Shigeki *et al. IPSJ Transactions on Computer Vision and Applications* (2018) 10:16
https://doi.org/10.1186/s41074-018-0052-9

IPSJ Transactions on Computer Vision and Applications

EXPRESS PAPER Open Access

Estimating 3D human shape under clothing from a single RGB image

Yui Shigeki[1†], Fumio Okura[1*†], Ikuhisa Mitsugami[2] and Yasushi Yagi[1]

Abstract

Estimation of naked human shape is essential in several applications such as virtual try-on. We propose an approach that estimates naked human 3D pose and shape, including non-skeletal shape information such as musculature and fat distribution, from a single RGB image. The proposed approach optimizes a parametric 3D human model using person silhouettes with clothing category, and statistical displacement models between clothed and naked body shapes associated with each clothing category. Experiments demonstrate that our approach estimates human shape more accurately than a prior method.

Keywords: Human shape modeling, Parametric 3D human model, Cloth-skin displacement modeling

1 Introduction

We propose a novel approach to estimate three-dimensional (3D) human body shapes under clothing from a single RGB image. Human shape is an essential element in the computer vision field. Humans are usually captured with clothing, that is, the actual (naked) human shape is concealed by the clothing and usually differs from the appearance. Virtual try-on systems, an important application of naked shapes, typically superimpose 3D models of clothing onto human shape [1]. Although it allows the consumer to visualize whether an item of clothing suits them with regard to design and color, it is difficult to determine how well the size of the item fits their actual body shape.

The estimation of naked human shape has been studied, but most approaches employ multi-view images or 3D scanners (e.g. [2–4]) for acquiring 3D shapes. These approaches are impractical for actualizing shape acquisition for virtual try-on at home, which requires easy input using commodity cameras. Our approach requires the input of only a single image and outputs parameters of 3D human body shape.

The proposed approach is built on a previous single-image human 3D modeling method (SMPLify [5]), which optimizes a parametric model of 3D human pose and shape [6] to fit joint positions acquired by a joint estimation method [7]. Joint-based optimization does capture aspects of shape information related to human skeletons such as the length of the arms and legs. However, in principle, joint locations do not include non-skeletal information such as musculature and fat distribution.

The proposed method estimates 3D human pose and shape including non-skeletal shape information under clothing. We optimize the parametric 3D human model using a single-image human silhouette with clothing region segmentation while considering pre-constructed statistics of the displacement by clothing (i.e., the distance between the naked and clothed contours) for each clothing category. The displacement modeling is a significant challenge in our study since collecting a dataset of image pairs of clothed and naked people is unfeasible. We, therefore, model the displacement based on clothed person shapes generated from naked silhouettes by a clothing simulator.

2 3D shape estimation under clothing

2.1 Overview

As shown in Fig. 1, the proposed approach optimizes the Skinned Multi-Person Linear (SMPL) [6] pose and shape parameters under clothing, with only a single RGB image as input. Similar to SMPLify [5], our approach optimizes the parameters of a SMPL model [6], which consists of 72 pose parameters (3 orientations for 23 joints + 3

*Correspondence: okura@am.sanken.osaka-u.ac.jp
†Yui Shigeki and Fumio Okura contributed equally to this work.
[1]The Institute of Scientific and Industrial Research, Osaka University, 8-1, Mihogaoka, Ibaraki, Osaka, Japan
Full list of author information is available at the end of the article

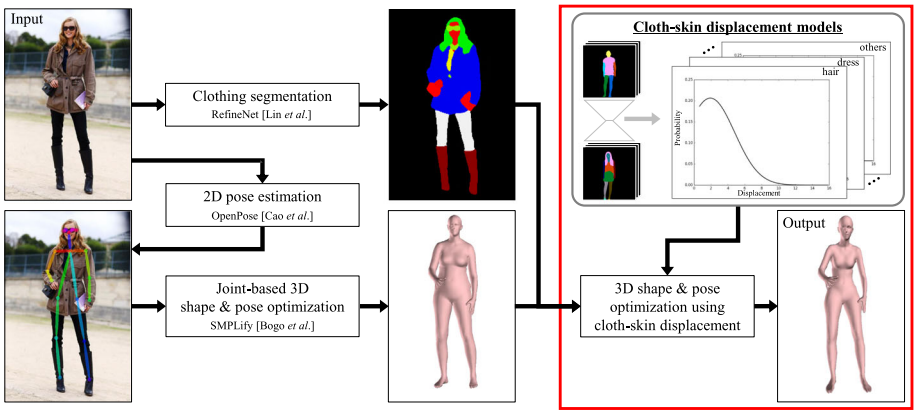

Fig. 1 Overview. Parametric model of 3D human pose and shape [6] is initialized using joint-based optimization [5] and adjusted to fit a person's silhouette with contour displacement

root orientation) and 10 linear shape coefficients. Initially using SMPLify [5] to obtain a joint-based optimization result, we further optimize the parameters using silhouette shape and the cloth-skin displacement model created for each clothing category.

2.2 Clothing segmentation

Given an input image, such as a photograph, we first perform a semantic segmentation to extract both a human silhouette and a clothing category. For this step, we utilize RefineNet [8], a semantic segmentation approach which successfully outputs high-resolution results for human part estimation. To train RefineNet, we utilize an image dataset with clothing segmentation, Clothing Co-Parsing (CCP) dataset [9] and Fashionista dataset [10], where each pixel is labeled by clothing categories. We re-classify the clothing labels into 11 categories: "background," "skin," "hair," "inner wear," "outer wear," "skirt," "dress," "pants," "shoes," "bag," and "others" and train RefineNet using 1500 images from the dataset.

2.3 Cloth-skin displacement modeling

Modeling the displacement between clothing and skin is a core part of this study. Given the impracticality of collecting a large dataset of pairs of clothed and naked person images, we employ an artificial dataset generated by a conditional variational auto-encoder, conditional sketch module (CSM) in [11], as shown in Fig. 2. We create image pairs of clothed and naked person silhouettes by inputting various silhouettes of the SMPL human body to the CSM network. For each image pair of clothed and naked person silhouettes, we compute the displacement on the clothed and naked silhouette contours. We create a distribution of the amount of displacement for each clothing category, based on the category labels of clothed silhouettes. We fit a truncated normal distribution pdf_c for the displacement distribution of each clothing category c, using maximum-likelihood estimation. The probability returned by $\mathrm{pdf}_\mathrm{c}(d)$ becomes zero when d is smaller than the lower bound $\alpha_{\mathrm{pdf}_\mathrm{c}}$, which is optimized via the maximum-likelihood estimation, since naked body contours are never on the exterior of clothing.

2.4 Fitting parametric human 3D model

Given a person silhouette with associated clothing category (see Section 2.2), joint locations, and cloth-skin displacement models (see Section 2.3), the proposed approach estimates the pose and shape through an optimization of SMPL model parameters. The initial SMPL parameters are acquired as the result of a joint-based optimization method, SMPLify [5], where joint locations on the input image are estimated by a CNN-based 2D joint estimation approach, OpenPose [12] trained using MS COCO dataset [13].

Here, SMPL consists of 72-dimensional pose (joint angles and root orientation) parameters θ and 10-dimensional linear shape coefficients β. The "ideal" pose can change during the optimization of the shape parameters β; we therefore jointly optimize both β and θ.

2.4.1 *Optimization*

Our objective function for the optimization is defined as follows:

$$E_{\mathrm{shape}}(\beta) = \lambda_\mathrm{s}E_\mathrm{s}(\beta,\theta) + \lambda_\mathrm{c}E_\mathrm{c}(\beta,\theta), \quad (1)$$

$$E_{\mathrm{pose}}(\theta) = \lambda_\mathrm{s}E_\mathrm{s}(\beta,\theta) + \lambda_\mathrm{c}E_\mathrm{c}(\beta,\theta) + \lambda_\mathrm{j}E_\mathrm{j}(\beta,\theta) + \lambda_\mathrm{a}E_\mathrm{a}(\theta) + \lambda_\mathrm{sp}E_\mathrm{sp}(\beta,\theta) + \lambda_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\theta). \quad (2)$$

E_{shape} and E_{pose} respectively denote the objective functions for optimizing shape β and pose θ parameters. Lambdas λ_s, λ_c, λ_j, λ_a, λ_sp, and $\lambda_{\boldsymbol{\theta}}$ are weights for each term. We alternatively minimize the objective terms: minimizing $E_{\mathrm{shape}}(\beta)$ using fixed θ and vice versa.

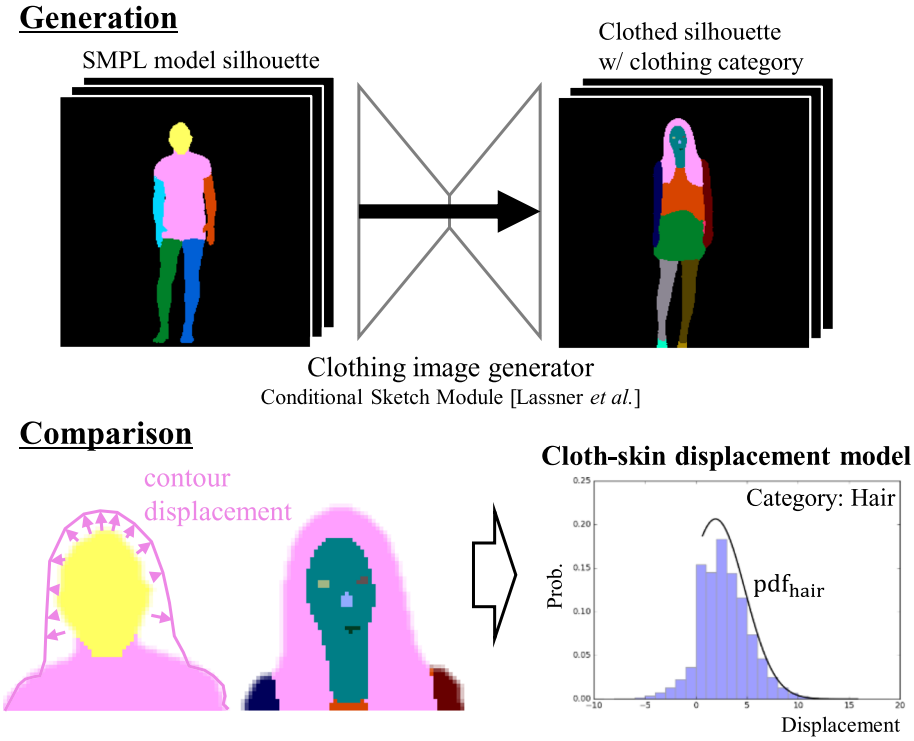

Fig. 2 Cloth-skin displacement modeling. Computing displacement between naked and clothed contours generated by CSM [11] and fitting a probability density function

$E_j(\beta,\theta)$, $E_a(\theta)$, $E_{sp}(\beta,\theta)$, and $E_{\boldsymbol{\theta}}(\theta)$ are cost terms identical to those utilized in SMPLify. The term $E_j(\beta,\theta)$ is a distance between 2D joints on the input image and the joints in the estimated SMPL model projected onto the image plane. For the other terms, refer to [5] for details.

The proposed approach employs cost terms for skin contours $E_s(\beta,\theta)$ and clothed contours $E_c(\beta,\theta)$. Let $\mathcal{S}_{\text{in}}$ be a point set on the person silhouette contour of the input image, which is fixed during the optimization, and $\mathcal{S}_{\text{SMPL}}$ be a point set on the corresponding SMPL silhouette contour, which is a variable that depends on β and θ to be optimized. The cost terms utilize nearest-neighbor correspondences from $\mathcal{S}_{\text{in}}$ to $\mathcal{S}_{\text{SMPL}}$,

$$\mathcal{S}_{\text{SMPL,c}} = \bigcup_{\mathbf{p}} \left(\text{NN}_{\mathcal{S}_{\text{SMPL}}}\left(\mathbf{p} \in \mathcal{S}_{\text{in,c}}\right)\right), \tag{3}$$

where $c \in \mathcal{C}$ is a region label for the foreground categories $\mathcal{C} = \{\text{skin, hair, ...}\}$[1]. Thus, $\mathcal{S}_{\text{in,c}} \subset \mathcal{S}_{\text{in}}$ denote input contour points labeled as category c. The mapping function $\text{NN}_{\mathcal{S}_{\text{SMPL}}}(\mathbf{p} \in \mathcal{S}_{\text{in,c}})$ acquires the nearest-neighbor point of $\mathbf{p} \in \mathcal{S}_{\text{in,c}}$ from $\mathcal{S}_{\text{SMPL}}$.

Skin contour cost E_s This term controls the behavior of SMPL silhouette contours $\mathcal{S}_{\text{SMPL,skin}}$, where the corresponding input contour points $\mathcal{S}_{\text{in,skin}}$ are labeled as skin region. The cost term penalizes the 2D Euclidean distance between the corresponding points in $\mathcal{S}_{\text{SMPL,skin}}$ and $\mathcal{S}_{\text{in,skin}}$:

$$E_s = \frac{1}{n_{\mathcal{S}_{\text{in}}}} \sum_{\mathbf{p} \in \mathcal{S}_{\text{in,skin}}} ||\mathbf{p} - \text{NN}_{\mathcal{S}_{\text{SMPL}}}(\mathbf{p})||, \tag{4}$$

where $n_{\mathcal{S}_{\text{in}}}$ denotes the number of points in $\mathcal{S}_{\text{in}}$, which normalizes the cost.

Cloth contour cost E_c The cost function E_c controls the behavior of contours not labeled as "skin" so that the contours located inside the input contour as much as the cloth-skin displacement described in Section 2.3. Letting $\mathcal{C}_{\text{cloth}} = \mathcal{C} - \{\text{skin}\}$, E_c is defined as the sum of cost terms for each clothing category, calculated based on contour distances:

$$E_c = \frac{1}{n_{\mathcal{S}_{\text{in}}}} \sum_{c \in \mathcal{C}_{\text{cloth}}} \sum_{\mathbf{p} \in \mathcal{S}_{\text{in,c}}} d_{\mathbf{p}}. \tag{5}$$

Here, let d_s denote the signed Euclidean distance between $\mathbf{p}$ and $\text{NN}_{\mathcal{S}_{\text{SMPL}}}(\mathbf{p})$, where the distance becomes positive if an input contour point $\mathbf{p}$ is outside of the contour of the corresponding SMPL contour point $\text{NN}_{\mathcal{S}_{\text{SMPL}}}(\mathbf{p})$. Accordingly, our distance function $d_{\mathbf{p}}$ which

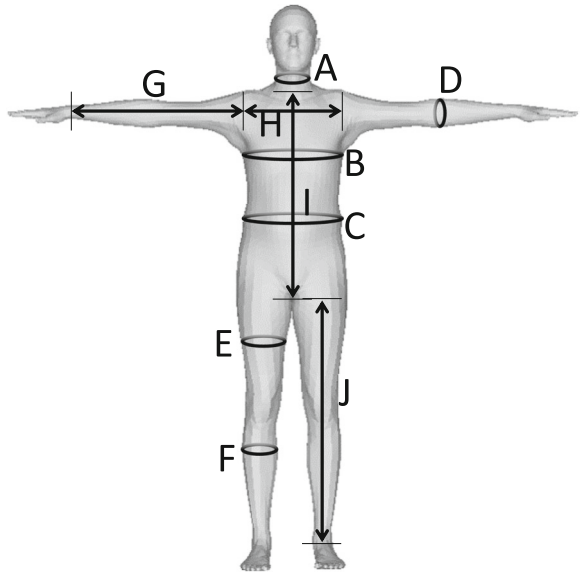

Fig. 3 Body measurements. Ten body measurements for the cloth-skin displacement modeling and the experiments

considering the cloth-skin displacement is defined as follows:

$$d_{\mathbf{p}} = \begin{cases} -\log\left(\mathrm{pdf}_c\left(d_{\mathrm{s}}\right)+\epsilon\right) & \left(d_{\mathrm{s}} \geq \alpha_{\mathrm{pdf}_c}\right) \\ \lambda_l(\alpha_{\mathrm{pdf}_c} - d_{\mathrm{s}}) - \log(\mathrm{pdf}_c(\alpha_{\mathrm{pdf}_c})+\epsilon) & \left(d_{\mathrm{s}} < \alpha_{\mathrm{pdf}_c}\right) \end{cases}, \quad (6)$$

where ϵ is a small constant to avoid $\log(0) = -\inf$. Here, $\mathrm{pdf}_{c\in\mathcal{C}_{\mathrm{cloth}}}$ denotes the truncated normal distribution modeled in Section 2.3, which returns the probability for a given cloth-skin displacement but truncated at α_{pdf_c}. We also define a function when d_{s} is smaller than α_{pdf_c} using the Euclidean distance weighted by λ_l to penalize the SMPL contour points outside the input silhouette. In the cost function, $d_{\mathbf{p}}$ forms the negative log-likelihood. Therefore E_{c} serves to change the SMPL parameters so that the contour displacement fits the pre-constructed displacement model.

3 Experiments

We performed qualitative and quantitative experiments to unveil the effect of the proposed approach.

3.1 Quantitative evaluation

3.1.1 Experimental settings

For evaluation, we used a dataset consisting of time-series 3D textured human shape acquired by 3D scanners [4], which includes the ground truth shape of the unclothed persons. We utilized 3D videos of four subjects in motion, where each subject wears two clothing variations: (1) T-shirt/long pants and (2) soccer outfit[2]. We generated input images by projecting selected frames from the dataset to the virtual camera of 860×860 resolutions, located at the front of the persons. From each 3D video sequence, we sampled five frames for single-image input. While the dataset provides detailed 3D shapes of human, the proposed approach use SMPL models. As the ground truth 3D models for this experiment, we, therefore, generated SMPL models fitted to the provided shape by minimizing distances between the 3D surface of two models.

We compared the following two approaches:

1 Optimization using joint positions [5] (SMPLify).
2 Optimization using joint, silhouette contours, and cloth-skin displacement model (proposed).

To evaluate shape accuracy, we translated the estimated SMPL human model using the unit pose, which is the same pose as the ground truth shape provided in the dataset. We measured the accuracy as the average error of the ten body measurements (shown in Fig. 3) in the estimated and the ground truth 3D models. For evaluation, we unified the overall height for each model because the two approaches do not estimate scale information.

3.1.2 Results

Table 1 shows the relative error of shape estimation by each approach. Our approach yielded better accuracy than that of SMPLify. Not only non-skeletal lengths (e.g., chest circumference), we found a few measurements related to human skeletons such as shoulder length are also estimated with higher accuracy than SMPLify.

3.2 Visual comparison

While results in the previous section demonstrate the improvements by the proposed approach, the dataset we employed for the quantitative evaluation [4] does not

Table 1 The relative error [%] for BUFF dataset [4]

Method	Clothing	A	B	C	D	E	F	G	H	I	J	Average
SMPLify	T-shirt, long pants	11.99	8.89	9.43	9.50	13.58	15.51	2.54	12.14	3.26	7.61	9.44
	Soccer outfit	13.46	10.92	7.23	11.15	17.32	18.80	3.18	13.88	3.42	8.51	10.79
	Average	12.72	9.90	8.33	10.32	15.45	17.15	2.86	13.01	3.34	8.06	10.12
Proposed	T-shirt, long pants	6.08	2.01	7.51	4.32	3.58	4.01	1.47	9.18	2.99	7.46	4.86
	Soccer outfit	6.49	3.89	5.25	4.66	5.51	10.01	3.12	9.34	3.90	8.47	6.06
	Average	6.29	2.95	6.38	4.49	4.54	7.01	2.30	9.26	3.45	7.97	5.46

For each approach, we averaged the error in five frames sampled from the 3D video sequences

Fig. 4 Visual comparison. From left to right: an input image, a result by SMPLify, and proposed approaches

include loose clothing and a variety of body shapes. One important advantage of the proposed approach is the adaptability to a variety of clothing types and body shapes. We, therefore, describe a visual comparison using a variety of photographs collected from fashion photographs.

Figure 4 shows one result of the two approaches described in Section 3.1. In comparison between the joint-based approach (SMPLify) and the proposed approach, joint-based optimization does not produce a body shape that represents musculature and fat distribution.

4 Conclusions

This paper has described the first approach that estimates human 3D pose and shape, including non-skeletal information from a single RGB image. We model the displacement between clothed and naked contours for each clothing category, using an artificial dataset created by an auto-encoder-based image generation method. The proposed approach optimizes a SMPL parametric human model through a likelihood-based cost function, using a cloth-skin displacement model, silhouette shape, and joint locations. Through the experiments, the proposed approach more accurately estimated shape coefficients as compared with the joint-based approach [5]. Extension of the proposed approach to unsynchronized multi-view input is an interesting and viable research direction.

Endnotes

[1] Because contours must belong to foreground regions, $\mathcal{C}$ does not include the "background" label.

[2] The original version of [4] includes five subjects in public, while footage of one subject wearing only a single clothing combination was unused in this experiment.

Acknowledgements
This work was partly supported by a cooperative research with Daikin Industries, Ltd.

Funding
The research being report in this publication was supported by Daikin Industries, Ltd., as a cooperative research.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. Note the datasets analyzed in the experiments should be obtained from the original authors.

Authors' contributions
YS played the key role in the implementation, experiments, and paper editing. FO conducted the algorithm design and mainly wrote and edited the paper. IM supported the experiments and played an important role in editing the paper. YY played an important role in the research design. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1] The Institute of Scientific and Industrial Research, Osaka University, 8-1, Mihogaoka, Ibaraki, Osaka, Japan. [2] Graduate School of Information Sciences, Hiroshima City University, 3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima, Hiroshima, Japan.

Received: 16 October 2018 Accepted: 23 November 2018
Published online: 27 December 2018

References

1. Yuan M, Khan IR, Farbiz F, Yao S, Niswar A, Foo MH (2013) A mixed reality virtual clothes try-on system. IEEE Trans Multimed 15(8):1958–1968
2. Balan A, Black MJ (2008) The naked truth: estimating body shape under clothing. In: Proc. European Conf. on Computer Vision (ECCV'08). Springer-Verlag Berlin Heidelberg, Marseille. pp 15–29
3. Song D, Tong R, Chang J, Yang X, Tang M, Zhang JJ (2016) 3D body shapes estimation from dressed-human silhouettes. Comput Graph Forum 35(7):147–156
4. Zhang C, Pujades S, Black MJ, Pons-Moll G (2017) Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17). IEEE, Honolulu. pp 4191–4200
5. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ (2016) Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Proc. European Conf. on Computer Vision (ECCV'16). Springer, Amsterdam. pp 561–578

6. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: a skinned multi-person linear model. ACM Trans Graph 34(6):248
7. Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B (2016) DeepCut: joint subset partition and labeling for multi person pose estimation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16). IEEE, Las Vegas. pp 4929–4937
8. Lin G, Milan A, Shen C, Reid I (2017) RefineNet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17). IEEE, Honolulu. pp 1925–1934
9. Yang W, Luo P, Lin L (2014) Clothing co-parsing by joint image segmentation and labeling. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14). IEEE, Columbus. pp 3182–3189
10. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2012) Parsing clothing in fashion photographs. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'12). IEEE, Providence. pp 3570–3577
11. Lassner C, Pons-Moll G, Gehler PV (2017) A generative model of people in clothing. In: Proc. IEEE Int'l Conf. on Computer Vision (ICCV'17). IEEE, Venice. pp 853–862
12. Cao Z, Simon T, Wei S, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17). IEEE, Honolulu. pp 7291–7299
13. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. In: Proc. European Conf. on Computer Vision (ECCV'14). Springer International Publishing, Zurich. pp 740–755