Graphics Capsule: Learning Hierarchical 3D Face Representations from 2D Images

Chang Yu[1,2], Xiangyu Zhu[1,2]*, Xiaomei Zhang[1,2], Zhaoxiang Zhang[1,2,3], Zhen Lei[1,2,3]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation,
Chinese Academy of Sciences
{chang.yu, xiangyu.zhu, zlei}@nlpr.ia.ac.cn
{zhangxiaomei2016, zhaoxiang.zhang}@ia.ac.cn

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

*Corresponding author.

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.

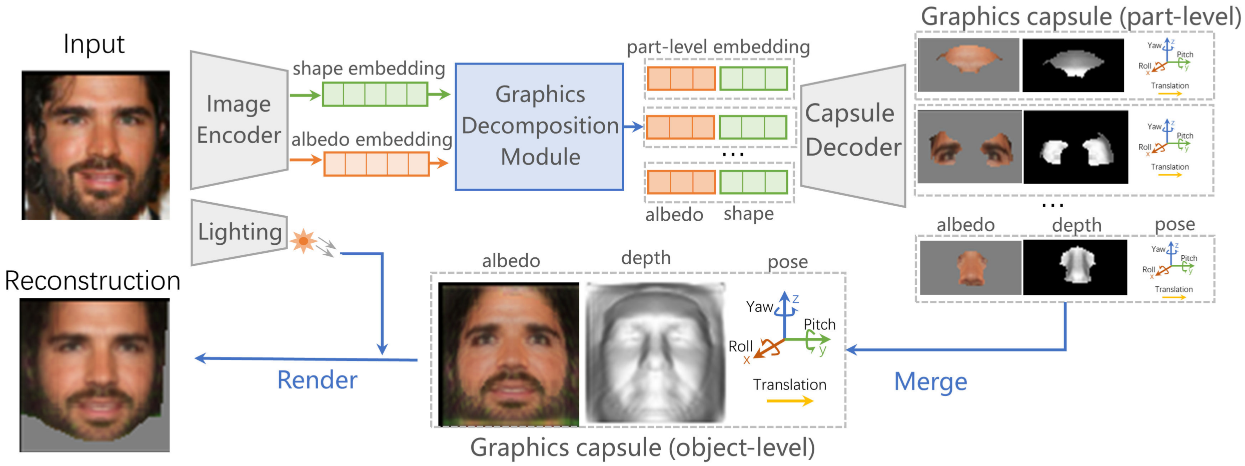


Figure 1. Overview of the Inverse Graphics Capsule Network (IGC-Net). The input image is first encoded to global shape and albedo embeddings and then sent to a decomposition module to get the spatially-decoupled part-level graphics capsules. Afterward, these capsules are decoded to get the explicit 3D 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.

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 pre-defined 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 $\mathbf{D} \in \mathbb{R}^{H\times W}$, an albedo map $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ and 3D pose parameters $\mathbf{p} \in \mathbb{R}^{1\times 6}$ (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 $\mathbf{I}$ into the shape and the albedo embeddings $\mathbf{f}_s$ and $\mathbf{f}_a$:

$$\mathbf{f}_s, \mathbf{f}_a = \text{ImageEncoder}(\mathbf{I}). \tag{1}$$

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:

$$\begin{aligned} \{\hat{\mathbf{e}}_s^1, ..., \hat{\mathbf{e}}_s^M\}, \{\hat{\mathbf{e}}_a^1, ..., \hat{\mathbf{e}}_a^M\} &= \text{GDM}(\mathbf{f}_s, \mathbf{f}_a), \\ \{\mathbf{D}_p^m, \mathbf{A}_p^m, \mathbf{p}_p^m\} &= \text{GraphicsDecoder}(\hat{\mathbf{e}}_s^m, \hat{\mathbf{e}}_a^m), \end{aligned} \tag{2}$$

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

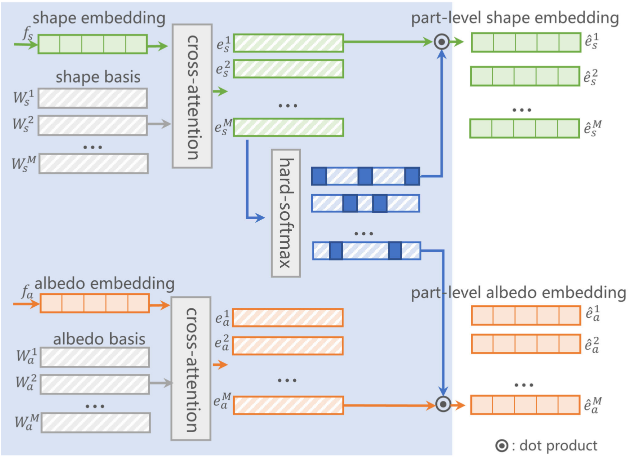


Figure 2. Illustration of the Graphics Decomposition Module (GDM). GDM is proposed to ensure that each part capsule presents a semantic-consistent part of objects.

depth to generate the global object capsule:

$$
\begin{aligned}
\mathbf{V}^m(i,j) &= \mathbf{1}_{m=\underset{n}{\operatorname{argmin}}(\mathbf{D}_p^n(i,j))}, \\
\mathbf{D}_o &= \sum\nolimits_m \mathbf{V}^m \odot \mathbf{D}_p^m, \\
\mathbf{A}_o &= \sum\nolimits_m \mathbf{V}^m \odot \mathbf{A}_p^m, \\
\mathbf{p}_o &= \frac{1}{M} \sum\nolimits_m \mathbf{p}_p^m,
\end{aligned}
\tag{3}
$$

where $V_{i,j}^m$ is the visibility map of the mth part capsule at the position (i,j), $\odot$ is the element-wise production, and $\Theta_o : \{\mathbf{D}_o, \mathbf{A}_o, \mathbf{p}_o\}$ is the object capsule. During assembly, a capsule is visible at (i,j) only when its depth is smaller than the others. The part-level depth and albedo maps are multiplied with their visibility maps and aggregated as one, respectively, and the object pose is the average of part poses. In the object capsule, both the depth $\mathbf{D}_o$ and albedo $\mathbf{A}_o$ are defined in the canonical space, and the pose $\mathbf{p}_o$ is used to project the 3D object to the image plane. Finally, by estimating the lighting $\mathbf{l}$ with another module similar to [34], the recovered image $\hat{\mathbf{I}}$ is generated by the differentiable rendering Λ [16]:

$$
\hat{\mathbf{I}} = \Lambda(\mathbf{D}, \mathbf{A}, \mathbf{p}, \mathbf{l}). \tag{4}
$$

When training IGC-Net, we can minimize the distance between the input image $\mathbf{I}$ and the reconstructed image $\hat{\mathbf{I}}$ following the analysis-by-synthesis strategy, so that the network parameters can be learned in an unsupervised manner.

3.2. Graphics Decomposition Module

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 $\{\mathbf{W}_s^m\}$ as the implicit part templates. Given the global embeddings $\mathbf{f}_s$ extracted in Eqn. 1, GDM performs cross attention between the global embedding and the basis to get M disentangled D dimensional embeddings:

$$
\mathbf{e}_s^m = \mathbf{f}_s \mathbf{W}_s^m, \quad m = 1, ..., M. \tag{5}
$$

To further reduce the entanglement between $\{\mathbf{e}_s^m\}$ 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:

$$
\begin{aligned}
&\hat{\mathbf{e}}_s^m = \mathbf{e}_s^m \odot \mathbf{M}_{[m,:]}, \\
&\mathbf{M}_{[:,d]} = \text{hard\_softmax}([\mathbf{e}_s^1(d), \mathbf{e}_s^2(d), ..., \mathbf{e}_s^M(d)]), \\
&\text{hard\_softmax}(\mathbf{e}) = \frac{\mathbf{e}}{\sum_i \mathbf{e}(i)} \odot \text{ onehot}(\frac{\mathbf{e}}{\sum_i \mathbf{e}(i)}),
\end{aligned}
\tag{6}
$$

where $\mathbf{M}_{M \times D}$ is the attention matrix, whose mth row is $\mathbf{M}_{[m,:]}$ and dth column is $\mathbf{M}_{[:,d]}$, $\mathbf{e}_s^m(d)$ is the dth dimension of the embedding $\mathbf{e}_s^m$, $\text{onehot}(\cdot)$ is the one-hot operation, and $\hat{\mathbf{e}}_s^m$ is the final part-level shape embedding. The same pipeline is applied to the albedo embeddings, where the only difference is that the attention $\mathbf{M}$ 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 $\hat{\mathbf{I}}$:

$$
\begin{aligned}
\mathcal{L}_{rec} = &-\frac{1}{|\Omega|} \sum \ln \frac{1}{\sqrt{2}\sigma} \exp -\frac{\sqrt{2}|\hat{\mathbf{I}} - \mathbf{I}|}{\sigma} \\
&-\frac{1}{|\Omega|} \sum \ln \frac{1}{\sqrt{2}\sigma} \exp -\frac{\sqrt{2}|\hat{\mathbf{I}}_{flip} - \mathbf{I}|}{\sigma},
\end{aligned}
\tag{7}
$$

where Ω 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:

$$\begin{aligned}\mathcal{L}_{rec} = &-\frac{1}{|\Omega^{(k)}|}\sum \ln \frac{1}{\sqrt{2}\sigma^{(k)}} \exp - \frac{\sqrt{2}|f^{(k)}(\hat{\mathbf{I}}) - f^{(k)}(\mathbf{I})|}{\sigma^{(k)}} \\ &- \frac{1}{|\Omega^{(k)}|}\sum \ln \frac{1}{\sqrt{2}\sigma^{(k)}} \exp - \frac{\sqrt{2}|f^{(k)}(\hat{\mathbf{I}}_{flip}) - f^{(k)}(\mathbf{I})|}{\sigma^{(k)}},\end{aligned} \tag{8}$$

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 $\hat{\mathbf{e}}_s^m$ of each capsule, which is formulated as:

$$\begin{aligned}\mathcal{L}_{contra} = &-\sum_{b=1}^{B}\sum_{m=1}^{M} \log \\ &\frac{\sum_{i\neq b}\exp(e_s^{m,(b)} \cdot e_s^{m,(i)}/\tau)}{\sum_{i\neq b}\exp(e_s^{m,(b)} \cdot e_s^{m,(i)}/\tau) + \sum_{j\neq m}\sum_{i\neq b}\exp(e_s^{m,(b)} \cdot e_s^{j,(i)}/\tau)},\end{aligned} \tag{9}$$

where B is the batch size, M is the number of part capsules, $\hat{\mathbf{e}}_s^{j,(i)}$ is the shape embedding of the jth part that belongs to the ith sample. $\mathcal{L}_{contra}$ maximizes the shape similarity between the same capsule across the samples and minimizes the similarity across different capsules. τ 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:

$$\mathcal{L}_{sparse} = \mathrm{std}(\sum_{i,j}\mathbf{V}_{i,j}^m), \tag{10}$$

where $\mathrm{std}(\cdot)$ calculates the standard deviation, $\mathbf{V}_{i,j}^m$ is the visibility map of the mth 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:

$$\mathcal{L}_{bg} = \|\mathbf{V}^{bg} - \widetilde{\mathbf{V}}\|, \;\; \widetilde{\mathbf{V}} = \mathbf{1}_{\mathbf{D}_o<\gamma}, \tag{11}$$

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

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

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{rec} + \lambda_{contra}\mathcal{L}_{contra} + \lambda_{sparse}\mathcal{L}_{sparse} \\ &+ \lambda_{bg}\mathcal{L}_{bg},\end{aligned} \tag{12}$$

where λ_{contra}, λ_{sparse} and λ_{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.

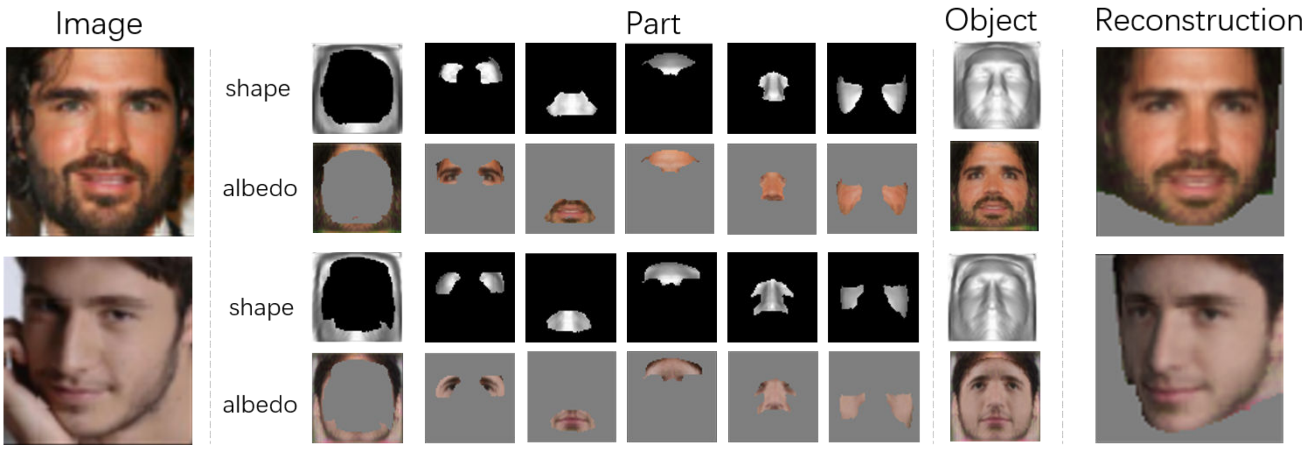


Figure 3. Illustration of the discovered face hierarchy with 3D descriptions. By incorporating shape and albedo information as cues, IGC-Net decomposes the images into six parts: background, eyes, mouth, forehead, nose, and cheek.

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.

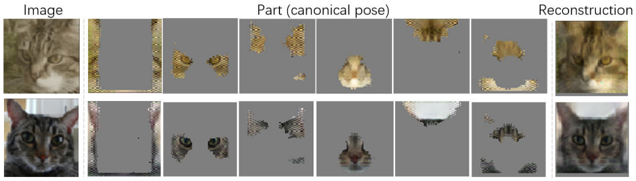
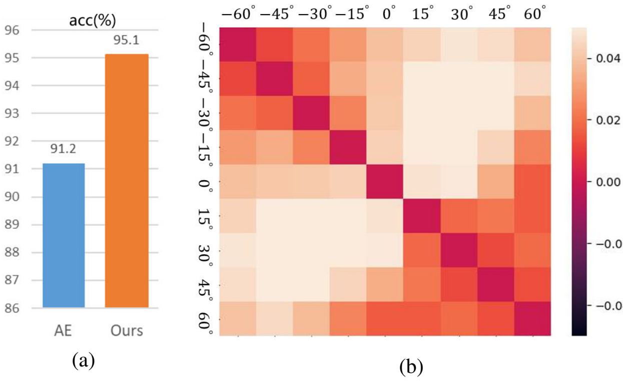


Figure 4. The discovered hierarchy of cats. The cat faces are split into background, eyes, ears, nose, forehead, and other skins.

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



(a)

(b)

Figure 5. The comparison of the representation consistency under different views between 2D representations and 3D representations on Multi-PIE. (a) The recognition accuracy. The representations are sent to a linear classifier for classification. (b) 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° vs -60°.

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 $\{\hat{\mathbf{e}}_s^m\}$ and albedo embeddings $\{\hat{\mathbf{e}}_a^m\}$ with trainable scalar $\{\omega_s^m\}$ and $\{\omega_a^m\}$ as the attention weights. The weight parameters $\{\omega^m \hat{\mathbf{e}}^m\}$ 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].

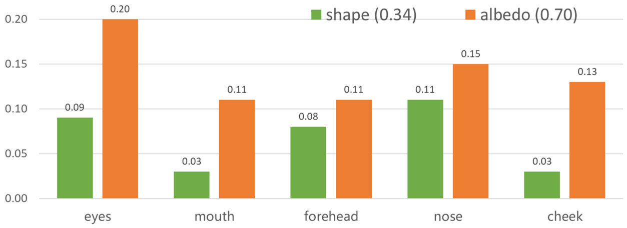


Figure 6. The importance of part-level graphics capsules for face recognition on Multi-PIE. On average, the albedo is more crucial than the shape when recognizing faces. The albedo of the eyes is the most important component and the shape of the nose is more important than the shape of other parts.

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.

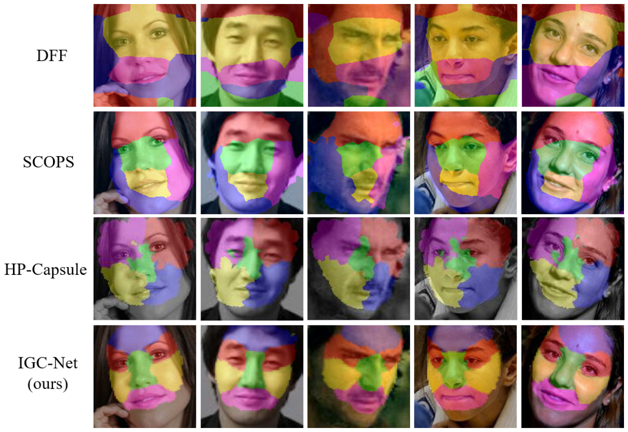


Figure 7. The qualitative comparison of unsupervised face segmentation on CelebA.

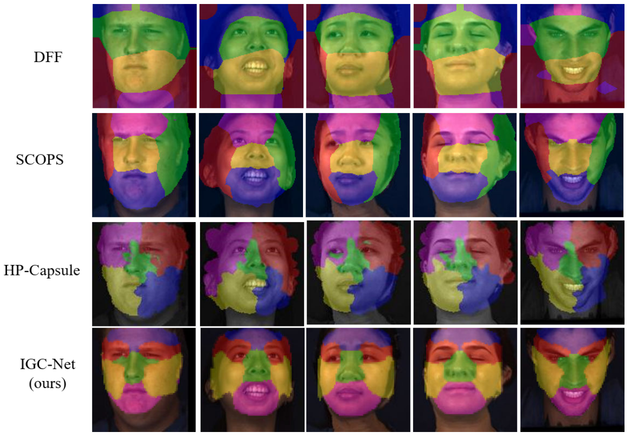


Figure 8. The qualitative comparison of unsupervised face segmentation on BP4D.

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, $\mathrm{NME}_{\mathrm{L}}$ treats the centroid of the segment maps as landmarks and uses linear mapping to convert them to human-annotated landmarks. $\mathrm{NME}_{\mathrm{DL}}$ incorporates a shallow network to directly predict the landmarks from the segment maps. Table 1 and

Table 1. The quantitative comparison of unsupervised face segmentation on CelebA. $\text{NME}_{\text{L}}(\%)$ and $\text{NME}_{\text{DL}}(\%)$ use the landmarks estimated from the segment maps to evaluate the semantic consistency of parts.

METHOD	NME_{L}	NME_{DL}
DFF [7]	22.78	27.27
SCOPS [14]	18.72	23.69
HP-Capsule [35]	21.25	25.27
IGC-Net (ours)	**11.84**	**18.88**

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

METHOD	NME_{L}	NME_{DL}
DFF [7]	18.85	12.26
SCOPS [14]	9.10	6.74
HP-Capsule [35]	8.81	6.10
IGC-Net (ours)	**6.35**	**4.32**

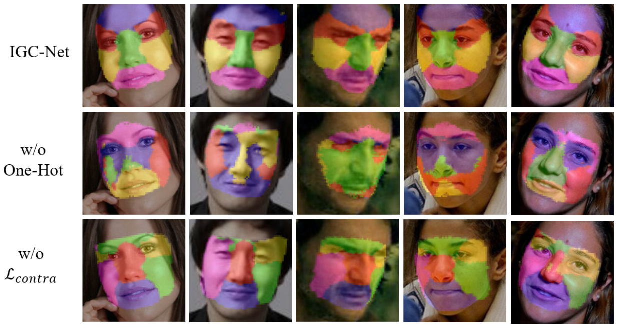


Figure 9. The qualitative ablation study on CelebA. It can be seen that the semantic consistency will be damaged without the one-hot operation in GDM (see Eqn. 6) and the $\mathcal{L}_{contra}$ (see Eqn. 9) is important for discovering parts with salient semantics.

Table 2 show the quantitative comparison results on CelebA and BP4D, which validate the effectiveness of our method.

Qualitative Comparison. The qualitative comparison results are shown in Figure 7 and Figure 8. It can be seen that our method performs better than other methods. The results of DFF don't successfully separate the foreground and the background. As for SCOPS, due to the lack of effective constraints, the segment maps of SCOPS are with some ambiguity, where the organs with salient semantics are assigned to different parts for different samples. For example, SCOPS sometimes takes the right eye as the green part (the fifth column in Figure 7) while sometimes splitting it from the middle (the first and the second column in Figure 7). The segment boundaries of HP-Capsule are clearer than DFF and SCOPS. However, as shown in the third column of Figure 7, limited by their 2D descriptions, HP-Capsule fails on the faces with large poses while our method performs well on these challenging samples.

Table 3. The quantitative ablation study on CelebA. The results show the importance of the one-hot operation in GDM and the semantic constraint $\mathcal{L}_{contra}$.

One-Hot	$\mathcal{L}_{contra}$	NME_{L}
	✓	19.10
✓		13.46
✓	✓	**11.84**

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 $\mathcal{L}_{contra}$ (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 $\mathcal{L}_{contra}$ 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

[1] H Clark Barrett. A hierarchical model of the evolution of human brain specializations. *Proceedings of the national Academy of Sciences*, 109(supplement_1):10733–10740, 2012. 2

[2] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 1

[3] Irving Biederman and Peter C Gerhardstein. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human perception and performance*, 19(6):1162, 1993. 1

[4] Anna Bodegård, Stefan Geyer, Christian Grefkes, Karl Zilles, and Per E Roland. Hierarchical processing of tactile shape in the human brain. *Neuron*, 31(2):317–328, 2001. 2

[5] Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020. 3

[6] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021. 2

[7] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352, 2018. 2, 7, 8

[8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[9] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986. 2

[10] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 5

[11] Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 3(3):231–250, 1979. 1

[12] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51. Springer, 2011. 2

[13] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018. 2

[14] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019. 2, 7, 8

[15] Iiro P Jääskeläinen, Enrico Glerean, Vasily Klucharev, Anna Shestakova, and Jyrki Ahveninen. Do sparse brain activity patterns underlie human cognition? *NeuroImage*, page 119633, 2022. 2

[16] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 4

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Adam Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton. Stacked capsule autoencoders. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3

[19] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. 2022. 1

[20] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Unsupervised part segmentation through disentangling appearance and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8355–8364, 2021. 2

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[22] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 1, 6

[23] David Marr and Herbert Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294, 1978. 1, 6

[24] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 3

[25] Tomaso Poggio and Shimon Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266, 1990. 1

[26] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017. 2

[27] Sara Sabour, Andrea Tagliasacchi, Soroosh Yazdani, Geoffrey Hinton, and David J Fleet. Unsupervised part representation by flow capsules. In *International Conference on Machine Learning*, pages 9213–9223. PMLR, 2021. 2

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[29] Manish Singh and Donald D Hoffman. Part-based representations of visual shape and implications for visual cognition. In *Advances in psychology*, volume 130, pages 401–459. Elsevier, 2001. 1

[30] NS Sutherland. The representation of three-dimensional objects. *Nature*, 278(5703):395–398, 1979. 6

[31] James W Tanaka and Diana Simonyi. The “parts and wholes” of face recognition: A review of the literature. *Quarterly Journal of Experimental Psychology*, 69(10):1876–1889, 2016. 2

[32] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 3

[33] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. In *European conference on computer vision*, pages 248–263. Springer, 2020. 7

[34] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 3, 4, 5

[35] Chang Yu, Xiangyu Zhu, Xiaomei Zhang, Zidu Wang, Zhaoxiang Zhang, and Zhen Lei. Hp-capsule: Unsupervised face part discovery by hierarchical parsing capsule network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4032–4041, 2022. 1, 2, 3, 5, 7, 8

[36] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 5

[37] Zhenyu Zhang, Yanhao Ge, Renwang Chen, Ying Tai, Yan Yan, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning to aggregate and personalize 3d face from in-the-wild photo collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14224, 2021. 3

[38] Zhenyu Zhang, Yanhao Ge, Ying Tai, Weijian Cao, Renwang Chen, Kunlin Liu, Hao Tang, Xiaoming Huang, Chengjie Wang, Zhifeng Xie, et al. Physically-guided disentangled implicit rendering for 3d face modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20353–20363, 2022. 3

[39] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1097–1106, 2019. 3