GigaHumanDet: Exploring Full-Body Detection on Gigapixel-Level Images

Chenglong Liu[1,2,*], Haoran Wei[3,*], Jinze Yang[1], Jintao Liu[1], Wenxi Li[4], Yuchen Guo[2,†], Lu Fang[2,†]

[1]University of Chinese Academy of Sciences
[2]BNRist, Tsinghua University
[3]MEGVII Technology
[4]Shanghai Jiao Tong University
{liuchenglong20, weihaoran18, yangjinze20, liujintao201}@mails.ucas.ac.cn
wenxi.li@sjtu.edu.cn, yuchen.w.guo@gmail.com, fanglu@tsinghua.edu.cn

Abstract

Performing person detection in super-high-resolution images has been a challenging task. For such a task, modern detectors, which usually encode a box using center and width/height, struggle with accuracy due to two factors: 1) Human characteristic: people come in various postures and the center with high freedom is difficult to capture robust visual pattern; 2) Image characteristic: due to vast scale diversity of input (gigapixel-level), distance regression (for width and height) is hard to pinpoint, especially for a person, with substantial scale, who is near the camera. To address these challenges, we propose GigaHumanDet, an innovative solution aimed at further enhancing detection accuracy for gigapixel-level images. GigaHumanDet employs the corner modeling method to avoid the potential issues of a high degree of freedom in center pinpointing. To better distinguish similar-looking persons and enforce instance consistency of corner pairs, an instance-guided learning approach is designed to capture discriminative individual semantics. Further, we devise reliable shape-aware bodyness equipped with a multi-precision strategy as the human corner matching guidance to be appropriately adapted to the single-view large scene. Experimental results on PANDA and STCrowd datasets show the superiority and strong applicability of our design. Notably, our model achieves 82.4% in term of AP, outperforming current state-of-the-arts by more than 10%.

Introduction

Person detection is a fundamental and critical task for human-centric visual analysis. Recently, the resolution has reached gigapixel level (*e.g.*, 25k×14k pixels) (Wang et al. 2020), posing a challenge for object detectors to cover the analysis at large-scale spatial range with clear local details.

Most modern detectors fall into the center-regression-guided type (Cai and Vasconcelos 2018; Tian et al. 2019; Hasan et al. 2021; Zhang et al. 2022), which locates objects via centers and prefers clear four boundaries to regress width/height (see Figure 1 (d)). But for the human-centric task in gigapixel images, it seems that the requirements of the above methods cannot be perfectly met. Numerous individuals exhibit various postures, leading to the instability of human center visual pattern (*e.g.*, Figure 1 (a): the center of the bending man lies on the ground while the center of the standing woman is occluded), further making it harder to determine four boundaries. Besides, because people are at different distances from the gigacamera (Yuan et al. 2017), there is vast variation in human size (*e.g.*, Figure 1 (c): from

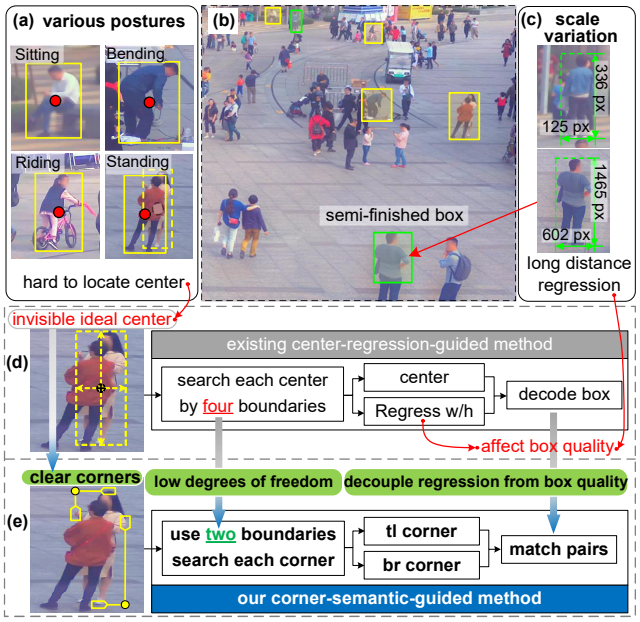

Figure 1: Part (b) shows a partial slice of the gigapixel-level image. Compared to center-regression-guided detectors in (d), our corner-semantic-guided model (e) enjoys the following strengths. 1) Our model locates persons by matching the top-left and bottom-right corners with low degrees of freedom, rather than centers, whose visual patterns are unstable or even invisible due to the diverse postures of flexible persons in (a). 2) Our model relies on a robust corner-matching algorithm to construct a box, so that the box quality is not affected by long-distance regression error in the single-view large scene (c) where the sizes of persons change drastically.

*These authors contributed equally.
†Corresponding authors.

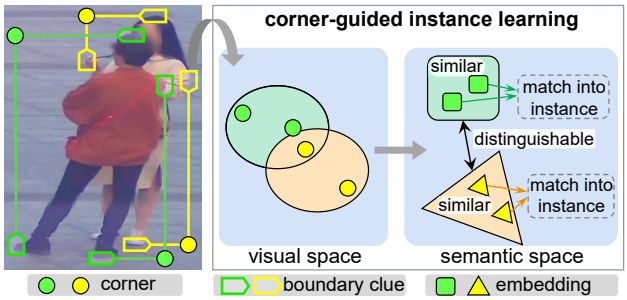

Figure 2: Framework of GigaHumanDet. By discriminative semantics and robust body embeddings, our model performs gigapixel-level full-body detection by matching corner pairs.

125×336 to 602×1465). Thus, long-range regression errors directly affect the quality of boxes produced by modern detectors that regress width/height, and the box may even be mistakenly considered a negative sample since IoU with the ground truth is under threshold (*e.g.*, yields a semi-finished box that only covers half of the body in Figure 1(b)).

Instead of the above detectors, another is the corner-semantic-guided type (Law and Deng 2018), which estimates two corners of the box via boundary clues. We find this type is more suitable for gigapixel-level human detection: As the best knowledge, the degree of freedom of determining a corner is lower than that of a center (a perfect center needs four borders while a corner only needs two), and manually generating a human annotation box via two corners is a strong prior knowledge for us. In Figure 1 (d) (e), although the ideal center sample is occluded, each corner point can still be located by two clear boundary clues. Further, the object representations of corner detectors do not rely on distance regression, so they are robust to multi-scale changes of the object and enjoy greater potential to generate higher-quality boxes in such large-scale single-view scenes.

However, existing corner-semantic-guided methods designed for common scenarios can not work well in the extremely large-scale fully-body detection task, especially their heuristic corner-matching algorithms. For example, CornerNet (Law and Deng 2018) determines corner pairs by local response similarity, which can cause severe confusion in detecting humans with similar appearances in the large spatial scale scene. CenterNet (Duan et al. 2019) predicts one more center point to filter out the false positives (FP), but it still can not perform well when applied to a gigapixel image, because the central regions of numerous FPs often cover the center point of a third person. Therefore, we may ask: *Can we devise a robust corner matching algorithm for gigapixel-level full-body detection?*

To answer the question and expand the applicability of corner-semantic-guided detectors for full-body detection on gigapixel-level images, we propose GigaHumanDet. GigaHumanDet predicts two decoupled corners of the bounding box, each of which is expressed explicitly and requires only two borders to be determined, improving the robustness in gigapixel crowds with various postures. To acquire discriminative corner semantics for similar-looking persons, we devise an instance-guided learning strategy as shown in Figure 2. For reliably matching corners, we devise shape-aware bodyness which encodes the body shape embeddings at the corresponding corner location. To further purify the body shape embedding and make it more tolerant to the drastic scale variation, a Gaussian-inspired multi-precision regression strategy is devised to alleviate the difficulty and inaccuracy of once long-range regression. Note that the distance regression is decoupled from the corner prediction, so the corner pair can decode an accurate object box without being affected by the distance regression error. Equipped with robust bodyness, GigaHumanDet can reap promising performance by pairing corners with consistent body shapes.

Experimental results on gigapixel-level PANDA(Wang et al. 2020) benchmark show GigaHumanDet yields a new SOTA accuracy, in terms of 82.4% on AP_{50}, boosting 60.8% than CornerNet baseline and surpassing other methods by more than 10%. Further, the competitive results on STCrowd verify the applicability of our method to general pedestrian detection. Our contributions can be summarized as follows:

- This paper unleashes the power of corner modeling approach on gigapixel-level full-body detection, and we demonstrate that it is more suitable and robust than center-regression-guided methods for this task.
- We design instance-guided learning and multi-precision strategy to acquire discriminative corner semantics. We propose shape-aware bodyness to provide reliable corner-matching guidance for large-scale dense scenes.
- GigaHumanDet achieves the SOTA accuracy on the gigapixel level detection task and outperforms other advanced methods by 10%.

Related Work

Object Detection on High-Resolution Images

Object detection on large-scale HR images has become a challenging task. A gigapixel-level human-centric PANDA dataset(Wang et al. 2020) is published and its resolution has reached 25,000×14,000. Due to wide FoV and high resolution, pedestrians have various postures as well as occlusions, and drastic scale changes exist, which degrades the accuracy of modern detectors developed on COCO(Lin et al. 2014).

Center-Guided and Corner-Guided Methods

The center-guided detectors utilize centers and width/height to encode boxes. Most of them(He et al. 2017; Cai and Vasconcelos 2018; Ge et al. 2021) take the center as a reference point and regress object size. Different from the above methods, the corner-guided detector(Law and Deng 2018; Duan et al. 2019) is proposed to estimate corners and match them to compose the final box. But for gigapixel images, humans with similar appearances cause great confusion when matching corners. Our GigaHumanDet employs the corner modeling method and tackles the matching problem by setting robust body-shape embedding for each corner.

Pedestrian Detection

The pedestrian detection task has been widely studied (Cai et al. 2016). Some full body detectors try to lift the accu-

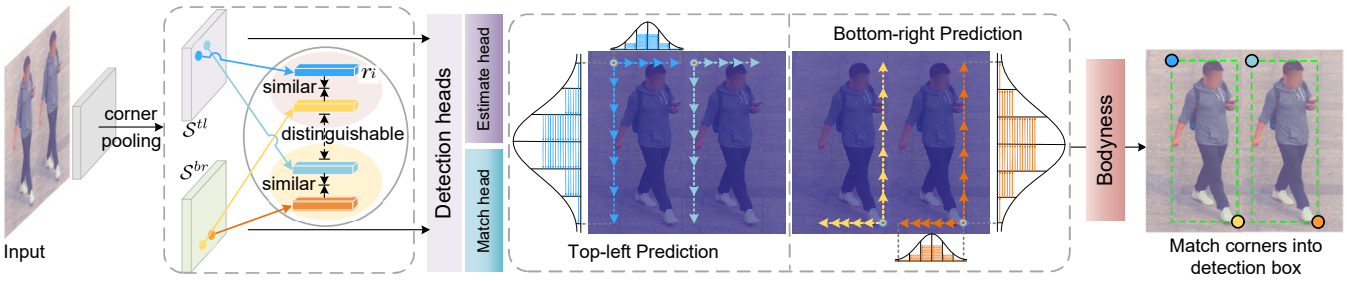

Figure 3: The pipeline of GigaHumanDet. First, the features with discriminative semantics are obtained using instance-guided learning. Then, the estimating head is responsible for outputting the top-left/bottom-right corner heatmaps. The matching head predicts the body shape embedding at each corner location via Gaussian-inspired multi-precision strategy. Finally, by the measurement of the shape-aware bodyness, a pair of corners with high affinity can be matched to yield the precise detection box.

racy using extra information (Mao et al. 2017; Brazil, Yin, and Liu 2017) or attention mechanism (Lin et al. 2018; Pang et al. 2019). Besides, several novel NMS algorithms(Huang et al. 2020; Liu, Huang, and Wang 2019) are devised to purify the dense boxes effectively. Note that the above works are about small-scale images and generally follow the center-regression-guided method. Different from them, we focus on conquering gigapixel images and effectively perform full-body detection using corner pairs for the first time.

Proposed Method

Given a gigapixel image $\mathcal{I}$, a full-body detection system is required to output the coordinates of all instance locations. GigaHumanDet predicts all top-left corners $\{(x_i^{tl}, y_i^{tl})\}_{i=0}^{m}$ and bottom-right corners $\{(x_i^{br}, y_i^{br})\}_{i=0}^{m}$, and match them into full-body boxes $\{(x_i^{tl}, y_i^{tl}, x_i^{br}, y_i^{br})\}_{i=0}^{m}$. m is the total number. Next, we will delve into each part of the model.

Instance-Guided Learning

The GigaHumanDet is required to efficiently obtain the discriminative semantic features for numerous persons with similar appearances in the single-view large scene, as illustrated in Figure 2. Therefore, we devise an instance-guided learning strategy to provide more accurate individual representations for the whole prediction process.

During training, the input image $\mathcal{I}$ is fed into the backbone and corner pooling layers to obtain deep feature maps with rich corner semantics, which is formulated as follows:

$$\mathcal{S}^{tl} = \delta^{tl}[\Phi(\mathcal{I})],\ \mathcal{S}^{br} = \delta^{br}[\Phi(\mathcal{I})] \tag{1}$$

where Φ denotes the backbone network. δ^{tl} and δ^{br} are the top-left and bottom-right corner pooling layers, respectively. $\mathcal{S}^{tl}$ and $\mathcal{S}^{br}$ are the corresponding feature maps. Then, at each corner location, we can extract i-th top-left/bottom-right corner embedding $\boldsymbol{t}_i/\boldsymbol{b}_i$ as follows:

$$\boldsymbol{t}_i = \mathcal{S}^{tl}(x_i^{tl}, y_i^{tl}),\ \boldsymbol{b}_i = \mathcal{S}^{br}(x_i^{br}, y_i^{br}) \tag{2}$$

where (x_i^{tl}, y_i^{tl}) and (x_i^{br}, y_i^{br}) are the ground-truth coordinates of the corner pair. Then, we can get the collections $\{\boldsymbol{t}_i\}_{i=1}^{m}/\{\boldsymbol{b}_i\}_{i=1}^{m}$ of corner embeddings for the m persons on the input image. As shown in Figure 3, similar-looking persons often co-occur in the visual space on the image. The embedding $\boldsymbol{t}_i/\boldsymbol{b}_i$ can represent the identity of the i-th instance. To purify each identity in the semantic space, the embeddings pair $\boldsymbol{t}_i/\boldsymbol{b}_i$ of the same instance is expected to be very similar and effectively distinguishable from the other embeddings of $(m-1)$ instances. To endow the vanilla corner embeddings with such discriminative power, we do so through the contrastive loss.

$$\{\boldsymbol{r}_i\}_{i=1}^{2m} = \{\boldsymbol{t}_i\}_{i=1}^{m} \cup \{\boldsymbol{b}_i\}_{i=1}^{m} \tag{3}$$

The collection of overall corner representations is obtained by Eq. 3. For each corner embedding $\boldsymbol{r}_i$, we constrain it by making it similar to its paired one and minimizing similarity to the remaining embeddings, which can be formulated as:

$$\mathcal{L}_{ins} = -\log \frac{\exp\left(\langle \overline{\boldsymbol{r}}_i, \overline{\boldsymbol{r}}_p \rangle / \tau\right)}{\sum_{j=1}^{2m} \exp\left(\langle \overline{\boldsymbol{r}}_i, \overline{\boldsymbol{r}}_j \rangle / \tau\right)} \tag{4}$$

where τ denotes the temperature hyper-parameter and is empirically set to 0.05. $\overline{\boldsymbol{r}}_i$ is the l_2 normalized embedding for a corner of the specific instance. $\overline{\boldsymbol{r}}_p$ is the corresponding positive corner, *i.e.*, the paired embedding of the same instance. Therefore, instance-guided unique semantic features are obtained and facilitate the subsequent prediction.

Estimate Human Corners

The GigaHumanDet adopts corner pooling(Law and Deng 2018) to enhance corner semantics, which helps to determine the top-left (bottom-right) corner point by looking along the top and left (bottom and right) boundary directions. Compared to the center prediction that requires all four borders of instance, our GigaHumanDet focuses on the corner that relies on just two boundaries, which is with lower degrees of freedom and more meaningful to human detection where some body parts are often invisible (see Figure 2).

We utilize heatmaps to predict corner keypoints, *i.e.*,

$$\mathcal{K}^{tl} = \Psi^{tl}[\mathcal{S}^{tl}],\ \mathcal{K}^{br} = \Psi^{br}[\mathcal{S}^{br}] \tag{5}$$

where Ψ^{tl} and Ψ^{br} are heatmap modules. $\mathcal{K}^{tl}$ and $\mathcal{K}^{br}$ are the predicted keypoint heatmaps. During training, each corner is mapped into a Gaussian region to reduce the penalty given to the negative samples near the true positive corner location. The distance-penalty-aware Focal Loss is adopted as the optimization objective, which is formulated as follows:

$$\mathcal{L}_c = -\frac{1}{N_p}\sum_{xy}\begin{cases}(1-p_{xy}')^{\alpha}\log(p_{xy}'), & \text{if } p_{xy}=1\\ (1-p_{xy})^{\beta}(p_{xy}')^{\alpha}\cdot\log(1-p_{xy}'), & \text{otherwise}\end{cases} \tag{6}$$

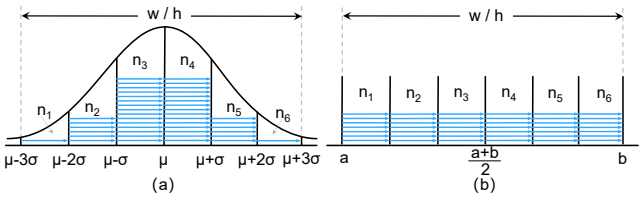

Figure 4: Illustration of multi-precision regression strategy. (a) is the Gaussian-inspired strategy. (b) is the uniform-inspired strategy. The w/h represents full-body width/height.

where p'_{xy} is the predicted value at the coordinate (x, y) on the heatmap and p_{xy} is the ground truth. The parameter α is used to adjust the weights of easy and hard samples and β controls the distance penalty reduction (we render α to 2 and β to 4 following CornerNet(Law and Deng 2018)).

Encode the Robust Body Shape Embedding to the Corner via Multi-Precision Strategy

In previous works(Law and Deng 2018; Duan et al. 2019), the self-supervised learned embeddings are used to determine whether a pair of corners belong to the same object, which depends on local response and presents serious confounds when dealing with people with similar appearances. Thus, we intend to encode the robust identity for the corner via the body shape (*i.e.*, width/height) embedding in a strong supervision manner. Specifically, we need to regress $(\boldsymbol{w}, \boldsymbol{h})$ at each top-left/bottom-right corner location. To alleviate the error brought by the long-distance regression, we devise a multi-precision strategy. In this case, the embeddings of human width/height can be formulated as follows:

$$\boldsymbol{w} = \left(w_1^1, w_2^1 \cdots w_{n_1}^1, w_1^2 \cdots w_{n_2}^2, \cdots, w_1^d, w_2^d \cdots w_{n_d}^d\right) \quad (7)$$

$$\boldsymbol{h} = \left(h_1^1, h_2^1 \cdots h_{n_1}^1, h_1^2 \cdots h_{n_2}^2, \cdots, h_1^d, h_2^d \cdots h_{n_d}^d\right) \quad (8)$$

$$s.t. \quad n_1 + n_2 + \cdots + n_k + \cdots + n_d = N, \quad k = 1, 2 \cdots d \quad (9)$$

where $\boldsymbol{w}, \boldsymbol{h} \in \mathbb{R}^N$. N is the dimension of width/height. We intend to divide the N components into d groups, and each group of n_d components is responsible for regressing the different partial distances of width/height. As a result, a single long-distance regression is transformed into a statistic for a collection including more accurate shorter regressions. As shown in Figure 4, two distribution-inspired strategies are designed to obtain the collection of multi-precision regressions. Both strategies imposes the following constraints on body shape embedding $\boldsymbol{w}/\boldsymbol{h}$:

$$\sum_{i=1}^{n_k} \xi(w_i^k) = N \cdot \left[F_w(w_i^0 + w_i^k) - F_w\left(\xi(k-2)w_i^0 + w_i^{k-1}\right)\right] \quad (10)$$

$$\sum_{i=1}^{n_k} \xi(h_i^k) = N \cdot \left[F_h(h_i^0 + h_i^k) - F_h\left(\xi(k-2)h_i^0 + h_i^{k-1}\right)\right] \quad (11)$$

where F_w/F_h is the cumulative distribution function (CDF) of the corresponding probability distribution and $\xi(\cdot)$ denotes the Heaviside function. w_i^k/h_i^k represents the partial human width/height of k-th group. w_i^0/h_i^0 corresponds to the starting point. In our experiments, we render d to 6 and the settings that fit the specific distribution are as follows:

- $Normal\left(\mu, \sigma^2\right)$: The number of components in each group is assigned by the CDF of the Gaussian distribution. μ and σ^2 are the mean and variance, respectively.

$$\begin{cases} \left[w_i^k\right]_{k=1}^6 \in [\sigma, \quad 2\sigma, \quad 3\sigma, \quad 4\sigma, \quad 5\sigma, \quad 6\sigma] \\ \sigma = w/6, \quad w_i^0 = \mu - 3\sigma \end{cases} \quad (12)$$

- $Uniform\,(a, b)$: The number of components in each group is assigned by the CDF of the uniform distribution. a and b are the two boundary values.

$$\begin{cases} \left[w_i^k\right]_{k=1}^6 \in \left[\frac{b-a}{6}, \frac{b-a}{3}, \frac{b-a}{2}, \frac{2(b-a)}{3}, \frac{5(b-a)}{6}, b-a\right] \\ b - a = w, \quad w_i^0 = a \end{cases} \quad (13)$$

Note that Equations 12 and 13 are detailed conditions for human width embedding $\boldsymbol{w}$, and those for height embedding $\boldsymbol{h}$ can be obtained in the same way. As a result, the network needs to produce two (top-left and bottom-right) regression maps with a $2N$-dimension (N-dimensional width and N-dimensional height). For training, we employ the Smooth L1 Loss(Girshick 2015) to learn the body shape embedding at ground-truth corner locations:

$$\begin{aligned} \mathcal{L}_e = \frac{1}{N_p} \sum_{j=1}^{N_p} (&\text{SmoothL1Loss}(\log(\boldsymbol{w_j}), \log(\hat{\boldsymbol{w}}_j)) \\ &+ \text{SmoothL1Loss}(\log(\boldsymbol{h_j}), \log(\hat{\boldsymbol{h}}_j))) \end{aligned} \quad (14)$$

where N_p denotes the head counts as well as the number of persons. $\boldsymbol{w_j}$ and $\boldsymbol{h_j}$ are the predicted body width/height embeddings, and $\hat{\boldsymbol{w}}_j$ and $\hat{\boldsymbol{h}}_j$ are the ground-truth ones. The function $\log(\cdot)$ maps the width/height into logarithmic space to make the learning process easier.

After training, we can decode the final body width/height by scaling their embeddings and obtaining the statistical mean of them:

$$w = \frac{1}{N} \sum_{k=1}^{d} \sum_{i=1}^{n_k} \alpha^k \cdot w_i^k, \quad h = \frac{1}{N} \sum_{k=1}^{d} \sum_{i=1}^{n_k} \alpha^k \cdot h_i^k \quad (15)$$

where α^k is the factor by which each component is scaled back to the corresponding width/height. Hence, the final body shape is the statistical mean of the multi-precision regression collection, which is shown in Figure 5. With this cooperation of shorter and longer distance regressions, the overall expected error decreases.

Notably, the regression of GigaHumanDet is very different from the popular detectors that yield the box via regression. Because our regression is completely independent of the quality of the detection box, it only enables the top-left/bottom-right corner to perceive human shape and provides effective guidance for the later matching process.

Match Corners via Shape-Aware Bodyness

After obtaining the predicted top-left and bottom-right corner heatmaps, we need to measure the affinity between the corner keypoints and match a pair to form a final detection box. Hence, we devise shape-aware bodyness to execute the corner-matching process. As shown in Figure 5, the key is

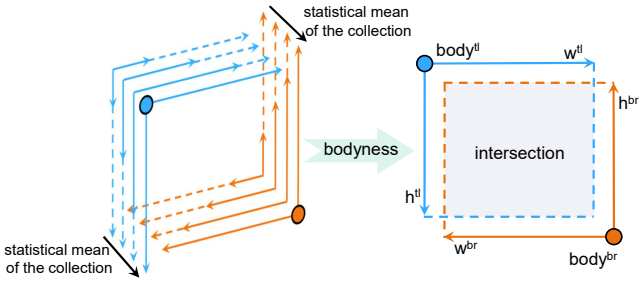

Figure 5: Illustration of bodyness. The final body shape (width/height) is the statistical mean of the multi-precision collection. Then, the top-left/bottom-right body boxes are decoded at corresponding corner locations. We devise bodyness based on the IoU and purify it by a shape-aware factor.

that human width/height at corner locations can decode the corresponding top-left/bottom-right body shape boxes and we define the shape-sware bodyness as follows:

$$\text{Bodyness} = \frac{body^{tl} \cap body^{br}}{body^{tl} \cup body^{br}} \cdot \overbrace{\exp\left(-2 \cdot (1-\lambda_s)^2\right)}^{\text{shape-aware term}} \tag{16}$$

$$\lambda_s = \sqrt{\frac{\min(w^{tl}, w^{br})}{\max(w^{tl}, w^{br})} \cdot \frac{\min(h^{tl}, h^{br})}{\max(h^{tl}, h^{br})}} \tag{17}$$

where $body^{tl}$ and $body^{br}$ denote the body boxes decoded by w^{tl}/w^{br} and h^{tl}/h^{br} at corner locations. λ_s is the shape-aware factor. Because the vanilla IoU is less sensitive to the shape, a shape-aware term is added to bodyness. If the shape similarity of two body boxes is small, then λ_s will be small, and the shape-aware term will help attenuate IoU. So the total bodyness score will be low. After calculating the bodyness of all estimated human corner pairs, those with higher bodyness are more likely to belong to the same body and the final precise detection box can be produced.

Experiments

Dataset and Evaluation Metrics

PANDA. We evaluate the proposed GigaHumanDet on the gigapixel-level human-centric dataset — PANDA (Wang et al. 2020). The images in PANDA cover various real-world scenes, including streets, markets, campuses, etc. It provides 13 scenarios for training and 5 scenarios for testing. Each scenario contains about 30 super high-resolution images and the image size reaches about 25k×14k.

STCrowd. STCrowd (Cong et al. 2022) is released recently and the total number of pedestrians is 219 K. There are 5263 and 2988 images with a size of 1280×720 in the training set and validation set, respectively.

Evaluation Metrics. We utilize AP (average precision) with the 0.5 and 0.75 IoU thresholds as the accuracy evaluation metric. The larger IoU threshold indicates that the detector can output higher-quality boxes. Objects of different sizes are divided as follows, small (<96×96 pixels), middle (96×96 to 288×288 pixels), and large (>288×288 pixels). And we also report AR (average recall) calculated under ten IoU thresholds (*i.e.*, 0.5:0.05:0.95). We also report #Pass (Fan et al. 2022), measuring the number of runs of the detector during the inference on gigapixel images.

Implementation Details

Training details. To prepare PANDA data for training, we downsample original gigapixel images by a factor of 4. Then, we decompose the down-sampled image using a 2,048×1,024 sliding window with a 0.25 overlap ratio. We train our model on four RTX 3090 GPUs with 4×24 GB RAM. The model with hourglass(Newell, Yang, and Deng 2016) or DLA-34(Yu et al. 2018) backbone is trained with a batch size of 43 for 40k iterations. The learning rate is set to 0.001, dropping 10× at the 30k iteration.

Testing details. For PANDA, we also downsample the original test gigapixel images and generate patches using the sliding window. The output heatmaps are filtered by the sigmoid function and 3×3 max pooling. Next, we decode the bodyness and match the top-left/bottom-right corners that belong to the same person. Finally, the results are performed Soft-NMS to remove redundant boxes. We only keep the top 500 detection boxes for each PANDA image.

Results on Gigapixel PANDA Dataset

Quantitative results. Quantitative results are reported in Table 1. Generally, GigaHumanDet achieves 82.4%, 51.8%, 80.9%, and 87.3% on AP_{50}, $\text{AP}^{\text{S}}_{50}$, $\text{AP}^{\text{M}}_{50}$, and $\text{AP}^{\text{L}}_{50}$, respectively, under Hourglass-104 backbone in single-scale testing, which is the state-of-the-art accuracy on the full body detection benchmark of the PANDA dataset. GigaHumanDet explores the pedestrian locations through corner keypoints with fewer degrees of freedom, rather than the more problematic centers, and utilizes robust body shape embeddings for effective corner matching. Therefore, GigaHumanDet lives up to expectations to enjoy a more promising performance than other popular detectors.

GigaHumanDet *vs.* center-guided methods. Compared with Cascade R-CNN(Cai and Vasconcelos 2018) which regresses dense center-based anchors to cover the entire image, GigaHumanDet lifts $\text{AP}^{\text{S}}_{50}$ by 29.1% (22.7% *vs.* 51.8%) thanks to avoiding the difficultly in learning unstable central visual features of humans with numerous postures. Even compared to DINO(Zhang et al. 2022) which adopts the advanced transformer architecture, the accuracy of GigaHumanDet is still far ahead with a 14.4% increase (*e.g.*, 68.0% *vs.* 82.4% on AP_{50}). This further demonstrates the excellence of the human corner estimating pipeline in such scenes where the partial boundaries are frequently not visible, and it is challenging to directly regress an accurate bounding box from the center to the four borders of the instance.

GigaHumanDet *vs.* advanced pedestrian detectors. We train some common pedestrian detectors on PANDA. CSP detects pedestrians by predicting center points and regressing scales, but it just reaches 61.5% on AP_{50} due to errors

Method	Backbone	Feature	#Pass	AP_{50}	AP^{S}_{50}	AP^{M}_{50}	AP^{L}_{50}
Center-regression-guided:							
Faster R-CNN (Ren et al. 2015)	ResNet-101	FPN	13,620	-	19.0	55.2	74.4
Cascade R-CNN (Cai and Vasconcelos 2018)	ResNet-101	FPN	-	-	22.7	57.9	76.5
RetinaNet (Lin et al. 2017)	ResNet-101	FPN	-	-	22.1	56.1	74.0
ATSS (Zhang et al. 2020)	ResNet-101	FPN	4,935	67.4	16.7	60.1	81.1
YOLOX (Ge et al. 2021)	CSPDarknet	FPN	4,935	69.5	21.1	63.5	81.5
DINO (Zhang et al. 2022)	ResNet-101	Encoder	4,935	68.0	18.0	63.0	79.6
ClusDet (Yang et al. 2019)	ResNet-50	FPN	7,871	71.8	21.9	69.6	78.2
PAN (Fan et al. 2022)	ResNet-50	FPN	3,671	71.5	25.6	71.9	76.8
CSP (Liu et al. 2019)	ResNet-101	FPN	4,935	61.5	28.1	64.1	64.3
Pedestron (Hasan et al. 2021)	HRNet	FPN	4,935	69.9	22.4	67.1	78.8
Corner-semantic-guided:							
CenterNet (Duan et al. 2019)	Hourglass-104	Single	4,935	26.0	6.7	20.7	33.4
CentripetalNet (Dong et al. 2020)	Hourglass-104	Single	4,935	62.6	27.3	60.1	69.6
CornerNet baseline (Law and Deng 2018)	Hourglass-104	Single	4,935	21.6	1.2	13.8	32.4
GigaHumanDet (Ours)	ResNet-50	Single	4,935	72.3	28.1	71.4	78.8
GigaHumanDet (Ours)	DLA-34	Single	4,935	78.2	41.4	76.8	84.2
GigaHumanDet (Ours)	Hourglass-52	Single	4,935	79.6	48.5	78.8	84.8
GigaHumanDet (Ours)	Hourglass-104	Single	4,935	82.4	51.8	80.9	87.3

Table 1: State-of-the-art comparisons in term of accuracy (%) on PANDA dataset. In this table, AP^{S}_{50}, AP^{M}_{50} and AP^{L}_{50} are for small, middle, and large objects, respectively. Note that all results of GigaHumanDet are obtained under single-scale testing.

Detector	Backbone	AP^{ST}_{50}
Faster R-CNN	ResNet-101	87.8
Cascade R-CNN	ResNet-101	88.7
RetinaNet	ResNet-101	88.8
DINO	ResNet-101	90.6
Pedestron	HRNet	90.5
CenterNet	Hourglass-104	84.2
CornerNet baseline	Hourglass-104	68.4
CentripetalNet	Hourglass-104	88.7
GigaHumanDet (Ours)	Hourglass-104	90.7

Table 2: Comparisons on general-resolution STCrowd.

of its single scale regression in the giga image. Cascade R-CNN with HRNet is an advanced baseline explored in Pedestron (Hasan et al. 2021). Nevertheless, under the same patch generation settings, it lags behind our GigaHumanDet by 12.5% on AP_{50}, firmly proving the superiority of our model.

GigaHumanDet *vs.* corner-guided methods. As shown in Table 1, CornerNet(Law and Deng 2018) has a poor AP_{50} of 21.6% because it is prone to mistakenly matching corners of two persons with a similar appearance. The AP_{50} of CenterNet(Duan et al. 2019) is stuck at 26.0% because its matching algorithm has low applicability in this scenario. With the robust body shape embedding and shape-aware bodyness, GigaHumanDet fully demonstrates its kingly demeanor in generating high-quality and high-reliability detection boxes, and yields a significant improvement of 60.8% on AP_{50} (from 21.6% to 82.4%) compared to the CornerNet baseline. The indicators of 51.8%, 80.9%, and 87.3% on AP^{S}_{50}, AP^{M}_{50}, and AP^{L}_{50} for objects of different area are clearly superior to other two corner-guided methods by large margins.

Visualization comparison. As shown in Figure 6, compared to CornerNet baseline which has poor accuracy due to numerous incorrect corner pairs, GigaHumanDet equipped with robust bodyness showcases reliable corner-matching results. Compared to center-guided YOLOX, DINO, and Cascade R-CNN, which suffer from regression errors, GigaHumanDet yields higher-quality boxes by decoupling regression from box encoding. Further, GigaHumanDet can find the adult and child missed by the other three center-visual-based methods, proving explicit corners with low degrees of freedom are more stable for gigapixel-level scenes.

Backbone	IGL	AP_{50}	AP_{75}	AP^{S}_{50}	AP^{M}_{50}	AP^{L}_{50}	AR
H-52	×	79.1	47.3	45.8	77.9	84.2	57.5
H-52	✓	79.6	48.0	48.5	78.8	84.8	57.9

Table 3: Effectiveness of instance-guided learning (IGL).

Results on General STCrowd

To show the generalization ability of GigaHumanDet on the common-resolution pedestrian detection task, we report the results of STCrowd(Cong et al. 2022) in Table 2. With robust corner-guided modeling, GigaHumanDet can still surpass other popular methods in the occluded crowd. Compared to CornerNet baseline, GigaHumanDet brings an AP_{50} improvement of 22.3% on STCrowd. All experimental results show that our GigaHumanDet enjoys strong applicability and can be a superior baseline model both in the gigapixel-level image and general image for full-body detection.

Ablation Study

Effectiveness of instance-guided learning. We devise an instance-guided learning (IGL) strategy to make corner semantics more discriminative. To verify the effectiveness of it, we retrain our model after removing the IGL loss (Eq.4). As shown in Table 3, IGL lifts 0.7% on AP_{75} and brings an impressive 2.7% increase on AP^{S}_{50}, proving that IGL can

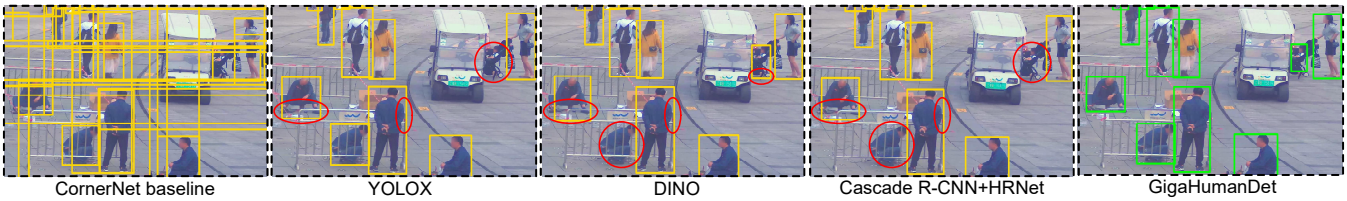

Figure 6: Visualization on PANDA. For clearer details, we zoom in on the local region of gigapixel images. Detection boxes are marked in yellow and green color. The red ellipses show that the object is missing or the box boundary is not precise enough.

Strategy	AP_{50}	AP_{75}	AP^{S}_{50}	AP^{M}_{50}	AP^{L}_{50}	AR
s.t. Gaussian	79.1	47.3	45.8	77.9	84.2	57.5
s.t. Uniform	78.8	46.8	45.4	77.8	83.9	57.1
single reg	78.1	46.0	45.3	77.5	83.1	56.9

Table 4: Effect of different regression strategies. These results are obtained under the Hourglass-52 backbone.

IoU	shape-aware	AP_{50}	AP_{75}	AP^{S}_{50}	AP^{M}_{50}	AP^{L}_{50}	AR
✓	×	78.3	47.1	45.6	77.0	83.8	56.9
✓	✓	79.1	47.3	45.8	77.9	84.2	57.5

Table 5: Effectiveness of the shape-aware term. These results are obtained under the Hourglass-52 backbone.

Threshold	0.3	0.35	0.4	0.45	0.5	0.55
AP_{50}	79.0	80.9	82.1	82.4	82.2	81.1

Table 6: Effect of different threshold values of bodyness.

Top-k	AP_{50}	AP_{75}	AP^{S}_{50}	AP^{M}_{50}	AP^{L}_{50}	AR
$k = 100$	80.6	55.8	47.8	77.6	86.7	62.3
$k = 200$	82.4	55.4	51.8	80.9	87.3	62.1
$k = 300$	81.6	54.9	51.3	80.2	87.0	61.4
$k = 400$	81.4	54.8	51.3	79.9	86.3	61.0
$k = 500$	80.8	54.7	50.9	79.4	86.1	60.7

Table 7: Comparisons of various number of keypoints kept on the predicted top-left/bottom-right heatmaps. These results are obtained under the Hourglass-104 backbone.

help acquire higher-quality boxes by learning more precise and distinguishable corners and body shape embeddings. And more accurate point-level semantics further purify corner estimation and matching, particularly for small bodies.

Effect of strategy for body shape embedding. As detailed in Section , two multi-precision strategies are proposed to encode body shape (width/height) for the gigapixel scene. Accordingly, we test the performance of GigaHumanDet when the regression collection is guided by the CDF of Gaussian and uniform distributions, respectively. As shown in Table 4, the model with the uniform-inspired strategy and H-52 backbone can yield a decent AP_{50} of 78.8%. Interestingly, we observe that the Gaussian-inspired strategy is 0.3% and 0.5% higher on AP_{50} and AP_{75} than the uniform-inspired one. If we replace the Gaussian multi-precision strategy with a common single-regression one, the accuracy will decrease by 1% on AP_{50}. The results indicate that the Gaussian-inspired one enjoys stronger robustness to match human corners and the body shape provided by its statistic of the collection is more precise. Therefore, the Gaussian-inspired multi-precision strategy is a better choice.

Analysis on the shape-aware term of bodyness. During matching human corner pairs using bodyness, we design a shape-aware term to directly measure the shape of bodies covered by two corners. As reported in Table 5, the proposed shape-aware term improves AP_{50} and AR by 0.8% and 0.6%, respectively. This demonstrates that shape-aware design is effective and can help bodyness perform better.

Effect of the threshold for bodyness. One threshold is required to keep these high-bodyness matching corner pairs to produce raw detection boxes. As shown in Table 6, we evaluate our method under the H-104 backbone by setting 6 different thresholds. And GigaHumanDet reaches the best AP_{50} with the threshold 0.45. Hence, we render the threshold of bodyness to 0.45 in all experiments.

Limitations for accuracy ceiling of GigaHumanDet. We firmly believe that our GigaHumanDet owns a high accuracy ceiling for the gigapixel-level human detection task. For example, a simple improvement solution is to adjust the number of corner keypoints to be retained on the output heatmaps and the results are shown in Table 7. As the number of reserved human corners increases from 100 to 500, the accuracy first increases and then decreases due to the inhomogeneity of crowd density in the large-scale spatial scenario. When the k value is set to 200, GigaHumanDet can yield the best AP_{50} of 82.4%, boosting of 1.8% than the case k =100. Since we have made progress with this rough solution, this fact shows that there is still a lot of room for optimization combined with head-counting approaches.

Conclusion

In this paper, we propose the corner-guided GigaHumanDet for human detection in gigapixel images and GigaHumanDet achieves state-of-the-art accuracy. The instance-guided learning and multi-precision strategies are devised to acquire discriminative body shape embeddings for each corner. Then, a reliable matching algorithm based on bodyness for gigapixel images is designed to yield precise boxes. We believe the proposed approaches enjoy promising potential for the gigapixel-level human-centric task.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2020AAA0105500, 2022ZD0119402), and in part by the National Natural Science Foundation of China (61971260, U21B2013).

References

Brazil, G.; Yin, X.; and Liu, X. 2017. Illuminating pedestrians via simultaneous detection & segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 4950–4959.

Cai, Z.; Fan, Q.; Feris, R. S.; and Vasconcelos, N. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, 354–370. Springer.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Cong, P.; Zhu, X.; Qiao, F.; Ren, Y.; Peng, X.; Hou, Y.; Xu, L.; Yang, R.; Manocha, D.; and Ma, Y. 2022. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19608–19617.

Dong, Z.; Li, G.; Liao, Y.; Wang, F.; Ren, P.; and Qian, C. 2020. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10519–10528.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6569–6578.

Fan, J.; Liu, H.; Yang, W.; See, J.; Zhang, A.; and Lin, W. 2022. Speed Up Object Detection on Gigapixel-Level Images With Patch Arrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4653–4661.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Hasan, I.; Liao, S.; Li, J.; Akram, S. U.; and Shao, L. 2021. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11328–11337.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Huang, X.; Ge, Z.; Jie, Z.; and Yoshie, O. 2020. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10750–10759.

Law, H.; and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, 734–750.

Lin, C.; Lu, J.; Wang, G.; and Zhou, J. 2018. Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of the European conference on computer vision (ECCV)*, 732–747.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, S.; Huang, D.; and Wang, Y. 2019. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6459–6468.

Liu, W.; Liao, S.; Ren, W.; Hu, W.; and Yu, Y. 2019. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5187–5196.

Mao, J.; Xiao, T.; Jiang, Y.; and Cao, Z. 2017. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3127–3136.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, 483–499. Springer.

Pang, Y.; Xie, J.; Khan, M. H.; Anwer, R. M.; Khan, F. S.; and Shao, L. 2019. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4967–4975.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9627–9636.

Wang, X.; Zhang, X.; Zhu, Y.; Guo, Y.; Yuan, X.; Xiang, L.; Wang, Z.; Ding, G.; Brady, D.; Dai, Q.; et al. 2020. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3268–3278.

Yang, F.; Fan, H.; Chu, P.; Blasch, E.; and Ling, H. 2019. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8311–8320.

Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2403–2412.

Yuan, X.; Fang, L.; Dai, Q.; Brady, D. J.; and Liu, Y. 2017. Multiscale gigapixel video: A cross resolution image matching and warping approach. In *2017 IEEE International Conference on Computational Photography (ICCP)*, 1–9. IEEE.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768.