Human-Art: A Versatile Human-Centric Dataset
Bridging Natural and Artificial Scenes

Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, Lei Zhang
1International Digital Economy Academy, 2The Chinese University of Hong Kong
{xju22, qxu}@cse.cuhk.edu.hk, {zengailing, wangjianan, leizhang}@idea.edu.cn
https://github.com/IDEA-Research/HumanArt

Abstract

Humans have long been recorded in a variety of forms since antiquity. For example, sculptures and paintings were the primary media for depicting human beings before the invention of cameras. However, most current human-centric computer vision tasks like human pose estimation and human image generation focus exclusively on natural images in the real world. Artificial humans, such as those in sculptures, paintings, and cartoons, are commonly neglected, making existing models fail in these scenarios.

As an abstraction of life, art incorporates humans in both natural and artificial scenarios. We take advantage of it and introduce the Human-Art dataset to bridge related tasks in natural and artificial scenarios. Specifically, Human-Art contains 50k high-quality images with over 123k person instances from 5 natural and 15 artificial scenarios, which are annotated with bounding boxes, keypoints, self-contact points, and text information for humans represented in both 2D and 3D. It is, therefore, comprehensive and versatile for various downstream tasks. We also provide a rich set of baseline results and detailed analyses for related tasks, including human detection, 2D and 3D human pose estimation, image generation, and motion transfer. As a challenging dataset, we hope Human-Art can provide insights for relevant research and open up new research questions.

1. Introduction

"Art is inspired by life but beyond it."

Human-centric computer vision (CV) tasks such as human detection [39], pose estimation [30], motion transfer [59], and human image generation [35] have been intensively studied and achieved remarkable performances in the past decade, thanks to the advancement of deep learning techniques. Most of these works use datasets [14, 23, 27, 39] that focus on humans in natural scenes captured by cameras due to the practical demands and easy accessibility.

However, besides being captured by cameras, humans are frequently present and recorded in various other forms, from ancient murals on walls to portrait paintings in digital form. However, existing state-of-the-art (SOTA) human detection and pose estimation models [64, 71] trained on commonly used datasets such as MSCOCO [27] generally work poorly on these scenarios. For instance, the average precision of such models can be as high as 63.2% and 79.8% on natural scenes but drops significantly to 12.6% and 28.7% on the sculpture scene. A fundamental reason is the domain gap between natural and artificial scenes. Also, the scarcity of datasets with artificial human scenes significantly restricts the development of tasks such as anime character image generation [9, 67, 74], character rendering [29], and character motion retargeting [1, 37, 68] in computer graphics and other areas. With the growing interest in virtual reality (VR), augmented reality (AR), and metaverse, this problem is exacerbated and demands immediate attention.

There are a few small datasets incorporating humans in artificial environments in the literature. Sketch2Pose [4] and ClassArch [33] collect images in sketch and vase painting respectively. Consequently, they are only applicable to the corresponding context. People-Art [61] is a human detection dataset that consists of 1490 paintings. It covers artificial scenes in various painting styles, but its categories are neither mutually exclusive nor collectively comprehensive. More importantly, the annotation type and image number in People-Art are limited, and hence this dataset is mainly used for testing (instead of training) object detectors.

Art presents humans in both natural and artificial scenes in various forms, e.g., dance, paintings, and sculptures. In this paper, we take advantage of the classification of visual arts to introduce Human-Art, a versatile human-centric dataset, to bridge the gap between natural and artificial scenes. Human-Art is hierarchically structured and includes high-quality human scenes in rich scenarios with precise
manual annotations. Specifically, it is composed of 50k images with about 123k person instances in 20 artistic categories, including 5 natural and 15 artificial scenarios in both 2D and 3D, as shown in Fig. 1. To support both recognition and generation tasks, Human-Art provides precise manual annotations containing human bounding boxes, 2D keypoints, self-contact points, and text descriptions. It can compensate for the lack of scenarios in prior datasets (e.g., MSCOCO [27]), link virtual and real worlds, and introduce new challenges and opportunities for human-centric areas.

Human-Art has the following unique characteristics:

- **Rich scenario**: Human-Art focuses on scenes missing in mainstream datasets (e.g., MSCOCO [27]), which covers most human-related scenarios. Challenging human appearances, diverse contexts, and various poses largely complement the scenario deficiency of existing datasets and will open up new challenges and opportunities.

- **High quality**: We guarantee inter-category variability and intra-category diversity in style, author, origin, and age. The 50k images are manually selected from 1,000k carefully collected images using standardized data collection, filtering, and consolidating processes.

- **Versatile annotations**: Human-Art provides carefully annotated images of 2D human keypoints, human bounding boxes, and self-contact points to support various downstream tasks. Also, we provide accessible text descriptions to enable multi-modality learning.

With Human-Art, we conduct comprehensive experiments and analysis on various downstream tasks including human detection, human pose estimation, human mesh recovery, image generation, and motion transfer. Although training on Human-Art can lead to a separate 31% and 21% performance boost on human detection and human pose estimation, results demonstrate that human-related CV tasks still have a long way to go before reaching maturity.

2. Related Work

Human-centric datasets with natural scenes: The main tasks in human-centric recognition are human detection and pose estimation. As summarized in Tab. 1, most existing datasets [2, 3, 11, 23, 28, 39, 62, 73] annotate humans in natural scenes with bounding boxes and keypoints. Among them, MSCOCO [27] is most widely used due to its diverse poses and complex scenes. Numerous deep models trained with it demonstrate high performance on various downstream tasks [15, 21, 36, 64]. Pedestrian detection datasets [26, 57, 65] can also be categorized as special human detection datasets focusing on small and hazy persons in congested situations.

Although widely used in computer vision tasks, exclusively focusing on natural scenes make models trained on these datasets fail in the artificial scenario.

Human-centric datasets with artificial scenes: Only a few small-scale datasets [4, 33, 61] involve artificial scenarios. Specifically, Sketch2Pose [4] focuses on the sketch scenario, and ClassArch [33] only includes ancient vase paintings. People-Art [61] contains both natural images and artificial images. It directly borrows the artistic painting styles from wikiart¹ for categorizing artificial scenes. Inter-category similarity causes confusion, especially when the images are not manually reviewed. Also, artworks beyond paintings (e.g., sculptures and digital arts) are ignored. More importantly, with the limited number of images and merely human bounding-box annotation, People-Art only supports small-scale human detection tasks.

Human-centric synthetic datasets: Various synthetic human body datasets [17, 19, 42, 60] are proposed in the literature. However, they are far less developed compared to artificial human face datasets [20, 75]. Generally speaking,

¹https://www.wikiart.org/
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image</th>
<th>Instance</th>
<th>Keypoint Number</th>
<th>Bbox</th>
<th>Pose</th>
<th>Self-Contact</th>
<th>Natural Scenario</th>
<th>Artificial Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC2012</td>
<td>8,174</td>
<td>17,132</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MSCOCO</td>
<td>66,808</td>
<td>273,469</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BodyHands</td>
<td>20,490</td>
<td>63,095</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>People-Art</td>
<td>1,490</td>
<td>3,870</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MSCI2[27]</td>
<td>58,945</td>
<td>156,165</td>
<td>17</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MPII[3]</td>
<td>24,920</td>
<td>40,522</td>
<td>16</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AI Challenger</td>
<td>240,000</td>
<td>448,776</td>
<td>14</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CrowdPose</td>
<td>20,000</td>
<td>~80,000</td>
<td>14</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OCHuman</td>
<td>4,731</td>
<td>8,110</td>
<td>17</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PoseTrack[2]</td>
<td>66,374</td>
<td>153,615</td>
<td>15</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HiEve[28]</td>
<td>49,820</td>
<td>1,099,357</td>
<td>14</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ClassArch[33]</td>
<td>1513</td>
<td>1728</td>
<td>17</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sketch2Pose[4]</td>
<td>808</td>
<td>14,772</td>
<td>18</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Human-Art (Ours)</td>
<td>50,000</td>
<td>123,131</td>
<td>21</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

1 Only calculate statistics of images that contain human bounding box annotation for detection;
2 Only calculate statistics of images that contain human keypoint annotation for human pose estimation.
3 Video-based datasets.

Table 1. Comparison of human-centric recognition datasets, including human detection and pose estimation tasks.

these datasets face the problems of unnatural interactions between background and humans and the lack of character diversities. For example, the anime/manga character dataset in [19] contains only 2631 different characters despite having more than a million images.

To sum up, it is essential and urgent to bridge the gap between natural and artificial scenes for human-centric computer vision tasks. This motivates Human-Art, a rich-scenario human-centric dataset containing sufficient high-quality images and versatile annotations.

3. The Human-Art Dataset

3.1. The Hierarchical Category Classification

As an abstraction of the natural world, art is a metaphorical expression of how people perceive the world.

This makes art a good point of penetration for scenes containing both natural and artificial scenarios.

According to [54], artistic presentations can be divided into eight classes: literature, music, architecture, painting, sculpture, drama, dance, and movies. Among them, painting, sculpture, drama, dance, and movies can be expressed in the form of images. To increase inter-category variability, as shown in Fig. 1, we further divide these classes into twenty categories wherein humans frequently appear:

- 5 types of natural human scenes: Acrobatics, Cosplay, Dance, Drama, and Movies;
- 3 types of 3D artificial human scenes: Garage Kits, Relief, Sculpture;
- 12 types of 2D artificial human scenes: Kids Drawing, Mural, Oil Painting, Shadow Play, Sketch, Stained Glass, Ukiyoe, Cartoon, Digital Art, Ink Painting, Watercolor, and Generated Images;

Compared to previous art-related models [12, 18, 25] that directly borrow classification criteria from some websites such as wikiart without examination, our classification criteria is more suitable for human-centric scenes. On the one hand, it enables us to easily collect a larger amount of high-quality human-related images. On the other hand, the inter-category variability of Human-Art is significant to achieve diverse images with negligible classification errors.

3.2. Data Collection

We design a standardized pipeline for high-quality image collection, including manual image collection & classification, filtering, and consolidating, as shown in Fig. 2.

- Manual collection & classification: All images in Human-Art are either manually selected from 27 im-
age websites that provide high-resolution images, self-collected from offline exhibitions, or generated by the popular models (e.g., Stable Diffusion [48]) to ensure high quality. We have carefully examined a large number of possible image sources and finally determined to use 27 high-quality image websites such as ukiyo-e² and Sonia Halliday Photo Library³. To ensure intra-category diversity, each category in Human-Art comes from multiple websites. For search-based image websites that are not precisely aligned with our classification, we conduct searching with multiple keywords. We also add images from Google and Bing searches to further increase the diversity. Finally, we collect around one million well-classified images with the above procedure.

- **Filtering:** Despite the above efforts, the quality of images crawled from the Internet is not fully guaranteed, and a large portion of these images do not contain human beings. To tackle the above issues, we manually screen the collected images twice, with each image examined by at least two people, obtaining around 200k high-quality images that include humans. Next, we apply a variety of human detection and pose estimation algorithms to these images. The objective is to identify and remove those images containing characters that are too simple to improve model performance, blurred characters, and crowded characters that are hard to label. This filtering step finally yields 50k images.

- **Consolidating:** We further consolidate the whole dataset with multiple resolutions: 512×512, 256×192, 32×32, and the original resolution, for ease of use in various kinds of downstream tasks.

Finally, we split the Human-Art dataset into training, validation, and testing sets with a ratio of 70%, 10%, and 20%, resulting in 35k, 5k, and 10k images in each group.

![Figure 3. Illustration of the provided annotations including 2D keypoints, bounding box, self-contact point, and text description.](image)

### 3.3. Data Annotation

As shown in Fig. 3, Human-Art provides rich annotations⁴, including human bounding-box, 21 human keypoints (with corresponding visible/included/invisible attribute), self-contact keypoints, and text information.

We follow MSCOCO [27] to define the first 17 keypoints and add 4 additional keypoints, left/right fingers and left/right toes, which is beneficial for 3D pose estimation and shape recovery by providing more comprehensive constraints [43, 70]. Self-contact keypoints [4, 38] also demonstrate benefits for 3D pose and shape estimation by disambiguating the body part depth unknown in 2D human pose representation, avoiding self-collisions and penetrations, and being easier to use than ordinal depth [44, 52, 76]. These annotations are especially important for humans in artworks because they suffer from more severe pose distortion and imprecise body shape due to artistic exaggeration. Self-contact keypoints are annotated as the center of the body contact surface. Although ambiguity such as distorted body proportions/shapes and incomplete/blurry human bodies exist in artworks, human annotators are capable of inferring the positions based on intuitive knowledge. Text descriptions are automatically crawled from image websites, which usually include comparatively accurate text descriptions. For images that do not contain corresponding text descriptions, we use BLIP-2 [22] to generate fake labels.

The annotations are performed by a professional team with standardized annotation and audit procedures. Including 35 data annotators and 12 data auditors, the annotation team is systematically trained before starting annotations to ensure high annotation quality and timely feedback. As shown in Fig. 2 (b), the entire labeling process goes through two plenary quality checks and two random quality checks to ensure an accuracy of at least 98%.

### 3.4. Dataset Statistics and Analysis

![Figure 4. Illustration of the data distribution of the 20 categories in Human-Art and that of MSCOCO via a popular clustering method, the uniform manifold approximation and projection (UMAP) [34].](image)

---

²http://ukiyo-e.org/

³http://www.soniahalliday.com/index.html

⁴We do not annotate for the Generated Image category because many images in this category do not have legitimate human body parts, as elaborated in Section 4.2.
<table>
<thead>
<tr>
<th>Detector</th>
<th>Faster R-CNN</th>
<th>YOLOX</th>
<th>Deformable DETR</th>
<th>DINO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting</td>
<td>val</td>
<td>val*</td>
<td>test</td>
<td>test*</td>
</tr>
<tr>
<td>MSCOCO [27]</td>
<td>52.2</td>
<td>51.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cartoon</td>
<td>8.8</td>
<td>37.9</td>
<td>7.0</td>
<td>33.5</td>
</tr>
<tr>
<td>Digital Art</td>
<td>18.8</td>
<td>46.4</td>
<td>17.8</td>
<td>44.2</td>
</tr>
<tr>
<td>Ink Painting</td>
<td>11.0</td>
<td>37.7</td>
<td>9.1</td>
<td>37.2</td>
</tr>
<tr>
<td>Kids Drawing</td>
<td>6.6</td>
<td>54.2</td>
<td>8.0</td>
<td>53.6</td>
</tr>
<tr>
<td>Mural</td>
<td>9.7</td>
<td>35.5</td>
<td>9.3</td>
<td>34.5</td>
</tr>
<tr>
<td>Oil Painting</td>
<td>15.9</td>
<td>41.1</td>
<td>13.7</td>
<td>37.5</td>
</tr>
<tr>
<td>Shadow Play</td>
<td>7.5</td>
<td>64.1</td>
<td>8.2</td>
<td>63.7</td>
</tr>
<tr>
<td>Sketch</td>
<td>2.6</td>
<td>48.8</td>
<td>2.4</td>
<td>55.7</td>
</tr>
<tr>
<td>Stained Glass</td>
<td>8.8</td>
<td>35.0</td>
<td>8.1</td>
<td>34.7</td>
</tr>
<tr>
<td>Ukiyo</td>
<td>12.7</td>
<td>51.9</td>
<td>12.7</td>
<td>50.3</td>
</tr>
<tr>
<td>Watercolor</td>
<td>14.8</td>
<td>42.8</td>
<td>14.2</td>
<td>42.2</td>
</tr>
<tr>
<td>Garage Kits</td>
<td>22.9</td>
<td>60.0</td>
<td>22.5</td>
<td>62.5</td>
</tr>
<tr>
<td>Relief</td>
<td>4.9</td>
<td>37.5</td>
<td>4.7</td>
<td>33.4</td>
</tr>
<tr>
<td>Sculpture</td>
<td>17.7</td>
<td>48.6</td>
<td>14.4</td>
<td>47.0</td>
</tr>
<tr>
<td>Acrobatics</td>
<td>17.0</td>
<td>49.7</td>
<td>17.0</td>
<td>53.4</td>
</tr>
<tr>
<td>Cosplay</td>
<td>31.2</td>
<td>52.8</td>
<td>31.3</td>
<td>56.7</td>
</tr>
<tr>
<td>Dance</td>
<td>17.0</td>
<td>46.6</td>
<td>18.4</td>
<td>49.3</td>
</tr>
<tr>
<td>Drama</td>
<td>24.3</td>
<td>46.0</td>
<td>24.8</td>
<td>48.7</td>
</tr>
<tr>
<td>Movie</td>
<td>26.3</td>
<td>36.5</td>
<td>25.0</td>
<td>37.2</td>
</tr>
</tbody>
</table>

Average          | 12.0 | 44.2 | 12.5 | 43.0 | 14.4 | 14.7 | 11.7 | 11.7 | 12.6 | 12.7 |

* the baseline results we provide by training on the assembly of MSCOCO [27] and Human-Art.

Table 2. Comparisons of average precision (AP) of wildly used object detection models, including Faster R-CNN [47], YOLOX [46], Deformable DETR [77], and recent SOTA DINO [71]. The best results are shown in bold and the worst results are highlighted with underlined font. All models are trained on MSCOCO [27]. Detailed settings for each model are provided in supplementary files.

4.1. Human-Centric Recognition

4.1.1 Human Detection

Human detection task [27, 40] identifies the bounding box of each person in a given image, which is fundamental for further human scene understanding. It is also a crucial step for downstream tasks such as top-down human pose estimation [63, 64, 69]. Most object detectors (e.g., YOLO [46], DETR [7], and DINO [71]) do not differentiate humans from other objects in detection. Recently, HBA [40] is specifically designed for human and hand detection.

Tab. 2 shows the performance of both CNN-based and Transformer-based widely-used detectors on the validation and test sets of Human-Art. All the pre-trained models have poor performance on artificial scenes, with average precision (AP) ranging from 11.7% to 14.7%, confirming the impact of the domain gap on the models’ generalization ability. For some natural scenarios with similar distribution to the MSCOCO dataset [27], e.g., Dance and Acrobatics, existing models achieve satisfactory performance. In particular, the stage backdrop in Acrobatics is usually clean, resulting in a higher AP value compared to that of other categories. At the same time, despite Shadow Play also having
a spotless background, its performance is among the lowest because of the huge texture disparity from the natural scene.

Moreover, we provide a baseline model by training Faster R-CNN on the assembly of MSCOCO [27] and Human-Art. The joint training procedure leads to about a 56% performance boost in Shadow Play and a 31% average improvement in all categories. Nevertheless, the performance of the baseline model on Human-Art is still relatively low, calling for future research on this topic.

### 4.1.2 2D Human Pose Estimation

Human Pose Estimation (HPE) is another basic task for human motion analysis, which can be divided into 2D HPE and 3D HPE, outputting 2D keypoints and 3D keypoints respectively. Hard poses, heavy occlusions, and confusing backgrounds make these tasks still quite challenging after years of research. Existing 2D HPE methods can be categorized into three types: top-down [41, 58, 64], bottom-up [6, 10], and one-stage [66]. Generally speaking, top-down approaches [41, 58, 64] usually have higher performance than other methods, provided that human detection performs correctly. However, they suffer from high computational costs. In contrast, bottom-up methods [6, 10] are efficient, especially for crowded scenes, but have relatively low accuracy. To trade off efficiency and effectiveness, one-stage methods (e.g. PETR [53], ED-Pose [66]) are proposed thanks to the emergent DETR-based models [77]. We provide results for these representative methods in Tab. 3.

Specifically, we show the quantitative results for widely used as well as the SOTA pose estimation methods on the validation and testing sets of Human-Art. Top-down human estimation depends heavily on the accuracy of human detection, leading to performance elevation when the ground-truth bounding box is given, as shown in Fig. 7 (a). Different from the human detection task, pose complexity has a bigger impact on results than the image background. Although cosplay typically includes a complex image background, simple postures ease the estimation procedure and lead to high estimation accuracy. Shadow Play still shows a low estimation accuracy due to the large shape and texture differences from humans in natural scenes. Some pose failure cases are shown in Fig. 7 (b).

Moreover, we provide a baseline model by training HRNet on the assembly of MSCOCO and Human-Art, resulting in an overall 21% boost in accuracy.

<table>
<thead>
<tr>
<th>Detector</th>
<th>Faster R-CNN + HRNet</th>
<th>YOLOX + ViTPose</th>
<th>HigherHRNet</th>
<th>ED-Pose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Setting</strong></td>
<td><strong>val</strong></td>
<td><strong>val</strong></td>
<td><strong>test</strong></td>
<td><strong>test</strong></td>
</tr>
<tr>
<td>MSCOCO [27]</td>
<td><strong>75.6</strong></td>
<td>77.6</td>
<td>77.2</td>
<td><strong>73.4</strong></td>
</tr>
<tr>
<td>Cartoon</td>
<td>9.7</td>
<td>37.6</td>
<td>64.7</td>
<td>7.3</td>
</tr>
<tr>
<td>Digital Art</td>
<td>22.6</td>
<td>59.6</td>
<td>74.4</td>
<td>25.8</td>
</tr>
<tr>
<td>Ink Painting</td>
<td>6.3</td>
<td>51.4</td>
<td>72.1</td>
<td>5.6</td>
</tr>
<tr>
<td>Kids Drawing</td>
<td>10.5</td>
<td>40.8</td>
<td>86.1</td>
<td>10.0</td>
</tr>
<tr>
<td>Mural</td>
<td>11.6</td>
<td>54.0</td>
<td>71.1</td>
<td>12.3</td>
</tr>
<tr>
<td>Oil Painting</td>
<td>31.6</td>
<td>65.7</td>
<td>78.1</td>
<td>28.5</td>
</tr>
<tr>
<td>Shadow Play</td>
<td>5.4</td>
<td>15.9</td>
<td>59.8</td>
<td>5.0</td>
</tr>
<tr>
<td>Sketch</td>
<td>6.3</td>
<td>44.1</td>
<td>73.1</td>
<td>6.7</td>
</tr>
<tr>
<td>Stained Glass</td>
<td>10.4</td>
<td>46.0</td>
<td>74.8</td>
<td>9.7</td>
</tr>
<tr>
<td>Ukiyoe</td>
<td>17.8</td>
<td>48.1</td>
<td>82.4</td>
<td>18.8</td>
</tr>
<tr>
<td>Watercolor</td>
<td>26.7</td>
<td>60.1</td>
<td>73.9</td>
<td>25.5</td>
</tr>
<tr>
<td>Garage Kits</td>
<td>45.2</td>
<td>57.5</td>
<td>86.7</td>
<td>44.5</td>
</tr>
<tr>
<td>Relief</td>
<td>10.5</td>
<td>57.3</td>
<td>78.7</td>
<td>7.9</td>
</tr>
<tr>
<td>Sculpture</td>
<td>36.4</td>
<td>65.9</td>
<td>81.0</td>
<td>38.5</td>
</tr>
<tr>
<td>Acrobatics</td>
<td>45.8</td>
<td>68.0</td>
<td>85.2</td>
<td>46.6</td>
</tr>
<tr>
<td>Cosplay</td>
<td>71.0</td>
<td><strong>81.1</strong></td>
<td>87.2</td>
<td>72.6</td>
</tr>
<tr>
<td>Dance</td>
<td>43.1</td>
<td>67.3</td>
<td>77.2</td>
<td>49.2</td>
</tr>
<tr>
<td>Drama</td>
<td>45.3</td>
<td>75.1</td>
<td>82.0</td>
<td>46.7</td>
</tr>
<tr>
<td>Movie</td>
<td>49.5</td>
<td>71.5</td>
<td>77.2</td>
<td>50.4</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>22.2</td>
<td>55.2</td>
<td>76.4</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Table 3. Comparisons of average precision (AP) of wildly used human pose estimation models, including top-down methods HRNet and ViTPose [58, 64], bottom-up method HigherHRNet [10], and one-stage method ED-Pose [66]. The best results are shown in **bold** and the worst results are highlighted with underlined font. Detailed settings for each model are provided in supplementary files.

...
Figure 5. Illustration of how the annotated self-contact points can benefit 3D human mesh recovery. (a), (c), and (e) show the human mesh outputs from three scenes without self-contact optimization. (b), (d), and (f) are optimized mesh results with self-contact points.

Figure 6. Failure cases of existing popular text-to-image diffusion models on human-centric generation. We highlight these cases by blue, yellow, and red arrows for the missing, redundant, and replaced body parts. We simply regard the natural human structure as more desirable, despite the fact that there is no right or wrong in art.

Figure 7. Failure cases of the pose estimator HRNet on Human-Art. The first column presents how human detection impacts top-down pose estimation. Red lines and points represent ground truth, lines and points in other color are detected results. w/o shows pose estimation results based on detected boxes, w means results with the ground truth bounding box. The right figures show (i) perspective, (ii) pose, (iii) shape, and (iv) texture in Human-Art are challenging to existing pose estimators (trained on MSCOCO).

4.1.3 Human Mesh Recovery

Statistical body models such as SMPL [31] show their convenient usage for animation, games, and VR applications. These models represent humans in a watertight and animatable 3D human body mesh with a small number of parameters, which largely simplifies 3D human mesh expression. However, depth ambiguities hinder the fidelity of 3D human mesh estimation from a monocular camera. To overcome this issue, similar to Sketch2Pose [4], we further provide self-contact annotations as additional information to facilitate reasonable depth optimization via the interpenetration penalty. By mapping the contact region onto the vertices of a rough SMPL model generated by Exemplar Fine-Tuning (EFT) [16] and then minimizing the distance among the contact vertices, visualization results in Fig. 5 show how annotating self-contact keypoints benefit 3D mesh recovery.

4.2. Image Generation

Image generation has experienced great advances in the past few years with eye-catching generative vision and language models such as unCLIP [45], Latent Diffusion [48] and Imagen [49]. Such progress is enabled by recent breakthroughs of generative models, as well as large-scale web-crawled datasets such as LAION [51]. However, despite their impressive capability of generating photorealistic and creative images, they often fail at faithfully respecting human structures, as shown in Fig. 6.

By offering fine-annotated artificial scenarios with Human-Art, we not only introduce a valuable supplement to existing datasets. More importantly, it provides a good prior for generating art images with plausible human poses. We show some examples produced by fine-tuning a diffusion model on Human-Art in Fig. 8. Note that Human-Art augments the art category Shadow Play, which is absent from SOTA generative models such as Stable Diffusion.
4.3. Motion Transfer

The motion transfer task aims to generate a new image or video of the source person by learning motion from target images while preserving the source character’s appearance.

Previous motion transfer methods can be roughly divided into two categories. Model-based methods [8, 32, 50] use off-the-shelf pose estimators to extract pose information and then use pose skeletons to drive the character. In contrast, model-free methods [55, 56, 59] can automatically detect character-agnostic implicit keypoint trajectories to transfer for arbitrary objects.

As shown in Fig. 9 (e), model-free approaches [56] easily fail on new scenes because of the unstable correspondence between source and driving images. Model-based methods show more stable performance on human motion transfer, but they highly rely on accurate pose estimation results. Present pose estimators are unsuitable for artificial scenes such as kids’ drawings, thus requiring training a pose estimation model specific to the scene. To illustrate how pose estimators with multi-scenario adaptability can benefit motion transfer tasks, we conduct experiments on the well-known EveryBodyDanceNow [8] without face enhancement. Although there are model-based models that generate better results than EveryBodyDanceNow, we choose this model to illustrate how poses influence motion transfer because it is widely accepted in the literature. Fig. 9 (c) shows the original motion transfer result with pose estimator OpenPose [5] trained on natural human scenes. By refining the pose estimator with Human-Art, Fig. 9 (d) shows how a better pose detection model greatly benefits the motion transfer task.

5. Conclusion and Discussions

In this paper, we have presented Human-Art, a rich-scenario human-centric dataset containing 50k high-quality images with versatile manual annotations, which serves as a new challenging dataset for multiple computer vision tasks, such as human detection, human pose estimation, body mesh recovery, motion transfer, and image generation. In our experiments, we provide comprehensive baseline results and detailed analyses for these tasks. We hope that this work will shed light on related research areas and open up new research questions.

Limitations and Future work: Images in Human-Art could be misused for generating fake images, which may bring negative social impact. Moreover, although downstream tasks have been extensively explored on Human-Art, we simply conduct experiments to reveal how and why existing methods often fail on our dataset, but did not offer a superior solution. Therefore, there is a significant gap to fill for these rich-scenario human-centric tasks, calling for novel solutions in future research. Specifically, future directions with Human-Art include but are not limited to 1) cross-domain human recognition algorithms that can adapt to different scenes with various human poses, shapes, textures, and image backgrounds; 2) trustworthy image generation with reasonable human body structure, especially controllable human image generation such as GLI-GEN [24] and ControlNet [72]; 3) Inclusive motion transfer algorithms across different scenes.

We plan to continuously expand Human-Art to support new scenarios. To facilitate future human-centric studies, we will make the training and validation set public with an easy-to-use data visualization platform. For the test set, we will provide a testing interface but withhold the data to prevent test information leakage.

6. Acknowledgements

This work was supported in part by the Shenzhen-Hong Kong-Macau Science and Technology Program (Category C) of the Shenzhen Science Technology and Innovation Commission under Grant No. SGDX2020110309500101 and in part by Research Matching Grant CSE-7-2022.
References


[9] Shuhong Chen and Matthias Zwicker. Improving the perceptual quality of 2d animation interpolation. In European Conference on Computer Vision (ECCV), 2022. 1


[17] David Kadin, Sebastian Risi, and Anders Sundnes Løvlie. Improving object detection in art images using only style transfer. In International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2021. 2


[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence


[33] Prathmesh Madhu, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas Maier, and Vincent Christlein. Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. Journal on Computing and Cultural Heritage (JOCCH), Nov. 2022. 1, 2, 3


[36] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Ac-


