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

Xuan Ju[1,2*†], Ailing Zeng[1*‡], Jianan Wang[1], Qiang Xu[2], Lei Zhang[1]
[1]International Digital Economy Academy, [2]The 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 scenes. We take advantage of it and introduce the **Human-Art** dataset to bridge related tasks in natural and artificial scenarios. Specifically, Human-Art contains* 50*k high-quality images with over* 123*k 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 on paper to computer-generated virtual figures 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

[*]Equal contribution.
[†]Work in part done during an internship at IDEA.
[‡]Corresponding author.

Figure 1. *Human-Art* is a versatile human-centric dataset to bridge the gap between natural and artificial scenes. It includes 20 high-quality scenes, including natural and artificial humans in both 2D representation (yellow dashed boxes) and 3D representation (blue solid boxes).

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., [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 $1,000$k carefully collected images using standardized data collection, filtering, and consolidating processes.
- **Versatile annotations**: *Human-Art* provides carefully manual annotations 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 *wikiart*[1] 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,

[1] https://www.wikiart.org/

Dataset	Image	Instance	Keypoint Number	Bbox	Pose	Self-Contact	Natural Scenario	Artificial Scenario
VOC2012[1] [11]	8,174	17,132	-	√	-	-	√	-
MSCOCO[1] [27]	66,808	273,469	-	√	-	-	√	-
BodyHands [39]	20,490	63,095	-	√	-	-	√	-
People-Art [61]	1,490	3,870	-	√	-	-	√	√
MSCOCO[2] [27]	58,945	156,165	17	√	√	-	√	-
MPII [3]	24,920	40,522	16	√	√	-	√	-
AI Challenger [62]	240,000	448,776	14	√	√	-	√	-
CrowdPose [23]	20,000	~80,000	14	√	√	-	√	-
OCHuman [73]	4,731	8,110	17	√	√	-	√	-
PoseTrack[3] [2]	66,374	153,615	15	√	√	-	√	-
HiEve[3] [28]	49,820	1,099,357	14	√	√	-	√	-
ClassArch [33]	1513	1728	17	√	√	-	-	√
Sketch2Pose [4]	808	14,772	18	√	√	√	-	√
Human-Art **(Ours)**	50,000	123,131	21	√	√	√	√	√

[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

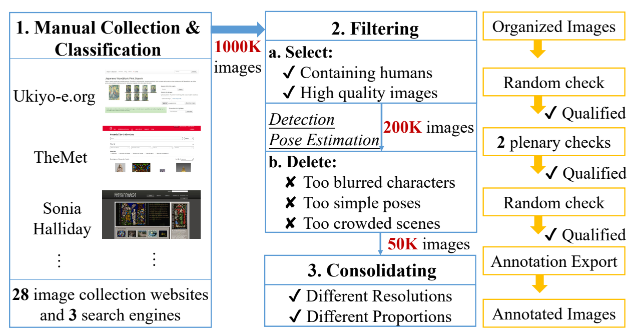
3.1. The Hierarchical Category Classification

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 Movie;
- 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;



(a) Data collection processes (b) Annotation

Figure 2. Data collection and annotation processes.

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 image 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-e[2] and Sonia Halliday Photo Library[3]. 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.

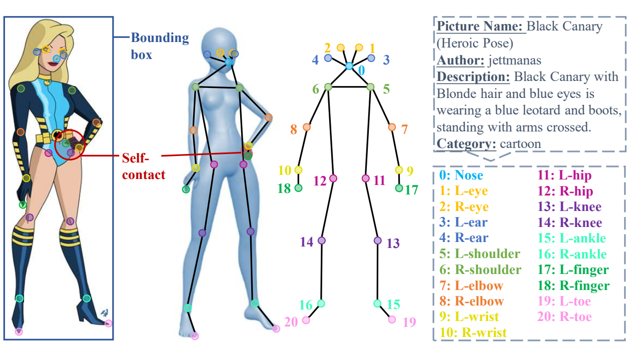


Figure 3. Illustration of the provided annotations including 2D keypoints, bounding box, self-contact point, and text description.

3.3. Data Annotation

As shown in Fig. 3, *Human-Art* provides rich annotations[4], 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 ensures 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

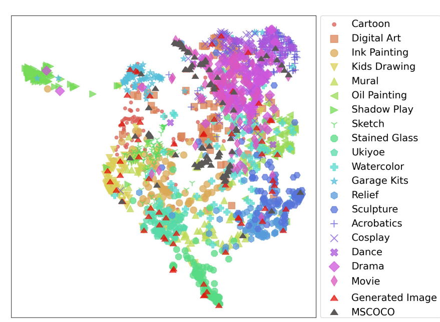


Figure 4. Illustration of the data distribution of the 20 categories in *Human-Art* and that of MSCOCO via a popular clustering method, the uniform manifold approximation and projection (UMAP) [34].

[2] https://ukiyo-e.org/
[3] http://www.soniahalliday.com/index.html

[4] We 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.

		Detector	Faster R-CNN				YOLOX		Deformable DETR		DINO	
		Setting	val	val *	test	test *	val	test	val	test	val	test
		MSCOCO [27]	**52.2**	51.6	-	-	**61.9**	-	**57.2**	-	**63.2**	-
Artificial Scene	2D Representation	Cartoon	8.8	37.9	7.0	<u>33.5</u>	10.8	9.2	7.9	6.7	8.7	8.1
		Digital Art	18.8	46.4	17.8	44.2	24.1	22.9	17.6	15.5	18.6	18.1
		Ink Painting	11.0	37.7	9.1	37.2	15.5	13.0	11.9	10.0	14.5	11.6
		Kids Drawing	6.6	54.2	8.0	53.6	6.8	11.5	5.6	7.2	6.8	8.2
		Mural	9.7	35.5	9.3	34.5	12.2	12.2	9.3	8.1	10.2	9.5
		Oil Painting	15.9	41.1	13.7	37.5	20.8	18.3	17.1	14.2	17.0	15.0
		Shadow Play	7.5	**64.1**	8.2	**63.7**	5.4	7.5	5.3	5.1	6.4	7.9
		Sketch	<u>2.6</u>	48.8	<u>2.4</u>	55.7	<u>4.6</u>	<u>5.2</u>	5.8	9.2	<u>3.6</u>	7.1
		Stained Glass	8.8	<u>35.0</u>	8.1	34.7	8.2	7.8	5.1	<u>4.6</u>	7.8	7.8
		Ukiyoe	12.7	51.9	12.7	50.3	13.1	12.8	8.5	8.4	11.4	11.4
		Watercolor	14.8	42.8	14.2	42.2	19.7	18.2	15.6	13.6	15.4	14.3
	3D Representation	Garage Kits	22.9	60.0	22.5	62.5	22.3	19.9	17.9	14.6	22.8	19.5
		Relief	4.9	37.5	4.7	33.4	8.4	9.1	<u>4.7</u>	5.7	4.4	<u>5.9</u>
		Sculpture	17.7	48.6	14.4	47.0	15.8	13.2	9.4	7.1	10.1	8.5
Natural Scene		Acrobatics	17.0	49.7	17.0	53.4	20.0	19.4	17.3	17.6	19.4	18.9
		Cosplay	31.2	52.8	**31.3**	56.7	38.0	**37.2**	34.6	**34.5**	37.2	**36.7**
		Dance	17.0	46.6	18.4	49.3	20.3	21.1	17.8	18.5	19.3	19.6
		Drama	24.3	46.0	24.8	48.7	27.4	27.5	15.4	25.8	27.8	16.7
		Movie	26.3	36.5	25.0	37.2	28.0	26.8	26.6	26.2	27.2	26.3
		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 <u>underlined font</u>. All models are trained on MSCOCO [27]. Detailed settings for each model are provided in supplementary files.

To demonstrate the inter-category variability and intra-category diversity of *Human-Art*, we randomly select 100 images from each category and also 100 images from the MSCOCO dataset. Then we use ResNet152 [13] to extract image features and UMAP [34] to reduce these features down to 2 dimensions for visualization. As shown in Fig. 4, most scenarios form their own clusters, and distinct types of images (e.g., Drama and Sculptures) lead to dramatically different distributions that are far apart. The distribution of MSCOCO is similar to the natural image categories in *Human-Art* and is included in the 20 categories' distribution. Moreover, the generated images are scattered across the distribution as expected.

Moreover, we calculate the invisible keypoints of *Human-Art*, and the results demonstrate a higher valid keypoint percentage than the MSCOCO dataset. More dataset statistics and analysis are in the supplementary file.

4. Experiments

We conduct comprehensive experiments on *Human-Art* with several popular human-centric tasks. In each subsection, we introduce the task and related methods first. Then, we present the corresponding results and analyses.

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

		Detector	Faster R-CNN + HRNet						YOLOX + ViTPose				HigherHRNet		ED-Pose	
		Setting	**val**	**val‡**	**val *‡**	**test**	**test ‡**	**test ***	**val**	**val ‡**	**test**	**test ‡**	**val**	**test**	**val**	**test**
		MSCOCO [27]	**75.6**	77.6	77.2	**73.4**	-	-	79.8	82.3	81.1	-	68.6	70.5	71.6	69.8
Artificial Scene	2D Representation	Cartoon	9.7	37.6	64.7	7.3	34.4	61.0	16.3	55.1	13.0	50.5	15.7	12.0	22.2	18.0
		Digital Art	22.6	59.6	74.4	25.8	61.2	75.7	29.0	69.9	31.9	72.2	42.5	44.3	43.5	45.6
		Ink Painting	6.3	51.4	72.1	5.6	48.0	72.4	8.9	59.8	9.2	58.2	26.8	20.9	28.2	24.9
		Kids Drawing	10.5	40.8	86.1	10.0	44.6	85.9	14.0	59.2	13.2	62.6	12.6	13.8	20.7	23.2
		Mural	11.6	54.0	71.1	12.3	53.6	71.6	15.9	50.1	16.4	51.6	30.6	32.0	34.6	35.1
		Oil Painting	31.6	65.7	78.1	28.5	62.2	75.6	39.6	73.4	36.7	70.5	54.4	51.1	56.2	51.7
		Shadow Play	<u>5.4</u>	<u>15.9</u>	<u>59.8</u>	<u>5.0</u>	<u>17.2</u>	<u>58.4</u>	<u>8.1</u>	<u>29.2</u>	<u>8.4</u>	<u>26.0</u>	<u>4.4</u>	6.5	<u>6.0</u>	<u>7.7</u>
		Sketch	6.3	44.1	73.1	6.7	57.2	79.4	9.1	61.3	10.9	71.3	13.6	<u>6.3</u>	12.0	12.2
		Stained Glass	10.4	46.0	74.8	9.7	45.1	73.1	12.0	59.1	12.1	58.1	26.6	23.1	27.6	25.6
		Ukiyoe	17.8	48.1	82.4	18.8	47.7	81.8	23.8	61.2	26.8	63.1	20.2	19.4	25.0	25.8
		Watercolor	26.7	60.1	73.9	25.5	57.6	73.4	36.4	71.0	36.1	69.0	48.9	43.4	50.6	45.6
	3D Representation	Garage Kits	45.2	57.5	86.7	44.5	61.4	**89.2**	52.5	76.2	50.6	77.0	37.4	34.7	47.9	44.1
		Relief	10.5	57.3	78.7	7.9	53.4	76.0	16.2	70.8	14.9	67.1	32.5	29.8	28.0	27.1
		Sculpture	36.4	65.9	81.0	38.5	64.0	78.5	34.9	78.5	34.2	73.7	33.5	35.2	45.9	46.7
Natural Scene		Acrobatics	45.8	68.0	85.2	46.6	68.4	83.2	69.1	86.8	66.3	83.9	58.6	57.4	41.4	44.4
		Cosplay	71.0	**81.1**	**87.2**	72.6	**81.9**	87.0	**80.0**	**90.3**	**81.7**	**88.8**	**78.1**	**77.8**	**79.6**	**79.7**
		Dance	43.1	67.3	77.2	49.2	70.1	80.1	57.3	81.5	61.5	83.8	51.4	62.4	53.6	61.2
		Drama	45.3	75.1	82.0	46.7	75.8	83.1	54.2	83.9	56.9	84.8	69.6	72.2	75.0	76.0
		Movie	49.5	71.5	77.2	50.4	72.2	76.2	57.6	76.8	56.5	78.6	64.9	65.8	69.2	68.2
		Average	22.2	55.2	76.4	24.1	55.4	76.0	28.7	67.7	30.7	67.5	34.6	36.3	37.5	39.2

‡ the top-down pose estimation results that use ground truth bounding box;
* the baseline results we provide by training on the assembly of MSCOCO [27] and *Human-Art*.

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 <u>underlined font</u>. Detailed settings for each model are provided in supplementary files.

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.

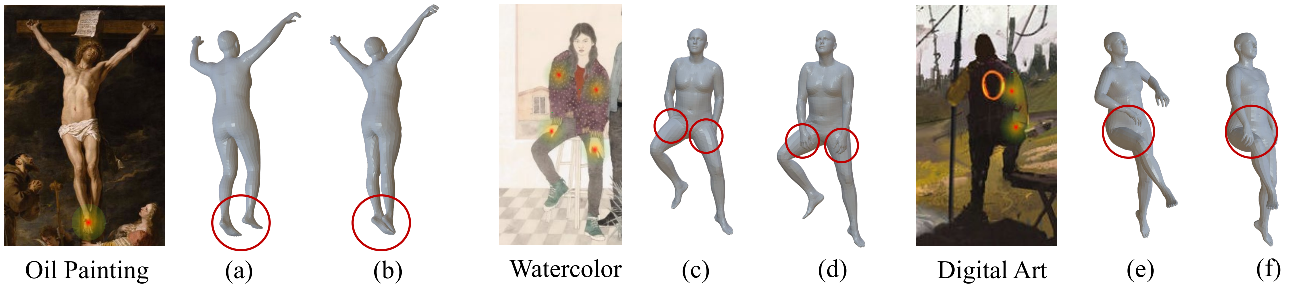


Figure 5. Illustration of how the annotated self-contact points can benefit 3D human mesh recovery. (a), (c), and (e) show the human mesh outputs from three scenes without self-contact optimization. (b), (d), and (f) are optimized mesh results with self-contact points.

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.

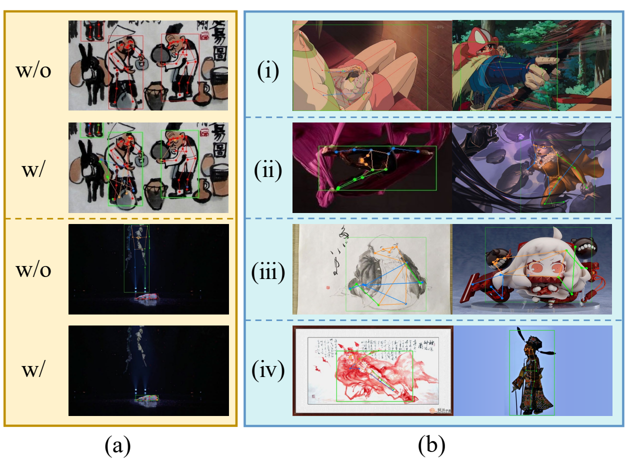


Figure 7. Failure cases of the pose estimator HRNet on *Human-Art*. The first column presents how human detection impacts top-down pose estimation. Red lines and points represent ground truth, lines and points in other color are detected results. *w/o* shows pose estimation results based on detected 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.

Figure 8. Example generations with five scenes from a diffusion generative model trained on *Human-Art*. Notably, Shadow Play is a novel scene for existing generative models.

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.

Figure 9. Visualization of model-based and model-free motion transfer results of intra-scene transfer (the 1st row) and inter-scene motion transfer (the 2nd and 3rd rows). The model-free method FOMM [56] severely fails due to unstable correspondence. In contrast, the model-based methods [8] are more suitable for multi-scene motion transfer. After using the pose estimation model pretrained on *Human-Art*, (d) shows better results than (c),

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[5], 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] https://sketch.metademolab.com/

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 GLIGEN [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

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020. 1

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5176, 2018. 2, 3

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 2, 3

[4] Kirill Brodt and Mikhail Bessmeltsev. Sketch2Pose: Estimating a 3d character pose from a bitmap sketch. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 1, 2, 3, 4, 7

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, 2019. 8

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 6

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 5

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5933–5942, 2019. 8

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

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5386–5395, 2020. 6

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 2, 3

[12] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1

[15] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[16] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 7

[17] David Kadish, 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

[18] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 3

[19] Pramook Khungurn and Derek Chou. Pose estimation of anime/manga characters: A case for synthetic data. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, pages 1–6, 2016. 2, 3

[20] Kangyeol Kim, Sunghyun Park, Jaeseong Lee, Sunghyo Chung, Junsoo Lee, and Jaegul Choo. AnimeCeleb: Large-scale animation celebheads dataset for head reenactment. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[21] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11025–11034, 2021. 2

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[23] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10863–10872, 2019. 1, 2, 3

[24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 8

[25] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The ArtBench dataset: Benchmarking generative models with artworks. 2022. 3

[26] Mei Kuan Lim, Ven Jyn Kok, Chen Change Loy, and Chee Seng Chan. Crowd saliency detection via global similarity structure. In *International Conference on Pattern Recognition (ICPR)*, pages 3957–3962. IEEE, 2014. 2

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

Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 1, 2, 3, 4, 5, 6

[28] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 2, 3

[29] Zuzeng Lin, Ailin Huang, Zhewei Huang, Chen Hu, and Shuchang Zhou. Collaborative neural rendering using anime character sheets. *arXiv preprint arXiv:2207.05378*, 2022. 1

[30] Wu Liu and Tao Mei. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys (CSUR)*, Mar. 2022. 1

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248:1–248:16, Oct. 2015. 7

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 8

[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

[34] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 4, 5

[35] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5084–5093, 2020. 1

[36] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2308–2317, 2022. 2

[37] Lucas Mourot, Ludovic Hoyet, François Le Clerc, François Schnitzler, and Pierre Hellier. A survey on deep learning for skeleton-based human animation. In *Computer Graphics Forum*, volume 41, pages 122–157. Wiley Online Library, 2022. 1

[38] Lea Muller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, June 2021. 4

[39] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4889–4899, 2022. 1, 2, 3

[40] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016. 6

[42] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[44] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7307–7316, 2018. 4

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 7

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 5

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 28. Curran Associates, Inc., 2015. 5

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 4, 7

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 7

[50] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision (ECCV)*, pages 596–613. Springer, 2020. 8

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 7

[52] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2325–2334, 2019. 4

[53] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11069–11078, 2022. 6

[54] Larry Shiner. *The invention of art: A cultural history*. University of Chicago press, 2003. 3

[55] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2377–2386, 2019. 8

[56] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems (NIPS)*, 32, 2019. 8

[57] Vishwanath Sindagi, Rajeev Yasarla, and Vishal MM Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020. 2

[58] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 6

[59] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3637–3646, 2022. 1, 8

[60] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2017. 2

[61] Nicholas Westlake, Hongping Cai, and Peter Hall. Detecting people in artwork with CNNs. In *European Conference on Computer Vision (ECCV)*, pages 825–841. Springer, 2016. 1, 2, 3

[62] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages 1480–1485. IEEE, 2019. 2, 3

[63] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, pages 466–481, 2018. 5

[64] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2022. 1, 2, 5, 6

[65] Zhewei Xu, Jiajun Zhuang, Qiong Liu, Jingkai Zhou, and Shaowu Peng. Benchmarking a large-scale fir dataset for on-road pedestrian detection. *Infrared Physics & Technology*, 96:199–208, 2019. 2

[66] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *International Conference on Learning Representations*, 2023. 6

[67] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. VToonify: Controllable high-resolution portrait video style transfer. *arXiv preprint arXiv:2209.11224*, 2022. 1

[68] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. TransMoMo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5306–5315, 2020. 1

[69] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. DeciWatch: A simple baseline for 10x efficient 2d and 3d pose estimation. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 5

[70] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. SmoothNet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 4

[71] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1, 5

[72] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 8

[73] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2Seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 889–898, 2019. 2, 3

[74] Qingyuan Zheng, Zhuoru Li, and Adam Bargteil. Learning to shadow hand-drawn sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7436–7445, 2020. 1

[75] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In *ACM International Conference on Multimedia (MM)*, pages 2264–2272, 2020. 2

[76] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2344–2353, 2019. 4

[77] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 5, 6