—— Supplementary Materials ——
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

Overview

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- More dataset statistics and analysis in Sec. A, including statistics and analysis of image source, keypoint visible/occlude/invisible attributes distribution, human size distribution, and annotation visualization.
- Experimental details in Sec. B, including implementation details of models used in our experiments, analysis of human detection and pose estimation results, and more evaluation results on *Human-Art*.
- More discussion about related datasets for multi-scenario generalization in Sec. C.

A. Dataset Statistic and Analysis

A.1. Image Sources

Human-Art is a comprehensive human-centric dataset with 50,000 images from 20 distinctive scenarios, where each scenario contains 2,500 images. As shown in the right part of Fig. 1, the images are selected from a total of 30 different image sources and we guarantee diversity of image sources for each scenario. Specifically, *Human-Art* include images collected from European, North American, East Asia, and South African authors, ranging from Before the Common Era to the 21st century with humans in different poses, shapes, and textures.

A.2. Keypoints Attribute

Human-Art follows MSCOCO [9] to annotate keypoints with visible/occlude/invisible attributes. The left part of Fig. 1 shows the percentage of the total visible, occluded, and invisible keypoints in all the annotated scenarios of *Human-Art* compared to MSCOCO. As can be observed, the invisible keypoints of the MSCOCO dataset reach a percentage of 63.2%, which is much higher than that of all the categories in *Human-Art*. We attribute it to the fact that MSCOCO does not only focus on human-centric scenes, despite the fact that it contains more than 250,000 humans. This resulted in a large percentage of small-scale, incomplete, and fuzzy humans, which can only be annotated using bounding boxes. At the same time, the percentage of visible keypoints of many categories in *Human-Art* is slightly low. This is because, on the one hand, artistic natural scenarios usually contain elaborate movements and fabric coverings, which obstruct the human body; on the other hand, images in artificial scenarios may have unclear lines or missing body pieces lost in history. Overall, *Human-Art* has a higher percentage of keypoint annotation than COCO, which can benefit the related tasks with more valid data.

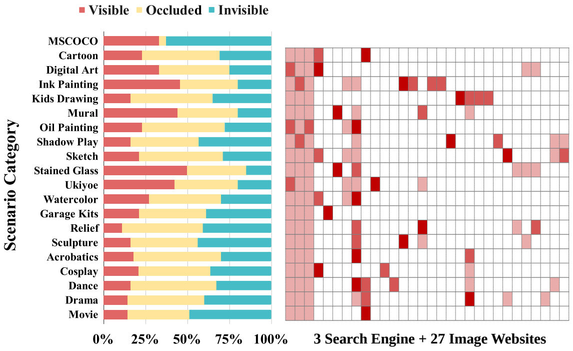


Figure 1. Statistical analyses on the visibility for all keypoints comparing our 20 scenes with MSCOCO [9] (left) and the distribution map of image sources of *Human-Art* (right).

A.3. Human Size Distribution

As shown in Fig. 2, human sizes in *Human-Art* are more evenly distributed than MSCOCO [9]. The average height of humans in *Human-Art* is 0.40 times the image's height, whereas in MSCOCO is 0.28 times, which shows

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

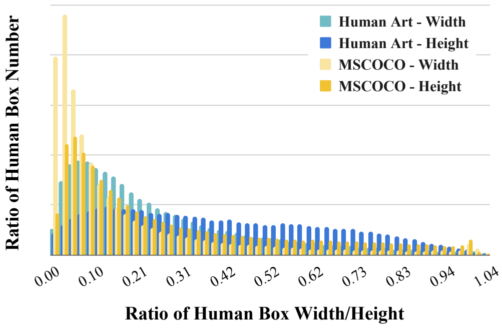


Figure 2. Distribution of $bounding\ box\ width/image\ width$ and $bounding\ box\ height/image\ height$. The horizontal axis shows the ratio of a human bounding box's height and width to the entire image. The vertical axis shows the percentage of human bounding boxes with the corresponding height and width ratio.

that *Human-Art* has fewer tiny humans than MSCOCO due to its human-centric image collecting process. The average ratio for human width is 0.15 and 0.25 in *Human-Art* and MSCOCO, respectively. For the reason that human beings are usually long and thin, the width ratio of humans is more concentrated in small proportions. A more balanced distribution of human sizes enables *Human-Art* to support downstream tasks required for various-sized humans. For example, motion transfer usually needs bigger and more detailed human figures to output characters with higher fidelity. However, image generation needs to generate humans in a variety of resolutions to satisfy the user's requirements. More interestingly, despite our relatively large as well as balanced human size, the poor performance from detection and pose estimation suggests that our dataset still offers difficulties beyond scale, such as appearance diversity, background variation, and pose complexity.

A.4. Annotation Visualization

We show the annotation quality and image diversity of *Human-Art* with the human bounding-box and keypoint annotation in Fig. 3. The diversity of *Human-Art* derives from the wide variations in painting techniques among categories as well as the variations in the human size, body shape, and character pose within each category. This makes *Human-Art* a more challenging dataset than previous real-world human datasets, necessitating higher generalization abilities from the detection and estimation model.

B. Exprimental Details

B.1. Implementation Details

For human detection, we provide baselines of Faster R-CNN [13], YOLOX [12], Deformable DETR [25], and DINO [23]. All the pretrained models we use are trained exclusively on MSCOCO [9]. And the training is implemented on the random shuffle of MSCOCO and *Human-Art*. The implementation details are explained as follows:

- Faster R-CNN: We choose Faster R-CNN on Feature Pyramid Networks [8] with ResNet-50 [4] as the backbone. For testing, the pertained model we use is trained on 8 NVIDIA GTX 1080 Ti GPUs for 12 epochs. For training, we trained on 8 NVIDIA RTX 3090Ti GPUs. Given that data volume has almost doubled, we trained 21 epochs rather than the original 12 epochs to guarantee model convergence.
- YOLOX: We choose YOLOX-L with an input size of 640x640. For testing, the pertained model we use is trained on 8 NVIDIA Tesla PG503-216 GPUs for 300 epochs. For training, we trained on 4 Nvidia Tesla A100 GPUs.
- Deformable DETR: We choose Two-Stage Deformable DETR with ResNet-50 [4] as the backbone. For testing, the pertained model we use is trained on 8 Nvidia Tesla v100 GPUs for 50 epochs. For training, we trained on 4 Nvidia Tesla A100 GPUs.
- DINO: We choose DINO-5scale with Swin-L [11] as the backbone. For testing, the pertained model we use is trained on Nvidia Tesla A100 GPU for 31 epochs.

For human pose estimation, we provide baselines of HRNet [14], ViTPose [20], HigherHRNet [2], and ED-Pose [21]. All the pretrained models we use are trained exclusively on MSCOCO [9]. The training is implemented on the random shuffle of MSCOCO and *Human-Art*. For top-down pose estimation methods, we use human detectors with settings the same as listed above in testing, and use augmentations of the ground truth bounding box (e.g., random flip, random bounding box center shift) in the testing and training stage. To make a fair comparison with COCO, testing and training only consider 17 human keypoints. The implementation details are explained as follows:

- HRNet: We choose HRNet-W48 with an input size of 256x192. For testing, the pertained model we use is trained on 8 Nvidia Tesla v100 GPUs for 210 epochs. For training, we trained on 4 Nvidia Tesla v100 GPUs.
- ViTPose: We choose ViTPose-H with an input size of 256x192. For testing, the pertained model we use is trained on 8 Nvidia Tesla v100 GPUs for 210 epochs.
- HigherHRNet: We choose HigherHRNet-W48 with an input size of 512x512. For testing, the pertained model we use is trained on 8 Nvidia Tesla v100 GPUs for 300 epochs. For training, we trained on 4 Nvidia Tesla A100 GPUs.
- ED-Pose: We choose ED-Pose with ResNet-50 [4] as the backbone. For testing, the pertained model we use is trained for 60 epochs.

For human mesh recovery, we use the same optimization strategies as in Sketch2Pose [1] with 17 human keypoints and self-contact keypoint. The 2D-to-SMPL model used in

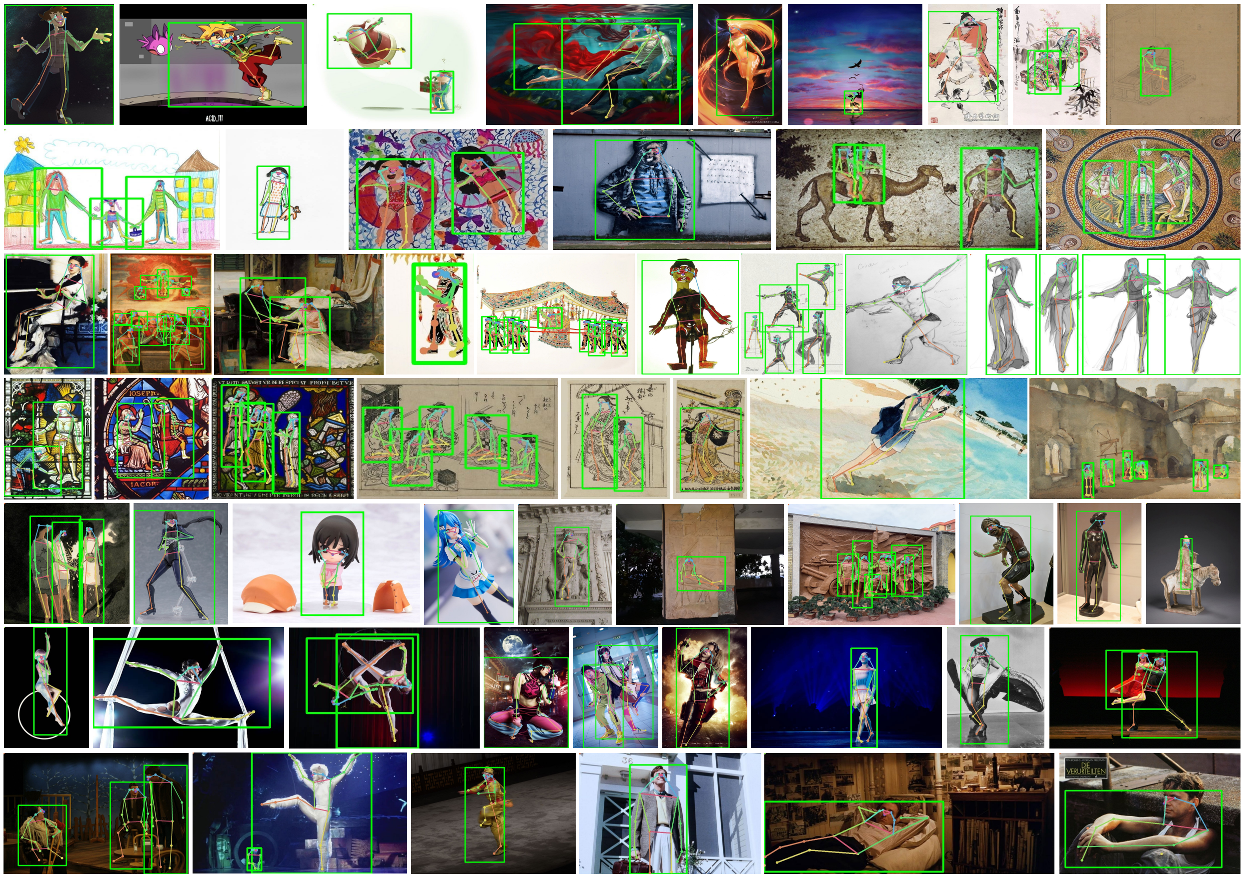


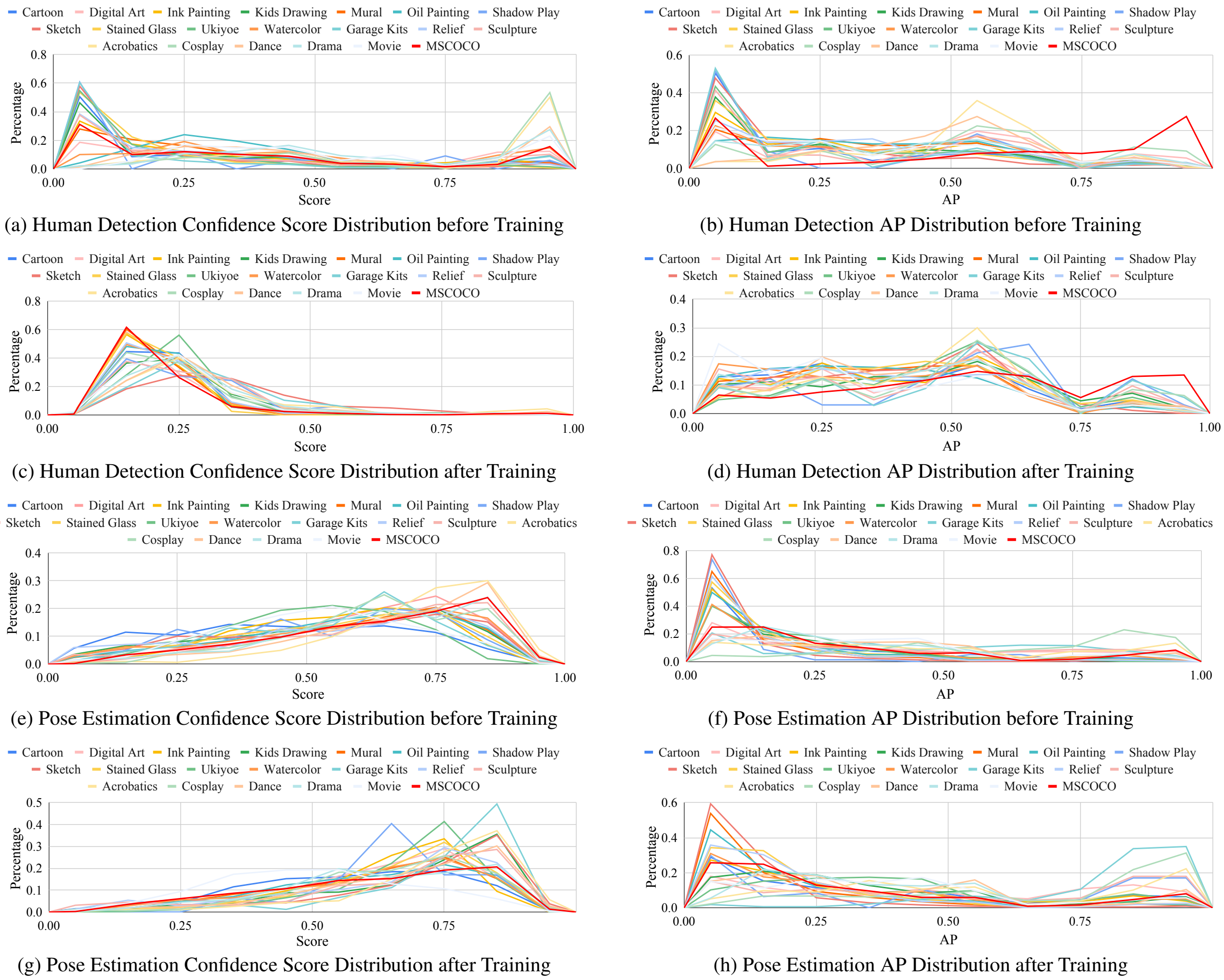
Figure 3. Annotated examples in *Human-Art*. We randomly select 3 images from each category to show the annotation quality and image diversity of *Human-Art*. Images in *Human-Art* are varied in terms of human shape, pose, texture and size.

Sketch2Pose [1] includes optimization of 2D bone tangents, body part contacts, and bone foreshortening. In the main paper, we contrast the visualization results of not using and using body part contacts in Fig. 6. Results show using self-contact keypoints benefits 3D mesh recovery by minimizing the 3D distance near the self-contact area. Noted that due to the influence of the other two optimization elements, bone tangents and bone foreshortening, body parts that do not directly connect to the self-contact area show different poses with and without the self-contact optimization (e.g. the elbow angle in Fig. 6 (a) and (b) in the main paper).

B.2. More Analyses of Human-centric Tasks

The confidence scores output from pose estimation or human detection model indicate how confident the model is for output results. The AP scores indicate the models' average prediction accuracy. We try to analyze why these current models underperform on our data based on the two metrics. Fig. 4 shows the variation of human detection model YOLOX [12] and pose estimation model HRNet [14]'s confidence score and AP distribution before and after training. We use the same trained model as in main paper.

We provide analyses from the following three aspects: (1) The contrast of confidence score and AP distribution. We discover that the model tends to be over-confident, where the confidence score distribution and AP distribution do not show a positive correlation. This issue has become more serious in artificial scenes of *Human-Art*. For instance, in Fig. 4 (e) and Fig. 4 (f), although the pose estimation model shows a relatively high confidence score in most images, a large proportion of the estimation outputs' AP scores range from 0 to 0.25. (2) The contrast between before training and after training. The recurring finding is that training can reduce the percentage of both low confidence scores and low AP scores, which is consistent with common sense. Another interesting finding is that, although the mean AP score on MSCOCO [9] is reduced after joint training, the percentage of the low scores is reduced as well, as shown in Fig. 4(b) and Fig. 4(d). This may be because the more evenly distributed human size and the richer depictions in *Human-Art* help the model to obtain better adaptability on hard poses in real-world scenarios. (3) The contrast between human detection and pose estimation tasks. After training, human detection shows a more uniform AP



(a) Human Detection Confidence Score Distribution before Training

(b) Human Detection AP Distribution before Training

(c) Human Detection Confidence Score Distribution after Training

(d) Human Detection AP Distribution after Training

(e) Pose Estimation Confidence Score Distribution before Training

(f) Pose Estimation AP Distribution before Training

(g) Pose Estimation Confidence Score Distribution after Training

(h) Pose Estimation AP Distribution after Training

Figure 4. Contrast of confidence score and AP distribution of human detection model YOLOX [12] and pose estimation model HRNet [14] before and after training on our proposed scenes and COCO. Specifically, (a)-(d) shows the distribution of pose estimation, and (e)-(h) shows the distribution of human detection. The horizontal axis of each figure shows the confidence score/AP intervals. The vertical axis of each figure shows the image percentage in each score/AP interval.

distribution along the horizontal axis. However, the pose estimation model shows concentrated distributions in low and high AP scores. This may be due to the differences in the two methods' targets. When human detection fails, valid interactions between ground truth and detected results may still exist. But caused by the interaction across different keypoints, pose estimation typically fails more severely.

B.3. Cross Dataset Results

Due to the page limit, we put cross-dataset experimental evaluation on *Human-Art* and Sketch2Pose in Table 1. MSCOCO, Sketch2Pose, and *Human-Art* have different keypoint definitions, thus we give out results on the 10/17 intersected keypoints of the three datasets (the three datasets have 10 intersected keypoints. MSCOCO and *Human-Art* have 17 intersected keypoints). Four training setting is shown in the table, where (1) HA(T) means using *Human-Art* for training, (2) HA(F) means using *Human-Art* to finetune the model pre-trained on MSCOCO, (3) SK(T) means using Sketch2Pose for training, (4) SK(F) means using Sketch2Pose to finetune the model pre-trained

Test Set	**HA(T)**	**HA(F)**	**SK(T)**	**SK(F)**
MSCOCO	63.5‡/ 62.4*	74.5‡/74.8*	13.5*	47.6*
Human-Art	65.7 / 63.6*	64.4 / 61.9*	14.3*	43.3*
Sketch2Pose	71.6*	70.5*	16.3*	68.9*

[1] **Training Strategies**: **HA** for *Human-Art* . **SK** for Sketch2Pose. **T** for training from scratch. **F** for fine-tuning from the HRNet pretrained on MSCOCO.

[2] * / ‡ means calculating on the 10 / 17 intersected keypoints of different datasets (*Human-Art* & Sketch2Pose / *Human-Art* & MSCOCO).

Table 1. AP results of pose estimation model HRNet on 3 test sets (the first column) under 4 training strategies (the first row). The best results on 10 intersected keypoints are shown in Red.

on MSCOCO. Results show that Sketch2Pose is not sufficient for the training of multi-scenario pose estimator and thus shows poor results. Both training and fine-tuning with *Human-Art* can lead to a relatively satisfactory accuracy, where fine-tuning with *Human-Art* has the highest AP in the natural scenario, and training with *Human-Art* has the best results on *Human-Art* and Sketch2Pose. Considering the commonness of natural humans in daily life, the recommended usage of *Human-Art* is still combining training *Human-Art* with MSCOCO.

C. Datasets for Multi-Scenario Generalization

Existing datasets [3, 5, 7, 15–18] that include multi-scenario are more often used in domain generalization. They focus on object classification or object detection tasks. Related methods try adapting classifiers or detectors from natural to artificial images [19, 22]. However, as demonstrated in Table 2, several limitations of these datasets make them hard to bridge natural and artificial human-centric tasks. First, the number of images and categories in these datasets is insufficient. Second, the number of downstream tasks they can support is constrained by the fact that they only have bounding box or object category annotations. Third, these datasets contain only a small percentage of scenes with humans and are not applicable to human-centric tasks. Besides, BAM! [5] is a large-scale dataset with 7 artificial categories targeted at image classification, but it uses untrustworthy model classifiers to label images instead of manually labeling, which may result in a lot of labeling mistakes.

Task	**Dataset**	**Image**	**Natural Scenario**	**Artificial Scenario**
Object Classification / Detection	Inoue N. et al. [5]	5,000	1	3
	Office-Home [15]	15,500	2	2
	PACS [7]	9,991	1	3
	People-Art [16]	1,490	1	42 *
	Photo-Art [18]	5,375	1	1
	***Human-Art* (Ours)**	**50,000**	**4**	**16**

* The 42 painting styles of People-Art [16] have different classification criteria from *Human-Art*, and these styles are encapsulated within the proposed 20 categories of *Human-Art*.

Table 2. Comparison of multi-scenario datasets that serve for general object classification and detection tasks.

By contrast, *Human-Art* is a full-scenario human-centric dataset inclusive of domain generalization tasks of both previous multi-scenario datasets such as human detection domain generalization, and other tasks such as human pose estimation, image generation, and image style transfer.

Previous methods solve domain gap problems of object detection by transferring knowledge from the source domain to the target domain. [16] fine-tune Faster R-CNN on People-Art to detect humans in artworks. H2FA R-CNN [19] proposes a Holistic and Hierarchical Feature Alignment R-CNN to enforce image-level alignment for object detection. [5] use image-level domain transfer and pseudo-labels from the source domain to train object detector SSD300 [10].

Previous works [6, 24] have explored domain generalization and adaptation for human keypoint detection in the natural scenario. However, to the best of our knowledge, no previous works involve multi-scenario human keypoint detection in both natural and artificial scenes.

In a word, no suitable domain adaptation and domain generalization method in the literature can be directly applied to *Human-Art* and we leave it to future work.

References

[1] 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. 2, 3

[2] 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. 2

[3] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 5

[4] 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. 2

[5] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5009, 2018. 5

[6] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2021. 5

[7] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5542–5550, 2017. 5

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 2

[9] 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

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 5

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[12] 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. 2, 3, 4

[13] 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. 2

[14] 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. 2, 3, 4

[15] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5018–5027, 2017. 5

[16] 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. 5

[17] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. BAM! the behance artistic media dataset for recognition beyond photography. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2017. 5

[18] Qi Wu, Hongping Cai, and Peter Hall. Learning graphs to model visual objects across different depictive styles. In *European Conference on Computer Vision (ECCV)*, pages 313–328. Springer, 2014. 5

[19] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2FA R-CNN: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14329–14339, 2022. 5

[20] 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. 2

[21] 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. 2

[22] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7097–7107, June 2022. 5

[23] 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. 2

[24] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 5

[25] 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. 2