$ of part capsules:

$$L_{sparse} = \text{std}(\sum_{i, j} V_{i,j}^{m}),$$

(10)

where $\text{std}(\cdot)$ calculates the standard deviation, $V^{m}_{i,j}$ is the visibility map of the $m$th capsule at the position $(i, j)$.

**Background Separation.** The prerequisite for unsupervised part discovery is separating foreground and background so that the network can focus on the objects. To achieve that, previous works incorporate salient maps or the ground-truth foreground masks during training. Instead, we use a specific part capsule to model the background. Note that the graphics capsule can recover the 3D information of the objects without any annotation, the foreground map can be easily estimated by setting a threshold to the depth:

$$L_{bg} = \|V^{bg} - \tilde{\mathbf{V}}\|_2, \quad \tilde{\mathbf{V}} = \mathbf{1}_{D_\gamma < \gamma},$$

(11)

where $V^{bg}$ is the visibility map of the part capsule that is used for background estimation, $\tilde{\mathbf{V}}$ is the external region of the object, $\mathbf{D}_\gamma$ is the depth of the object, and $\gamma$ is the threshold for locating the external region.

The final loss functions to train IGC-Net are combined as:

$$L = L_{rec} + \lambda_{contra} L_{contra} + \lambda_{sparse} L_{sparse} + \lambda_{bg} L_{bg},$$

(12)

where $\lambda_{contra}$, $\lambda_{sparse}$ and $\lambda_{bg}$ are the hyper-parameters to balance different loss functions.

### 4. Experiments

#### Implement Details

The image encoder, the capsule decoder, and the lighting module of IGC-Net are composed of convolutional neural networks. We set the number of the part-level graphics capsules $M = 6$, where one of them is used to model the background. Besides, the hyper-parameters for loss combination are set to be $\lambda_{contra} = 10^{-5}$, $\lambda_{sparse} = 10^{-1}$, $\lambda_{bg} = 10^{-1}$. For optimization, we use the Adam optimizer [17] with $10^{-4}$ learning rate to train the networks on a GeForce RTX 3090 for 60 epochs. More training and evaluation details are provided in the supplementary material.

**Datasets.** Following the recent study for the unsupervised face part discovery [35], we evaluate IGC-Net on BP4D [36] and Multi-PIE [10]. Both of these two datasets are captured in the controlled environment. To further validate the capability of tackling the images under real-world scenarios, we adopt the CelebA [21] for experiments, which contains over 200K in-the-wild images of real human faces. In the experiments, BP4D and CelebA are used to evaluate the unsupervised face segmentation and Multi-PIE is used for the interpretable face analysis.

#### 4.1. The Discovered Face Hierarchy

Due to the 3D representations embedded in the graphics capsules, IGC-Net successfully builds the hierarchy of in-the-wild faces that are captured under varied illuminations and poses, shown in Figure 3. By incorporating shape and albedo information as cues, the face images are naturally decomposed into six semantic-consistent parts: background, eyes, mouth, forehead, nose, and cheek, without any human supervision. Each part is described with a specific graphics capsule, which is composed of a set of interpretable parameters including pose, view-independent shape, and view-independent albedo. These parts are assembled by their depth to generate the object capsules as the object-centered representations, building a bottom-up face hierarchy. We also try to discover other numbers of facial parts by controlling $M$ and get reasonable results, shown in the supplementary material.
To show the potential of IGC-Net, we extend our method to image collections of cat faces. Compared with human faces, cat faces are more challenging as cats have more varied textures than humans. The results are shown in Figure 4. It can be seen that the cats are split into background, eyes, ears, nose, forehead, and other skins.

4.2. Analysis of Hierarchical 3D Face Representation

The graphics capsules learned by IGC-Net provide a face hierarchy with explicit graphics descriptions, which gives a plausible way to materialize Marr’s theory [22, 23] that the purpose of vision is building hierarchical 3D representations of objects for recognition. In this section, we apply IGC-Net to validate the advantages of such hierarchical 3D descriptions and uncover the face recognition mechanism of neural networks.

3D Representation vs. 2D Representation. As Marr’s theory reveals [23, 30], the brain should construct the observer-independent object-centered representations of the objects. To evaluate the view-independence of 2D and 3D representations, we compare our method with a 2D autoencoder with the architecture and the training strategy same as ours.

Specifically, both the models are trained on CelebA and tested on the Multi-PIE dataset with yaw variations from -60 to 60. When performing recognition, the embeddings of the autoencoder, and the depth and albedo embeddings of our method are sent to a linear classifier for face recognition to evaluate the consistency of the representations under different views. The results are shown in Figure 5. Firstly, it can be seen from Figure 4(a) that the 3D presentation achieves better accuracy (95.1% vs. 91.2%) in this cross-view recognition task. Secondly, we further analyze the representation consistency across views by computing the similarity matrix of representations under different views. The similarity matrices of the 2D and 3D representations are subtracted and shown as a heatmap. The score higher than 0 indicates 3D is better than 2D.
are subtracted (3D minus 2D) and shown as a heatmap in Figure 4(b). It can be seen that our method shows better consistency, especially when matching images across large views, i.e., $30^\circ$ vs $-60^\circ$.

**Shape vs. Albedo.** To show the potential of our method for interpretable analysis, we design an experiment to explore which part-level graphics capsule is crucial for face recognition and which component (shape or albedo) of the capsule is more important. Specifically, we assign the part-level shape embeddings $\{\hat{e}^m\}$ and albedo embeddings $\{\hat{e}^a\}$ with trainable scalar $\{\omega^m\}$ and $\{\omega^a\}$ as the attention weights. The weight parameters $\{\omega^m, \omega^a\}$ are sent to a linear classifier for face recognition. After training with L1 penalization for sparsity, the attention weights of part capsules are shown in Figure 6. By summarizing the attention weights of different parts, we can see that the albedo ($\omega = 0.70$) is more important than the shape ($\omega = 0.34$) for face recognition. Besides, the part-level attention weights also show that the albedo of the eyes is the most important component and the shape of the nose is more important than the shape of other parts, which is consistent with the previous conclusions [33, 35].

![Figure 6. The importance of part-level graphics capsules for face recognition on Multi-PIE. On average, the albedo is more crucial than the shape when recognizing faces. The albedo of the eyes is the most important component and the shape of the nose is more important than the shape of other parts.](image)

### 4.3. Unsupervised Face Segmentation

To execute the quantitative and qualitative evaluation, we treat the silhouettes of parts as segment maps and apply them to the unsupervised face segmentation task. Note that there is no ground truth for the unsupervised part-level segmentation, the key of this task is to evaluate the semantic consistency of the parsing manners. The following experiments show the superiority of our method.

**Baselines.** Learning to segment the face parts from the unlabeled images is a challenging task as parts are difficult to be described by math. In this paper, we compare our method with the state-of-art methods for unsupervised face segmentation, including DFF [7], SCOPS [14] and HP-Capsule [35]. To discover the semantic parts, DFF proposes to execute the non-negative matrix upon the CNN features, which need to optimize the whole dataset to get the segment results. SCOPS proposes a framework with the concentration loss to constrain the invariance of images between TPS transformation. However, due to the lack of effective constraints, their results tend to assign similar silhouettes to different parts. HP-Capsule proposes a bottom-up schedule to aggregate parts from subparts, whose parts are described with interpretable parameters. However, their descriptions are defined in the 2D space, limiting their capacity to tackle faces with large poses.

**Quantitative Comparison.** Following the previous work [35], we utilize the Normalized Mean Error (NME) of the landmarks predicted by segment maps to evaluate the quality of the parsing manners. Specifically, $\text{NME}_L$ treats the centroid of the segment maps as landmarks and uses linear mapping to convert them to human-annotated landmarks. $\text{NME}_{DL}$ incorporates a shallow network to directly predict the landmarks from the segment maps. Table 1 and

![Figure 7. The qualitative comparison of unsupervised face segmentation on CelebA.](image)

![Figure 8. The qualitative comparison of unsupervised face segmentation on BP4D.](image)
Table 1. The quantitative comparison of unsupervised face segmentation on CelebA. NME_L(%) and NME_DL(%) use the landmarks estimated from the segment maps to evaluate the semantic consistency of parts.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>NME_L</th>
<th>NME_DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFF [7]</td>
<td>22.78</td>
<td>27.27</td>
</tr>
<tr>
<td>SCOPS [14]</td>
<td>18.72</td>
<td>23.69</td>
</tr>
<tr>
<td>HP-Capsule [35]</td>
<td>21.25</td>
<td>25.27</td>
</tr>
<tr>
<td>IGC-Net (ours)</td>
<td>11.84</td>
<td>18.88</td>
</tr>
</tbody>
</table>

Table 2. The quantitative comparison of unsupervised face segmentation on BP4D.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>NME_L</th>
<th>NME_DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFF [7]</td>
<td>18.85</td>
<td>12.26</td>
</tr>
<tr>
<td>SCOPS [14]</td>
<td>9.10</td>
<td>6.74</td>
</tr>
<tr>
<td>HP-Capsule [35]</td>
<td>8.81</td>
<td>6.10</td>
</tr>
<tr>
<td>IGC-Net (ours)</td>
<td>6.35</td>
<td>4.32</td>
</tr>
</tbody>
</table>

Figure 9. The qualitative ablation study on CelebA. It can be seen that the semantic consistency will be damaged without the one-hot operation in GDM (see Eqn. 6) and the contrastive semantic constraint $L_{contra}$ (see Eqn. 9) is important for discovering meaningful parts.

4.4. Ablation Studies

The basis of building the hierarchy of objects is to learn the parts with explicit semantics and keep semantic consistency across different samples. In this section, we perform the ablation study to show the importance of the one-hot operation in the GDM (see Eqn. 6) and the semantic constraint $L_{contra}$ (see Eqn. 9) for discovering meaningful parts. Figure 9 shows the qualitative ablation study on CelebA. In the second row of Figure 9, it can be seen that, without the one-hot operation to prevent the information leakage of different parts, the semantic consistency across samples will be damaged. The third row of Figure 9 shows that the contrastive semantic constraint $L_{contra}$ is important for the discovery of parts with salient semantics. Without such constraint, the segmentation of the important organs such as the eyes will have ambiguity. These conclusions are also validated by the quantitative ablation study shown in Table 3.

5. Conclusion and Discussion

In this paper, we propose the IGC-Net to learn the hierarchical 3D face representations from large-scale unlabeled in-the-wild images, whose primitive is the graphics capsule that contains the 3D representations with explicit meanings. By combining depth and albedo information as cues, IGC-Net successfully decomposes the objects into a set of part-level graphics capsules and constructs the hierarchy of objects by assembling the part-level capsules into object-level capsules. IGC-Net reveals how the neural networks, oriented at visual perception, understand faces as a hierarchy of 3D models. Besides, the part-level graphics descriptions can be used for unsupervised face segmentation and interpretable face analysis. Experiments on CelebA, BP4D, and Multi-PIE validate the effectiveness and the interpretability of our method.

Acknowledgement. This work was supported in part by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects #62276254, #62176256, #62206280, the Youth Innovation Promotion Association CAS (#Y2021131), and the InnoHK program.
References


