Junyao Sun\textsuperscript{1} and Qiong Liu\textsuperscript{1}

\textsuperscript{1}Affiliation not available

April 01, 2024
Abstract—The inconsistency between training strategies and task objectives is a principal constraint impeding DensePose development. DensePose aims to minimize the surface distance between the estimations and the ground truth, thereby establishing a correspondence between two-dimensional images and three-dimensional humans. However, the discrete IUV optimization strategy employed by the existing pipeline deviates from the goal of minimizing surface distance, thus yielding sub-optimal results. To solve this problem, we propose Geodesic-consistent RCNN (GC RCNN), which models the intrinsic interdependence within the IUV to enhance human surface understanding and facilitate surface distance optimization. GC RCNN incorporates the adaptively semantic enhancement module (ASEM) and the geodesic-consistent loss (GCL) into the traditional DensePose estimation framework. ASME enhances human surface semantics by extracting distinct IUV features and dynamically modeling the feature-level IUV interdependence. GCL extracts the surface distance information by quantifying the degree of deviation between the estimated IUV distribution and the ground truth. The distribution deviation is utilized to refine the IUV optimization process, thereby aiding in minimizing surface distance. Empirical analyses conducted on the DensePose-COCO dataset validate the superior performance of GC RCNN, which surpasses the base model by a margin of up to 3.3% AP.

Index Terms—Dense Human Pose Estimation, Feature Enhancement, Loss Function.

I. INTRODUCTION

DENSEPOSE [1] is a fundamental task for human understanding, which is a cornerstone for advancements in intelligent transportation systems [2]–[4]. It facilitates many three-dimensional visual applications, including but not limited to texture/pose transfer [5], [6], 3D human reconstruction [7], and human image animation [8]. DensePose effectively bridges the dimensional divide by establishing the correspondences between two-dimensional image pixels and three-dimensional vertices on the SMPL human mesh [9]. Through implementing various enhancement modules, such as multi-scale feature fusion [10], [11], knowledge transfer [12], data augmentation [13], [14], re-balanced strategies [15], and quality estimation [16], DensePose has witnessed significant strides. Nonetheless, constrained by the discretized IUV optimization strategy, DensePose still suffers from inconsistency between training strategies and task objectives.

DensePose aims to establish the correspondence between two-dimensional images and three-dimensional humans. Thus, minimizing the geodesic distance between the estimation and ground truth is used as the core objective to precisely map 2D human pixels onto the 3D human surface. However, the current pipeline decouples the task of minimizing geodesic distance into independent optimizations for I, U, and V, where the ‘I’ identifies the specific surface category the pixel belongs to, and U and V identifies the position on that surface category’s unfolded map. The discrete IUV optimization strategy may deviate from the core objective of minimizing geodesic distance due to the substantial dependency of UV on I. Figure 1 intuitively explains this phenomenon with a visualization case. With the ground truth on surface 23 (23,0.81,0.6) and an initial position (24,0.2,0.6), the discrete IUV optimization strategy will always prompt the UV towards the position (0.81,0.6). When ‘I’ fails to achieve optimality, such as being misclassified as surface 24 due to appearance similarities, the optimization strategy directed towards (0.81,0.6) inadvertently increases the geodesic distance between the prediction and the ground truth.

To mitigate the deviation in optimization, DensePose RCNN [1] adopts a strategy that models the UV distribution with multiple distinct single-modal models, thereby strengthening the coherence within IUV estimates. It trains multiple UV regression functions, each dedicated to regressing UV coordinates pertinent to its designated ‘I’. During each training iteration, only the UV estimation yielded by the regression function corresponding to the ground-truth surface category is considered for the computation of loss. However, during inference, the UV coordinates yielded by the regression function corresponding to the predicted surface category are utilized to establish the correspondence. Finally, it results in a phenomenon where the predicted UV coordinates erroneously fit with an inappropriate regression function due to surface misclassification.

Currently, there are two solutions to the inconsistency problem: 1) Improving the surface classification by introducing human part information. KTN [12] extracted human part
knowledge from the keypoints classifier and integrated it into the surface classifiers via a predefined relational knowledge graph. However, the part/pose knowledge only delineates the sparse spatial positioning of human components rather than providing comprehensive information about the human surface. We have summarized the misclassifications for each surface category, as shown in Figure 2. Misclassifications frequently arise among surface categories corresponding to the same human part, such as the left upper arm (surface 15 and 17) and the right upper arm (surface 16 and 18). Therefore, the key to enhancing surface classification efficiency is leveraging human surface information rather than relying solely on keypoints data. 2) Designing a geodesic distance-based optimization strategy. Since the mapping from IUV to vertices is a non-differentiable query operation, Human GPS [17] and CSE [18] opt to replace the IUV with the per-pixel embeddings to represent the corresponding vertex. These embeddings facilitate the direct utilization of geodesic distances between corresponding 3D vertices as the loss function. However, this approach necessitates a predefined human embedding model, and perhaps owing to the heightened task complexity, the embedding setting fails to yield significant performance improvements within the existing model architecture.

These challenges motivate us to study two problems: (1) how to design a straightforward yet effective module to enhance human surface semantics information, and (2) how to equip the IUV optimization strategy to minimize surface distance. Based on the analysis above, we think the individual I, U, and V alone inadequately capture the holistic human surface information. Therefore, we proposed Geodesic-consistent RCNN (GC RCNN), which models the intrinsic interdependence within the IUV to enhance human surface understanding and facilitate surface distance optimization.

GC RCNN incorporates the adaptively semantic enhancement module (ASEM) and the geodesic-consistent loss (GCL) into the traditional DensePose estimation framework. ASEM enhances human surface semantics by extracting the distinct I, U, and V features and dynamically modeling the feature-level IUV interdependence among these features. The current pipeline implicitly learns the human surface information by sharing IUV features, which will deliver sub-optimal results because of the misalignment between classification and localization task domains [19]. Intuitively, the surface classification feature learns the surface category semantics of each pixel and has regional similarities. In contrast, the UV regression feature learns the local surface position semantic restricted to the corresponding surface partition and promotes regional coherence to engender seamless UV coordinates. ASEM proposed the separation semantic head to explicitly extract surface classification and UV regression features with two individual branches for learning surface category semantics and local surface position semantics, respectively. Then, the human surface semantic can be represented by combining all these distinct features. In particular, ASEM introduces the modulation module to adaptively adjust the relative contribution among the shared feature, surface classification feature, and UV regression feature based on the inputs and outputs, dynamically modeling the feature-level IUV interdependence. All branches are designed to be lightweight and minimize additional computational demands.

The geodesic-consistent loss (GCL) is a differentiable loss function that approximates the surface distance by quantifying the degree of deviation between the estimated IUV distribution and the ground truth. As shown in Figure 1, we observed the semantic inconsistency across UV coordinate systems for different surfaces. When the left head is inaccurately classified, the UV estimation will fit the non-ground truth surfaces' regression function, resulting in biased UV coordinates. Take, for instance, the example at the 2D coordinates of the left eye (marked by a green dot at (0.81,0.6)), where the regression model for the right head incorrectly assigns UV coordinates corresponding to the right eye (indicated by a blue dot at (0.2,0.6)). Meanwhile, other surface regression functions will produce disorganized UV values due to insufficient eye understanding. The overall UV deviation and uncertainty subtly represent the surface distance between the estimation and ground truth. Drawing on this insight, GCL measures the degree of deviation between the estimated IUV distribution and ground truth to refine IUV optimization, thereby aiding in minimizing surface distance. Specifically, we model the UV estimation with a multi-modal Gaussian distribution, in which the variance represents the uncertainty of the UV estimation. GCL computes the KL divergence between the estimated IUV distribution and the ground truth as the IUV distribution deviation to approximate surface distances. It then prompts the classifier to suppress responses to the wrong surface categories based on the IUV distribution deviation and minimize the UV distribution deviation to advance the precision of the UV regression functions.

We evaluate our GC RCNN on the DensePose-COCO dataset. Compared with DensePose RCNN, we achieve up to 3.3% AP and 1.5% AP improvement with ResNet-50 and ResNet-101-FPN-DL as the backbone. In brief, the contributions of the paper are:

1) Propose Geodesic-consistent RCNN (GC RCNN) to alleviate the inconsistency in DensePose, which introduces the adaptively semantic enhancement module (ASEM)
2) ASME extracts distinct IUV features and models the feature-level IUV interdependence dynamically to enhance human surface semantics. GCL assists in surface distance optimization based on the IUV distribution deviation.

3) Extensive experiments demonstrate the effectiveness of our model. GC RCNN significantly outperforms the base models by up to 3.3% AP on the DensePose-COCO dataset.

II. RELATED WORK

A. DensePose

Motivated by Mask RCNN [20], current state-of-the-art methods of DensePose adopted the two-stage pipeline. DensePose RCNN [1] replaced the mask head with a densepose head and modeled the UV distribution with multiple single-modal models to fit the discretized surface partition. However, it results in the predicted UV fitting the wrong regression function once a surface is misclassified. Several outstanding works solve this problem by introducing enhancements to improve surface classification, such as multi-scale feature fusion [10], [11], knowledge transfer [12], and re-balanced strategies [15]. Multi-scale feature fusion is an effective way to obtain instance details. Parsing R-CNN [10] proposed the proposals separation sampling, which performed RoIPool only on the finest level to keep the details of instances. AMA-Net [11] introduced the adaptive multi-path aggregation algorithm to perceive multi-scale visual information. Knowledge transfer can boost performance with limited supervision. AMA-Net [11] proposed a task transformer to benefit the performance of the densepose estimator by leveraging 2D human parsers. KTN [12] introduced a structural body knowledge graph to transform 2D human parsers trained from sufficient annotations to 3D human surface parsers. SimPose [21] mitigated the ”inter-domain Covariate Shift” between the 2.5D DensePose estimation task and the 3D human surface normal estimation task. Data imbalance is one of the major factors limiting the algorithm performance of DensePose. PoiseNet [15] proposed the adaptive equalization loss and block-based landmark localization to re-balance inter-class and intra-class learning. Despite the advantages offered by current technology, achieving optimal surface classification remains a formidable challenge. We summarized the misclassifications for each surface category and found that misclassifications frequently arise among surface categories corresponding to the same human part. Therefore, leveraging human surface information is critical to improve surface classification performance. Given these insights, we proposed ASME to enhance human surface semantics by extracting the explicit I, U, and V features and dynamically modeling the feature-level IUV interdependence.

B. Geodesic distance-based optimization

Recently, many works have solved the above problem by directly minimizing the geodesic distances between corresponding 3D vertices, such as CSE [18], Human GPS [17], and BodyMap [22]. Since the mapping from IUV to vertices is a non-differentiable query operation, CSE [18] opt to replace the IUV with the per-pixel embeddings to represent the corresponding vertex. These embeddings facilitate the direct utilization of geodesic distances between corresponding 3D vertices as the loss function. [22] then extended CSE to other non-rigid objects, such as animals. Human GPS [17] also maps each pixel to an embedding, where the embedding distances reflect the geodesic distances. It designs four geodesic losses to push features apart according to their geodesic distances on the surface. BodyMap [22] takes the CSE prediction as input and builds a transformer-based architecture to provide a finer correspondence. However, all these methods necessitates a predefined human embedding model, and perhaps owing to the heightened task complexity and reduced interpretability, the embedding setting fails to yield significant performance improvements within the existing model architecture. It motivates us to study the problem of how to equip the IUV optimization strategy to minimize surface distance. We finally proposed the geodesic-consistent loss to facilitate surface distance optimization based on the IUV representation.

III. GEODESIC-CONSISTENT RCNN

Geodesic-consistent RCNN (GC RCNN) enhances human surface understanding and facilitates surface distance optimization by modeling the intrinsic interdependence within the IUV components. It achieves this through a dual approach: enhancing features and guiding the learning process via the Adaptive Semantic Enhancement Module (ASEM) and the Geodesic-Consistent Loss (GCL), respectively. Figure 3 shows the overall pipeline of GC RCNN. In this section, we begin with a brief overview of the DensePose and then introduce the ASEM and GCL in detail.

A. DensePose RCNN

DensePose RCNN [1] adopted the ResNet and FPN as the backbone model and applied RPN to generate proposals. RoIAlign is applied to extract RoI features. Then, eight convolution layers are stacked to extract the instance-level densepose features. Four parallel deconvolution and interpolation operations upsample these features to generate a mask for instance segmentation \( M_S \in \mathbb{R}^{B \times 2 \times H' \times W'} \), a mask for surface segmentation \( M_I \in \mathbb{R}^{B \times 25 \times H' \times W'} \), a U coordinate...
Fig. 3. The architecture of Geodesic-consistent RCNN. GC RCNN integrates the adaptively semantic enhancement module (ASEM) and the geodesic-consistent loss (GCL) to enhance interdependence within the IUV components.

Fig. 4. The architecture of the adaptively semantic enhancement module. It proposes the separation semantic head and the modulation module to enhance human surface semantics.

map $M_U \in \mathbb{R}^{B \times 25 \times H' \times W'}$, and a V coordinate map $M_V \in \mathbb{R}^{B \times 25 \times H' \times W'}$, respectively. During training, cross-entropy loss is used for learning $M_S$ and $M_I$ and SmoothL1 loss is used for $M_U$ and $M_V$. During inference, the IUV maps can be extracted according to Algorithm 1. Firstly, the instance mask and surface mask are extracted by the argmax on the $M_S$ and $M_I$, and the UV maps are extracted based on the surface mask. Finally, we remove the background to obtain the final IUV image $\hat{S}$.

B. Adaptively semantic enhancement module

The adaptively semantic enhancement module (ASEM) enhances human surface semantics by extracting distinct IUV features and dynamically modeling the feature-level IUV interdependence. The current pipeline uses the shared features for IUV estimation, which will deliver sub-optimal results because of the misalignment between classification and localization task domains [19]. To solve the misalignment in the shared feature, ASEM proposed the separation semantic head to explicitly extract surface classification and UV regression features with two individual branches for learning surface category semantics and local surface position semantics, respectively. Then, the modulation module is proposed to adaptively adjust the relative contribution among the shared feature, surface classification feature, and UV regression feature based on the inputs and outputs, dynamically modeling the feature-level IUV interdependence, dynamically modeling the feature-level IUV interdependence. The architecture of ASEM is shown in Figure 4.

Separation semantic head. The separation semantic head extracts three critical semantic features: the shared feature, the surface category semantics feature, and the local surface position semantics feature, employing three dedicated branches for this purpose, respectively. Our separation semantic head is simple: as shown in Figure 4, the shared feature is extracted
by a sharing branch similar to the DensePose RCNN [1]. Then, the surface category and local surface position semantics features are extracted via two distinct 3 × 3 convolution layers, each succeeding the shared feature. The surface category semantics feature is used to estimate the surface category, and the local surface position semantics feature estimates the multi-modal UV distribution. Finally, all these features are fed into the modulation module, where their respective contributions to IUV estimation are dynamically adjusted.

**Modulation module.** The modulation module is proposed to integrate the surface information in the features from the separation semantic head for IUV estimation. The insight is to imbue the surface classification and UV regression branches with relevant surface information by aligning the shared feature with the specific requirements of each task. Motivated by the dynamic convolution [23], we view the surface category semantics and the local surface position semantics features as guidance for extracting aligned surface semantics features from the shared feature. Specifically, the surface category semantics feature and the local surface position semantics feature are fed to two parallel squeeze-and-excitation blocks to generate the category semantics perception filter and position semantics perception filter for aligning the sharing features. The two filters adaptively align the semantic intensity of the output according to the input features. Finally, the aligned surface category semantics feature is added to the surface category semantics feature to facilitate pixel-wise classification to the appropriate surface category. Similarly, the aligned surface position semantics feature is combined with the local surface position semantics feature to facilitate the estimation of the UV distribution. The pipeline of the modulation module is given in Figure 4.

C. Geodesic-consistent loss

Geodesic-consistent loss (GCL) is a differentiable loss function that approximates the surface distance by quantifying the degree of deviation between the estimated IUV distribution and the ground truth. The existing framework employs an independent optimization strategy for I, U, and V components. For example, the classification unit is optimized through a cross-entropy loss function, while the 24-way U or V regression units are optimized via the SmoothL1 loss function. Notably, only the UV estimation yielded by the regression unit corresponding to the ground truth I is considered for the computation of loss. As mentioned, the current independent loss setup may lead to the UV estimation fitting the wrong regression unit in misclassification cases. Therefore, the geodesic-consistent loss is a holistic loss that considers all IUV estimations, aiming to optimize surface distance directly. The insight is calculating the degree of deviation between the estimated IUV distribution and the ground truth to facilitate IUV optimization.

**Measurement of IUV distribution deviation.** Our observations revealed the semantic inconsistency across UV coordinate systems for different surfaces. When the surface is inaccurately classified, the UV estimation will fit the non-ground truth surfaces’ regression function, resulting in biased UV coordinates. Specifically, when the surface category I is misclassified to a category within the same human part, the U estimation tends to exhibit mirror symmetry with the ground truth, given the semantic similarities between the two surface categories in the 2D image (e.g., left head and right head). Moreover, when I is misclassified to the other surface category, the UV estimation often exhibits high uncertainty due to the difficulty of their regression units to respond to the ground-truth surface category. Based on this observation, the geodesic-consistent loss uses the overall UV deviation and uncertainty among all surfaces to approximate the surface distance between the estimation and ground truth.

We use the Kullback-Leibler Divergence between the estimated UV distribution and the ground-truth UV distribution to measure the distribution deviation for each surface, as shown in Equation (1)

\[
\mathbb{D}_\text{KL}(P||Q) = \int_{-\infty}^{+\infty} p(x) \log \left( \frac{p(x)}{q(x)} \right) dx
\]

where the \(q(x)\) is the estimated UV distribution, and the \(p(x)\) is the ground-truth UV distribution. Motivated by [24], we assume that the UV distribution follows the Gaussian distribution \(\mathcal{N}(\mu, \sigma)\). Therefore, the estimated UV distribution \(Q \sim \mathcal{N}(\mu, \sigma)\), where the \(\mu\) is the estimated UV coordinates and the \(\sigma\) is the estimated uncertainty. The ground-truth UV distribution \(P \sim \mathcal{N}(\bar{\mu}, \hat{\sigma})\), where \(\bar{\mu}\) is the ground-truth UV coordinates and \(\hat{\sigma}\) is a predefined hyperparameter describing the uncertainty of ground truth. Then, the KL Divergence between the estimated UV distribution and the ground-truth UV distribution is:

\[
\mathbb{D}_\text{KL}(P||Q) = \int_{-\infty}^{+\infty} p(x) \log \left( \frac{p(x)}{q(x)} \right) dx
\]

\[
= \frac{1}{2} \log(2\pi\sigma^2) + \frac{\hat{\sigma}^2 + (\mu - \bar{\mu})^2}{2\sigma^2} - \frac{1}{2}(1 + \log 2\pi\hat{\sigma}^2)
\]

\[
= \log \frac{\sigma}{\hat{\sigma}} + \frac{\hat{\sigma}^2 + (\mu - \bar{\mu})^2}{2\sigma^2} - \frac{1}{2}
\]

(2)

**Surface classification.** For each surface category \(i\), we have a \(\mathbb{D}_\text{KL}(P||Q)\) to measure the distribution deviation between its corresponding UV distribution and the ground-truth UV distribution. A large KL divergence suggests a significant disparity between the corresponding surface category and the ground truth, indicating a probable misclassification. Conversely, a smaller KL divergence implies a closer resemblance to the ground truth. We introduce a weighting mechanism for the classification loss based on the KL divergence to aid in optimizing surface classification.

The original classification loss is a cross-entropy loss and is formulated as follows:

\[
L_I = -\sum_{i=1}^{C} y_i \log(p_i).
\]

Where \(C\) is the number of surface categories. The multinomial distribution \(p\) is calculated by re-weighting \(\text{Softmax}(z)\) from the network outputs \(z\). The ground-truth distribution \(y\)
uses one-hot representation and has \( \sum_{j=1}^{C} y_j = 1 \). Formally, for the ground-truth surface \( c \) of a sample,
\[
y_j = \begin{cases} 
1, & \text{if } j = c; \\
0, & \text{otherwise}.
\end{cases}
\]  

(4)

The derivative of the \( L_I \) with respect to network’s output \( z \) is formulated as follows:
\[
\frac{\partial L_I}{\partial z_i} = p_i - y_i.
\]  

(5)

For a surface \( i \) that is not the ground truth \( (y_i \neq 1) \), \( L_I \) generates negative gradients to force the classifier for the surface \( i \) to output low classification scores \( (p_i = \text{Softmax}(z_i)) \). Based on our analysis, for a surface \( i \) with a larger KL divergence, \( L_I \) should generate more negative gradients to output lower classification scores. Therefore, the classification GCL can be formulated as follows:

\[
L_{GCL-I} = -\sum_{i=1}^{C} y_i \log(p_i).
\]  

(6)

\[
\tilde{p}_i = \frac{e^{w_i z_i}}{\sum_{k=1}^{C} w_k e^{z_k}}.
\]

Where the \( w_k \) is the KL Divergence between the estimated UV distribution and the ground-truth UV distribution:

\[
w_k = \mathbb{D}_{KL}(P||Q) = \log \frac{\sigma_k}{\sigma} + \frac{\sigma^2 + (\mu_k - \hat{\mu})^2}{2\sigma_k^2} - \frac{1}{2}
\]  

(7)

Finally, the gradient of the loss function with respect to \( z_i \) can be formulated as Equation (8).

\[
\frac{\partial L_{GCL-I}}{\partial z_i} = w_i(\tilde{p}_i - y_i)
\]  

(8)

UV regression. Rather than model the UV distribution as the multiple individual models, we opt to represent it as a multi-modal Gaussian model. This approach allows for the jointly optimization of UV coordinates and uncertainty \( \sigma \). The \( \sigma \) is generated through the incorporation of an additional deconvolution layer following the UV branch.

MDN \cite{25} dynamically partitions the target space to prevent convergence towards an average target. Specifically, it predicts the probability density of UV coordinates conditioned on the input image, as shown in Equation (9).

\[
p(\mu|x) = \sum_{i=1}^{C} \alpha_i \phi_i(\mu|x).
\]  

(9)

Where \( C \) is the number of components. The mixing coefficient \( \alpha_i \) indicates the probability that the component \( i \) responds to the input. The probability density function \( \phi_i(\mu|x) \) is the estimated Gaussian distribution for component \( i \):

\[
\phi_i(\mu|x) = \frac{1}{2\pi\sigma_i} e^{-\frac{||\mu_i - \hat{\mu}||^2}{2\sigma^2_i}}
\]

(10)

Finally, the mixture density loss is the negative log-likelihood of the \( p(\mu|x) \), as shown in Equation (11). It jointly optimizes the \( \mu_i, \sigma_i, \) and \( \alpha_i \) with the \( L_{MDN} \).

\[
L_{MDN} = -\ln \sum_{i=1}^{C} \alpha_i \phi_i(\mu|x)
\]

\[
= -\ln \sum_{i=1}^{C} \alpha_i \left( \frac{1}{2\pi\sigma_i} e^{-\frac{||\mu_i - \hat{\mu}||^2}{2\sigma^2_i}} \right)
\]

(11)

Our method can be seen as a specialized case of MDN. Specifically, we predefined the components as the human surfaces and replaced the \( \alpha_i \) with the surface classification probability \( p_i \). We also use the Gaussian distribution as the probability density function for these components. However, instead of optimizing all parameters with a single loss function, we design the classification GCL to optimize the surface classification probability \( p_i \) and the regression GCL to diminish the UV distribution deviation of the ground truth surface while enlarging the UV distribution deviation of other surfaces.

Therefore, the regression GCL can be formulated as Equation (12).

\[
L_{GCL-UV} = -\ln \sum_{i=1}^{C} p_i \phi_i(\mu)
\]

\[
= -\ln \sum_{i=1}^{C} p_i \left( \frac{1}{2\pi\sigma_i} e^{-\frac{||\mu_i - \hat{\mu}||^2}{2\sigma^2_i}} \right)
\]

(12)

Finally, we replace the classification loss with our geodesic-consistent loss \( L_{GCL-I} \) and add the \( L_{GCL-UV} \) to optimize the UV distribution. The total loss is written as:

\[
L_{total} = \lambda_1 * L_{GCL-I} + \lambda_2 * L_{GCL-UV} + \lambda_3 * L_{UV}.
\]  

(13)

Where \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are balanced hyperparameters.

IV. EXPERIMENTS

A. Datasets

DensePose-COCO dataset. Constrained by a complex annotation procedure, DensePose-COCO stands as the sole large-scale outdoor DensePose dataset presently available, serving as a solid benchmark for the advancement of DensePose. The DensePose-COCO dataset gathers annotations for dense correspondences between 2D human images and 3D human surface. It delineates human parts to be isomorphic to a plane and partitions the limbs and torso into frontal-back parts, totaling 24 parts. The DensePose-COCO dataset has collected annotations for 50K humans, providing approximately 5M manually annotated correspondences.

B. Experimental settings

1) Evaluation Metrics: For DensePose evaluation, we report the metric of the average precision (AP) and recall (AR) based on the geodesic point similarity (GPS) and masked geodesic point similarity GPSm. The GPS is formulated by
Equation (14), where $P$ is the set of annotated points on each instance, $i_p$ is the ground truth 3D human vertex at point $p$, and $\hat{i}_p$ is the predicted 3D human vertex at point $p$. The GPSm is calculated by $GPSm = \sqrt{GPS \cdot I}$, where $I$ is the IoU between the ground truth and the predicted instance mask. $\kappa$ is a normalizing parameter and is used to make $GP = 0.5$ if the geodesic distance $g(i_p, \hat{i}_p)$ is the half-size of the body part. Therefore, we use different $\kappa$ for different human parts. The AP and AR are reported with the GPS and GPSm ranging from 0 to 0.95, corresponding to the range of geodesic distances between 0 and half-size of the body segment.

$$GPS = \frac{P}{\kappa} \sum_{p \in P} exp\left(-\frac{g(i_p, \hat{i}_p)}{2\kappa^2}\right).$$  \hspace{1cm} (14)

2) Implementation details.: Our models are built on an Ubuntu server with 8 GeForce GTX 1080 Ti GPUs. For the training of GC RCNN, the batch size is set as 2 per GPU, and the initial learning rate is 0.01. It is trained for 130K iterations. The learning rate is decreased to $1e^{-3}$ and $1e^{-4}$ at the 100K-th and 120k-th iteration. All models are optimized by stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.0001. For the hyperparameters in Equation (13), we set $\lambda_1 = 5$, $\lambda_2 = 0.001$, and $\lambda_3 = 0.01$.

C. Comparison with State-of-the-art Approaches

Table I shows the comparison between our GC RCNN and the SOTA methods on the DensePose-COCO minival-2014 dataset. GC RCNN outperforms the DensePose RCNN by 3.3% and 1.5% with ResNet-50 and ResNet-101-FPN-DL as the backbone, respectively. Compared with PoseNet [15], which employs a balanced training strategy, GC RCNN achieves 64.7%, suppressing PoseNet by 1.9% AP gains with our interdependent IUV training strategy. QANet [27] combined the bound box score, mIoU score, and pixel score to rank the proposal and achieved 63.7% AP, while GC RCNN achieved 64.7% AP with the original scoring criterion. AMA-Net [11] adaptively fused multi-level features, achieving 64.1% AP. GC RCNN outperforms AMA-Net by 0.7% AP with the original FPN setting. KTN [12] transformed the human structure knowledge from keypoints to benefit the DensePose estimation but was inferior to GC RCNN by 0.5% AP. DS RCNN proposed to estimate the GPS score to rank the proposals and achieve 66.2% AP. GC RCNN outperforms DS RCNN by 0.5% AP with the original classification score. These results demonstrate the effectiveness of GC RCNN in learning human surface information by modeling IUV interdependence to benefit DensePose. Furthermore, GC RCNN suppressed DensePose RCNN by 3.3% AP, 1.3% $AP_{50}$, and 3.4% $AP_{75}$, respectively. More improvement in $AP_{75}$ indicates that GC RCNN yields IUV results with higher-quality

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>$AP_{gpsm}$</th>
<th>AR</th>
<th>$AR_{50}$</th>
<th>$AR_{75}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP RCNN(R50)*</td>
<td>61.4</td>
<td>90.1</td>
<td>69.8</td>
<td>63.7</td>
<td>69.6</td>
<td>93.6</td>
<td>77.4</td>
</tr>
<tr>
<td>DP RCNN(R101-FPN-DL)*</td>
<td>65.2</td>
<td>92.0</td>
<td>73.6</td>
<td>66.7</td>
<td>72.9</td>
<td>95.0</td>
<td>80.6</td>
</tr>
<tr>
<td>Parsing R-CNN [10]</td>
<td>55.0</td>
<td>87.6</td>
<td>59.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SimPose [21]</td>
<td>57.2</td>
<td>88.4</td>
<td>67.3</td>
<td>-</td>
<td>66.4</td>
<td>95.1</td>
<td>77.8</td>
</tr>
<tr>
<td>Slim DensePose [13]</td>
<td>57.3</td>
<td>88.4</td>
<td>63.9</td>
<td>-</td>
<td>65.8</td>
<td>92.6</td>
<td>73.0</td>
</tr>
<tr>
<td>RP RCNN [26]</td>
<td>58.1</td>
<td>88.3</td>
<td>64.6</td>
<td>58.1</td>
<td>72.2</td>
<td>95.1</td>
<td>81.1</td>
</tr>
<tr>
<td>CSE(R50) [18]*</td>
<td>62.3</td>
<td>90.3</td>
<td>69.8</td>
<td>63.9</td>
<td>70.2</td>
<td>93.6</td>
<td>77.1</td>
</tr>
<tr>
<td>PoiseNet(R50) [15]</td>
<td>62.8</td>
<td>91.0</td>
<td>71.3</td>
<td>64.5</td>
<td>70.5</td>
<td>94.1</td>
<td>78.1</td>
</tr>
<tr>
<td>QANet [27]</td>
<td>63.7</td>
<td>92.7</td>
<td>73.4</td>
<td>63.7</td>
<td>69.5</td>
<td>94.8</td>
<td>79.0</td>
</tr>
<tr>
<td>AMA-Net [11]</td>
<td>64.1</td>
<td>91.4</td>
<td>72.9</td>
<td>-</td>
<td>71.6</td>
<td>94.7</td>
<td>79.8</td>
</tr>
<tr>
<td>UV RCNN [28]</td>
<td>65.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CSE(R101-FPN-DL) [18]*</td>
<td>65.6</td>
<td>91.9</td>
<td>73.2</td>
<td>67.0</td>
<td>73.7</td>
<td>95.2</td>
<td>80.9</td>
</tr>
<tr>
<td>KTN [12]*</td>
<td>66.2</td>
<td>91.9</td>
<td>75.0</td>
<td>65.9</td>
<td>74.3</td>
<td>95.8</td>
<td>82.4</td>
</tr>
<tr>
<td>Confidence Estimation [24]</td>
<td>66.2</td>
<td>92.2</td>
<td>73.4</td>
<td>67.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS RCNN [16]</td>
<td>66.2</td>
<td>91.9</td>
<td>74.7</td>
<td>67.0</td>
<td>72.3</td>
<td>94.9</td>
<td>80.2</td>
</tr>
<tr>
<td>GC RCNN(R50)</td>
<td>64.7</td>
<td>91.4</td>
<td>73.2</td>
<td>65.8</td>
<td>72.1</td>
<td>95.6</td>
<td>79.8</td>
</tr>
<tr>
<td>GC RCNN(R101-FPN-DL)</td>
<td>66.7</td>
<td>92.3</td>
<td>75.1</td>
<td>67.1</td>
<td>74.7</td>
<td>96.0</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Table II Ablation study of the individual component of GC RCNN.

<table>
<thead>
<tr>
<th>ASME</th>
<th>GCL</th>
<th>AP</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>$AP_{gpsm}$</th>
<th>AR</th>
<th>$AR_{50}$</th>
<th>$AR_{75}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td></td>
<td>61.4</td>
<td>90.1</td>
<td>69.8</td>
<td>63.7</td>
<td>69.6</td>
<td>93.6</td>
<td>77.4</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>63.7</td>
<td>91.6</td>
<td>72.2</td>
<td>65.1</td>
<td>70.8</td>
<td>94.6</td>
<td>78.5</td>
</tr>
<tr>
<td>✓ ✓</td>
<td></td>
<td>62.8</td>
<td>90.6</td>
<td>71.4</td>
<td>63.9</td>
<td>70.5</td>
<td>94.3</td>
<td>78.5</td>
</tr>
<tr>
<td>✓ ✓</td>
<td></td>
<td>64.7</td>
<td>91.4</td>
<td>73.2</td>
<td>65.8</td>
<td>72.1</td>
<td>95.6</td>
<td>79.8</td>
</tr>
</tbody>
</table>
TABLE III
ABLATION STUDY OF THE INDIVIDUAL COMPONENT OF ASEMM.

<table>
<thead>
<tr>
<th>SSH</th>
<th>MM</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APgpsm</th>
<th>AR</th>
<th>AR50</th>
<th>AR75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>61.4</td>
<td>90.1</td>
<td>69.8</td>
<td>63.7</td>
<td>69.6</td>
<td>93.6</td>
<td>77.4</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>62.9</td>
<td>91.1</td>
<td>71.6</td>
<td>64.7</td>
<td>70.4</td>
<td>94.3</td>
<td>78.2</td>
<td></td>
</tr>
<tr>
<td>✓ ✓</td>
<td>63.7</td>
<td>91.6</td>
<td>72.2</td>
<td>65.1</td>
<td>70.8</td>
<td>94.6</td>
<td>78.5</td>
<td></td>
</tr>
</tbody>
</table>

TABLE IV
COMBINATION STRATEGIES FOR MODULATION MODULE.

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APgpsm</th>
<th>AR</th>
<th>AR50</th>
<th>AR75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>61.4</td>
<td>90.1</td>
<td>69.8</td>
<td>63.7</td>
<td>69.6</td>
<td>93.6</td>
<td>77.4</td>
</tr>
<tr>
<td>F_s + F_i/F_{uw}</td>
<td>63.4</td>
<td>91.2</td>
<td>71.3</td>
<td>65.2</td>
<td>64.4</td>
<td>70.8</td>
<td>94.5</td>
</tr>
<tr>
<td>dc(F_s, F_i/F_{uw})</td>
<td>64.1</td>
<td>91.6</td>
<td>73.0</td>
<td>65.5</td>
<td>71.6</td>
<td>94.9</td>
<td>79.4</td>
</tr>
<tr>
<td>dc(F_s, F_i/F_{uw}) + F_i/F_{uw}</td>
<td>64.7</td>
<td>91.4</td>
<td>73.2</td>
<td>65.8</td>
<td>72.1</td>
<td>95.6</td>
<td>79.8</td>
</tr>
</tbody>
</table>

TABLE V
ABLATION STUDY OF THE NUMBER OF CONVOLUTION LAYER IN SEPARATION SEMANTIC HEAD.

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APgpsm</th>
<th>AR</th>
<th>AR50</th>
<th>AR75</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>61.4</td>
<td>90.1</td>
<td>69.8</td>
<td>63.7</td>
<td>69.6</td>
<td>93.6</td>
<td>77.4</td>
</tr>
<tr>
<td>(1.1)</td>
<td>62.6</td>
<td>91.1</td>
<td>70.1</td>
<td>64.3</td>
<td>70.4</td>
<td>94.1</td>
<td>78.0</td>
</tr>
<tr>
<td>(2.2)</td>
<td>63.7</td>
<td>91.6</td>
<td>72.2</td>
<td>65.3</td>
<td>71.6</td>
<td>94.6</td>
<td>78.5</td>
</tr>
<tr>
<td>(3.3)</td>
<td>63.5</td>
<td>91.2</td>
<td>72.0</td>
<td>65.1</td>
<td>70.8</td>
<td>94.5</td>
<td>79.7</td>
</tr>
<tr>
<td>(4,4)</td>
<td>63.8</td>
<td>91.8</td>
<td>72.6</td>
<td>65.6</td>
<td>71.6</td>
<td>95.0</td>
<td>79.7</td>
</tr>
</tbody>
</table>

human surface information.

D. Ablation Study

Individual Component Contribution. GC RCNN integrates the adaptively semantic enhancement module (ASEM) and the geodesic-consistent loss (GCL) to the standard DensePose estimation framework. Table III evaluates the effects of each module. We use the DensePose RCNN with ResNet-R50 as the base model. ASEM improved performance to 63.7% AP, 2.3% AP higher than the base model. Moreover, ASEM achieved consistent improvements across all evaluation metrics, which proves the effectiveness of modeling explicit IUV interdependent representation. GCL improved performance to 62.8% AP, 1.4% AP higher than the base model. Specifically, it obtains 1.6% AP75 gains and 0.5% AP50 gains. More performance gains on AP75 validate the effectiveness in minimizing surface distance. Finally, combining ASEM and GCL, GC RCNN improves the performance to 64.7% AP and 72.1% AR, 3.3% AP, and 2.5% AR gains compared with the base model.

Adaptively semantic enhancement module. The ASME proposes the separation semantic head (SSH) to explicitly learn multi-semantic features and the modulation module (MM) to adjust the contribution proportion of these features adaptively. Table III evaluates the effects of the components of ASEM. Compared with the base model, the separation semantic head improved performance to 62.9% AP, which is 1.5% AP higher than the base model. Moreover, the separation semantic head has yielded consistent enhancements across all evaluative metrics, proving the effectiveness of explicitly learning multi-semantic features. The modulation module is structurally dependent on the separation semantic head, and then we evaluate the combination performance. With the separation semantic head and modulation module, ASME further improved performance to 63.7% AP, 2.3% AP higher than the base model, and 0.8% AP higher than the separation semantic head. It proves that merging the aligned shared features enhances the human surface semantics.

Individual Component Contribution in modulation module. We study multiple combination strategies for the modulation module. As shown in Table IV, F_s + F_i/F_{uw} directly adds the shared feature F_s and the surface category semantic feature F_i or the local surface position semantic feature F_{uw} for IUV estimation, respectively. dc(F_s, F_i/F_{uw}) uses the aligned global semantic features from the dynamic convolution for IUV estimation. Compared with the base model, F_s + F_i/F_{uw} improves the performance to 63.4% AP, achieving 2% AP gains. Our aligned global semantic feature dc(F_s, F_i/F_{uw}) achieves 64.7% AP, with 3.3% AP gains compared with the base model. We finally chose the dc(F_s, F_i/F_{uw}) + F_i/F_{uw} as our combination strategy for the modulation module to achieve the optimal.

The number of convolution layers in separation semantic head. As shown in Figure 4, the separation semantic head stacks the parallel two 3 × 3 convolution layers to extract the surface category semantic feature and the local surface position semantic feature. We perform the ablation study to prove that the lightweight design is sufficient to extract semantic features. As shown in Table V, the (0,0) shows the performance of the base model, which only uses the shared feature to estimate IUV. The performance achieves 1.2% AP gains when we stack one convolution layer and 2.3% AP gains when we stack two convolution layers. Compared with using two convolution layers, using three convolution layers decreases the performance by 0.2% AP, and using four convolution layers only achieves 0.1% AP gain. Moreover, both settings...
TABLE VI
ABLAST STUDY OF THE INDIVIDUAL COMPONENT OF THE GEODESIC-CONSISTENT LOSS.

<table>
<thead>
<tr>
<th>Component</th>
<th>AP 50</th>
<th>AP 75</th>
<th>AP g诚m</th>
<th>AR 90</th>
<th>AR 75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>61.4</td>
<td>90.1</td>
<td>69.8</td>
<td>63.7</td>
<td>69.6</td>
</tr>
<tr>
<td>✓</td>
<td>62.0</td>
<td>90.3</td>
<td>70.1</td>
<td>63.9</td>
<td>69.6</td>
</tr>
<tr>
<td>✓</td>
<td>62.8</td>
<td>90.6</td>
<td>71.4</td>
<td>63.9</td>
<td>70.5</td>
</tr>
</tbody>
</table>

TABLE VII
ABLATION STUDY OF THE GROUND-TRUTH VARIANCE.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>AP 50</th>
<th>AP 75</th>
<th>AP g诚m</th>
<th>AR 90</th>
<th>AR 75</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>64.5</td>
<td>91.4</td>
<td>72.9</td>
<td>65.7</td>
<td>71.6</td>
</tr>
<tr>
<td>0.01</td>
<td>63.66</td>
<td>90.97</td>
<td>72.69</td>
<td>65.0</td>
<td>71.2</td>
</tr>
<tr>
<td>0.001</td>
<td>64.7</td>
<td>91.4</td>
<td>73.2</td>
<td>65.8</td>
<td>72.1</td>
</tr>
</tbody>
</table>

TABLE VIII
PARAMETERS AND FLOPS COMPARISON WITH THE SOTA METHODS.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>Flops</th>
<th>FLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP RCNN</td>
<td>17.7M</td>
<td>13.9G</td>
<td>-</td>
</tr>
<tr>
<td>DensePose RCNN</td>
<td>59.6M</td>
<td>206.4G</td>
<td>0.0828</td>
</tr>
<tr>
<td>ASEM</td>
<td>27.2M</td>
<td>21.4G</td>
<td>-</td>
</tr>
<tr>
<td>ASEM MM</td>
<td>4.9M</td>
<td>3.7G</td>
<td>-</td>
</tr>
<tr>
<td>ASEM</td>
<td>32.1M</td>
<td>25.1G</td>
<td>-</td>
</tr>
<tr>
<td>GCL</td>
<td>0.2M</td>
<td>0.2G</td>
<td>-</td>
</tr>
<tr>
<td>GC RCNN</td>
<td>74.3M</td>
<td>218.0G</td>
<td>0.1409</td>
</tr>
</tbody>
</table>

Ground-truth variance \( \hat{\sigma} \) setting. We model the ground-truth UV as a Gaussian distribution \( Q \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}) \) to compute the KL divergence between the estimated IUV distribution and the ground-truth IUV distribution. The \( \hat{\mu} \) is the ground-truth UV coordinates, and the \( \hat{\sigma} \) is the variance describing the uncertainty. Theoretically, the uncertainty associated with the ground truth should diminish towards zero ad infinitum. In order to simplify the training strategy and ensure stability, we predefined the \( \hat{\sigma} \) as a hyperparameter. Table VII shows the ablation study of the \( \hat{\sigma} \). \( \hat{\sigma} = 0.001 \) improves the performance most. A substantial \( \hat{\sigma} \) fails to conform to the ground-truth UV distribution, whereas a small \( \hat{\sigma} \) causes the KL divergence in Equation (2) to be dominated by \( \hat{\sigma} \). We finally set \( \hat{\sigma} \) to 0.001 to achieve the best performance.

Geodesic-consistent loss. The geodesic-consistent loss models the UV distribution as a multi-modal gaussian distribution, in which an extra deconvolution layer generates the variance \( \hat{\sigma} \). We use the KL divergence to measure the distribution deviation of each surface. Then the classification GCL and the regression GCL are proposed based on the KL divergence between the IUV distribution to optimized the surface distance directly. Table VII evaluates the effects of the classification GCL and the regression GCL. Compared with the base model, the regression GCL improved performance to 62.0% AP, 0.6% AP higher than the base model. Moreover, the regression GCL achieved consistent slight improvements across all evaluation metrics. The reason is that the regression GCL is only used for UV optimization and still suffers from serious surface misclassification, which is the main factor causing most failure cases, as we mentioned in the introduction. The classification GCL is structurally dependent on the regression GCL to optimize the \( \hat{\sigma} \), and we then evaluate the combination performance. With both geodesic-consistent losses, we further improved performance to 62.8% AP, 1.4% AP higher than the base model, and 0.8% AP higher than the regression GCL. It proves the effectiveness of optimizing IUV based on the IUV distribution deviations.

To achieve consistent slight improvements across all evaluation
cases, as we mentioned in the introduction. The classification GCL is structurally dependent on the regression GCL to optimize the \( \hat{\sigma} \), and we then evaluate the combination performance. With both geodesic-consistent losses, we further improved performance to 62.8% AP, 1.4% AP higher than the base model, and 0.8% AP higher than the regression GCL. It proves the effectiveness of optimizing IUV based on the IUV distribution deviations.

E. Visualization

We present the visualization of the surface classification and UV isocontours of GC RCNN in various complex scenarios: (1) unusual postures, (2) image truncation, (3) occlusion, and (4) crowd. Figure 5 shows the qualitative comparison between the DensePose RCNN and our GC RCNN, revealing that: (1) GC RCNN mitigates the ambiguity in surface categorization induced by unusual postures. For example, as illustrated in Figure 5 (row 1), DensePose RCNN struggles with distinguishing the surface category of the lower limb for the sitting instance, resulting in irregular surface segmentation and disordered UV isocontours. Conversely, GC RCNN accurately segmented the surface categories within this area and generated smooth UV isocontours. (2) GC RCNN can handle irregular surface edges. As illustrated in Figure 5 (row 2), DensePose RCNN struggles to accurately distinguish surface categories at the edges of truncated images, producing tangled UV isocontours. GC RCNN alleviates this problem by efficaciously modeling surface interrelations within two-
dimensional imagery. (3) GC RCNN can generate regionally consistent IUV estimates. As illustrated in Figure 5 (row 3), DensePose RCNN tends to identify the area as the background for the area occluded by the umbrella pole. In contrast, GC RCNN tends to ignore the occlusion, ensuring the regional homogeneity of the human surface. (4) GC RCNN reduces missed detections in crowded scenes. As illustrated in Figure 5 (row 4), GC RCNN identifies instances missed by DensePose.
RCNN, facilitating smoother IUV estimates.

**Limitation.** We visualize the failure cases of GC RCNN in Figure 6. It can be seen that: (1) GC RCNN is still prone to missed detections for severely occluded instances in crowded scenes, as shown in Figure 6 (row 1); (2) GC RCNN may confuse people with surrounding objects, as shown in Figure 6 (row 2). GC RCNN has difficulty detecting small targets whose body parts are severely occluded. For example, as shown in Figure 6 (row 1), GC RCNN missed detection for people with only visible heads. A more robust object detector for small objects may provide assistance to DensePose. In order to ensure the completeness and smoothness of the estimation, GC RCNN may erroneously detect surrounding objects that do not belong to the instance as its body parts. For example, as shown in Figure 6 (row 2), the elephant’s trunk is mistakenly identified as a part of the human surface when it intersects with the right limb of a person. Similarly, for a baby in a stroller, GC RCNN incorrectly detects the stroller as part of the baby, substituting for the occluded lower limbs. These failure cases reflect the shortcomings of GC RCNN in learning human overall structure. Leveraging human structural knowledge is proven to be an efficient means of pose estimation, but more in-depth research is needed in DensePose.

**V. Conclusion**

In this paper, we focus on modeling intrinsic interdependency within the IUV components to meet the challenge of the inconsistency between training strategies and task objectives. We propose Geodesic-consistent RCNN (GC RCNN) to enhance human surface understanding and facilitate surface distance optimization, which integrates the adaptively semantic enhancement module (ASEM) and the geodesic-consistent loss (GCL) to the standard DensePose estimation framework. ASEM enhances human surface semantics by extracting distinct IUV features and dynamically modeling the feature-level IUV interdependence. It proposed the separation semantic head to learn multi-semantic features explicitly, and the modulation module to adaptively adjust the relative contribution among these features. GCL is a differentiable loss function that approximates the surface distance by quantifying the degree of deviation between the estimated IUV distribution and the ground truth. The distribution deviation is used to facilitate IUV optimization, thereby aiding in minimizing surface distance. GC RCNN has verified its effectiveness and advancement on several challenging benchmarks and can be a cornerstone and help future research in DensePose. However, our GC RCNN is still prone to missed detections for severely occluded instances in crowded scenes and may confuse people with surrounding objects. Leveraging human structural knowledge is proven to be an efficient means of pose estimation, but more in-depth research is needed in DensePose. Our future work will focus on this aspect.

**VI. Acknowledgements**

This work is supported by Guangdong Basic and Applied Basic Research Foundation (Grant No.2414050001530).

**VII. References**

**References**


**Biography**

**Junyao Sun** is a Ph.D. candidate in the School of Software Engineering, South China University of Technology, China. She received the B.S. degree in South China University of Technology in 2018. Her research focuses on pose estimation and its sub-tasks, dense human pose estimation. **Qiong Liu** received her BE degree in automation dept. from Tsinghua University in 1982, and MS and Ph.D. degrees respectively in automation dept. in 1985 and School of Biomedical Engineering in 1996 from Chongqing University. Then She was a professor in school of software engineering, South China University of Technology. Her main research interests now are computer vision application technology using machine leaning and pattern recognition etc.