Domain-Adaptive 2D Human Pose Estimation via Dual Teachers in Extremely Low-Light Conditions

Yihao Ai[1⋆], Yifei Qi[1⋆], Bo Wang[2,3], Yu Cheng[1], Xinchao Wang[1], and Robby T. Tan[1]

[1] Department of Electrical and Computer Engineering, National University of Singapore
[2] Department of Computer and Information Science, University of Mississippi
[3] CtrsVision
{yihao, e0315880}@u.nus.edu, hawk.rsrch@gmail.com, e0321276@u.nus.edu, {xinchao, robby.tan}@nus.edu.sg

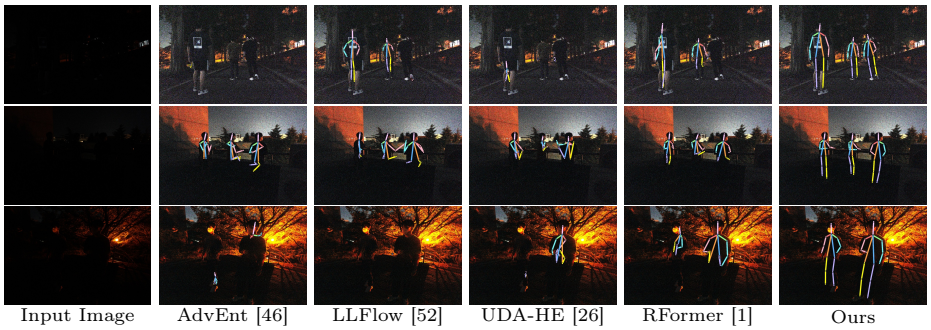


Fig. 1: Qualitative comparison involving two image enhancement methods [1, 52], two domain adaptation methods [26, 46], and our proposed method on ExLPose-OCN [29]. The human pose results for two image enhancement methods are obtained by initially enhancing the input images from low-light to well-lit and subsequently applying an off-the-shelf human pose estimation method [14] to predict the poses. The first column shows the original low-light images, which serve as input for both the baselines and our method. The images in the subsequent columns have been brightened solely for visualization purposes and are not utilized by any of the methods.

Abstract. Existing 2D human pose estimation research predominantly concentrates on well-lit scenarios, with limited exploration of poor lighting conditions, which are a prevalent aspect of daily life. Recent studies on low-light pose estimation require the use of paired well-lit and low-light images with ground truths for training, which are impractical due to the inherent challenges associated with annotation on low-light images. To this end, we introduce a novel approach that eliminates the need for low-light ground truths. Our primary novelty lies in leveraging two complementary-teacher networks to generate more reliable pseudo labels, enabling our model achieves competitive performance on extremely low-light images without the need for training with low-light ground truths.

⋆ Equal contribution.

Our framework consists of two stages. In the first stage, our model is trained on well-lit data with low-light augmentations. In the second stage, we propose a dual-teacher framework to utilize the unlabeled low-light data, where a center-based main teacher produces the pseudo labels for relatively visible cases, while a keypoints-based complementary teacher focuses on producing the pseudo labels for the missed persons of the main teacher. With the pseudo labels from both teachers, we propose a person-specific low-light augmentation to challenge a student model in training to outperform the teachers. Experimental results on real low-light dataset (ExLPose-OCN) show, our method achieves **6.8%** (**2.4** AP) improvement over the state-of-the-art (SOTA) method, despite no low-light ground-truth data is used in our approach, in contrast to the SOTA method. Our code is available at DA-LLPose.

Keywords: Human pose estimation · Low-light · Domain adaptation

1 Introduction

Human 2D pose estimation is a fundamental task in computer vision and important for many downstream tasks like human activity recognition [59], virtual tryon [12], motion capture [11], augmented reality [55], or Metaverse [21]. The existing methods and benchmarks [31,34,60] primarily focus on well-illuminated scenarios, where human subjects exhibit both clear visibility and high recognizability in the majority of cases. However, poor lighting conditions such as low light or nighttime are a prevalent aspect of daily life, which signifies the importance of low-light human pose estimation research. Despite the significance of this research area, its exploration has been constrained by the lack of suitable low-light datasets or benchmarks.

A previous dataset for human pose estimation under poor lighting conditions primarily revolves around nighttime scenarios [9, 10], where human subjects generally remain recognizable due to the presence of light sources such as lamps or car lights. Thus, these nighttime datasets do not focus on extremely low-light conditions, which pose severe challenges to existing human pose estimation methods. Recently, Lee et al. [29] proposed a new benchmark for human pose estimation in extremely low-light conditions, which covers different levels of low-light conditions, and a low-light pose estimation method, which requires paired well-lit and low-light images with ground truths for training. However, the scarcity of paired images with ground truths in practical scenarios makes the application of this method impractical. To this end, we aim to develop an innovative domain adaptive method for low-light human pose estimation, utilizing well-lit ground-truth data only. Our goal is to achieve performance on par with the SOTA methods that rely on both low-light and well-lit ground truths.

To the best of our knowledge, we are the first to propose a solution to estimate human poses in extremely low-light conditions utilizing well-lit ground-truth data only. This is an extremely challenging task because the visibility of human subjects is poor and low-light images are corrupted due to severe low-light noise. The intensity and contrast are much lower in the low-light images

(e.g., average pixel intensity of well-lit and low-light images are 90.5 and 2.0). Without paired well-lit and low-light images with ground-truth data, two types of existing methods could be used for low-light pose estimation: low-light image enhancement [1, 52] and domain adaptation [26, 46]. However, these two groups of methods have their own limitations. The former is difficult to restore the extremely low-light images faithfully and the latter still lacks a specific solution for bridging the gap between well-lit and low-light conditions.

In response to these challenges, we introduce an innovative domain-adaptive dual-teacher framework for human pose estimation that facilitates knowledge transfer from well-lit to low-light domains. This innovation enables our model to exceed state-of-the-art (SOTA) methods by eliminating the requirement for training with low-light ground truths, which are often impractical to obtain. Ideally, cross-domain knowledge transfer is achievable using a single teacher, as demonstrated by existing methods [25, 26, 44]. However, a single center-based teacher, designed to predict the entire human pose based on each person's center, may encounter difficulties in detecting individuals in low-light images due to inherent architectural limitations that focus solely on the detection of human centers. Consequently, this approach may fail to identify individuals when their centers cannot be detected, resulting in a limited number of pseudo labels. To generate a more comprehensive set of human pose predictions for use as pseudo labels, we propose the incorporation of a complementary teacher: a keypoint-based bottom-up human pose estimation model. This model is capable of predicting partial human poses by grouping detected keypoints, offering a complementary approach to the primary center-based teacher.

Our method comprises two stages: pre-training and dual-teacher knowledge acquisition. During the first stage, we perform supervised training with labeled well-lit data alone. This initial step equips both teachers to recognize human poses in well-lit images and their corresponding low-light augmentations. In the subsequent second stage, our focus shifts to harnessing knowledge directly from unlabeled low-light images. To challenge the student model during the second stage, we introduce a novel augmentation technique called Person-specific Degradation Augmentation (PDA). PDA is selectively applied to images with pseudo labels generated by the teachers. Given that low-light conditions pose a significant challenge in pose estimation, our degradation augmentations primarily darken the individuals, making them less distinguishable from their backgrounds to mimic realistic low-light scenarios. Such a simple, flexible and effective degradation neither necessitates the camera calibration as various physical-based augmentations did [53, 54], nor relies on an additional network, which would cost extra computational resources [39]. And it facilitates the student model's exploration of knowledge in the low-light domain, thereby encouraging it to surpass the performance of the teachers in the second stage.

Fig. 1 shows three images from ExLPose-OCN [29] and the human pose estimation results of the existing methods [1,26,46,52] and our proposed method. Compared with the existing methods, our method can effectively estimate human poses in extremely low-light conditions. Specifically, our method is capable of

handling low-contrast (last row) and partially visible (second row) persons. For the relatively visible persons (first row), our method can produce more accurate pose estimation. In summary, our major contributions are listed as follows.

- We investigate a challenging yet practical task, i.e., estimating human poses in extremely low-light conditions with well-lit ground-truth data only. To the best of our knowledge, we are the first to propose a solution in this challenging setting.
- We introduce a novel domain-adaptive dual-teacher framework, utilizing both center-based and keypoint-based teachers to generate enriched pseudo labels for effective student model training, eliminating the need for low-light ground truths, which are often impractical to obtain. To our knowledge, this represents a significant advancement in low-light human pose estimation.
- Our proposed method outperforms the SOTA method by **6.8%** (**2.4** AP) on the real low-light dataset, ExlPose-OCN. Notably, our method utilizes well-lit ground-truth data alone, while the SOTA method is trained with paired well-lit and low-light images with ground truths.

2 Related Work

2D Human Pose Estimation Two primary paradigms exist within 2D human pose estimation: top-down and bottom-up approaches. In general, top-down methods, benefiting from applying human detection first [5, 17, 43, 45, 56], tend to outperform their bottom-up counterparts [3, 6, 27, 37]. Early approaches of bottom-up estimation [3,37] are keypoint-based, detecting keypoints for each individual in an image before grouping them to construct individual human poses. Recently, Geng et al. propose a center-based bottom-up method [14], which predicts human centers instead of grouping individual keypoints, and regresses offsets between a human center and the keypoints for each individual. This method sets a new trend among the bottom-up methods [47,48,58]. However, existing 2D pose estimation methods focus on well-lit conditions, primarily due to the prevalence of well-lit images within the benchmark datasets [31, 34, 60]. Additionally, since existing methods are fully supervised, they demonstrate poor performance when applied directly to low-light images after being trained on well-lit data.

Domain Adaptive Pose Estimation Domain adaptive pose estimation is potentially useful for low-light conditions. A number of domain adaptation methods are proposed in human pose [26,38,41], animal pose [2,30,36] and hand pose estimations [16, 22, 23]. However, these methods are not designed for human pose estimation in extremely low-light conditions. Source-free domain adaptive human pose methods are developed [38, 41] to transfer knowledge from synthetic to real data in the absence of the source data. This is in opposition to low-light scenarios, where well-lit (source) domain data is typically available. Kim et al. propose a unified framework for domain adaptive human pose estimation [26], where the student is challenged to learn the target-domain knowledge by applying affine transform-based augmentations. However, this augmentation is not tailored for the severe lighting conditions.

Low-light Human Pose Estimation Apart from the domain adaptation-based methods, fully-supervised approaches are developed as well. Crescitelli et al. propose a nighttime human pose estimation method that combines images from both RGB and infrared cameras [9, 10]. However, the reliance on infrared cameras restricts the practical applicability of this method. Recently, Lee et al. [29] propose a top-down human pose estimation approach for extremely low-light conditions. Nevertheless, this method depends on the use of paired well-lit and low-light images with ground-truth data during training, which are impractical for real-world applications due to the scarcity of the paired data and the inherent challenges associated with annotating low-light images. In the context of semi-supervised learning for human pose estimation, the utilization of multiple teachers, especially in dual-teacher configurations [19, 57], is related to our work. However, semi-supervised learning typically leverages both labeled and unlabeled data under the assumption that they come from similar distributions. This assumption differs significantly from the low-light setting, underscoring the necessity for our innovative domain-adaptive dual-teacher framework.

Low-light Image Enhancement Low-light image enhancement might be used to enhance low-light images prior to human pose estimation. Early approaches primarily rely on gamma correction [40] or histogram equalization [4,7]. More advanced methods restore the low-light images based on the Retinex theory [32,51]. Unfortunately, these methods tend to introduce additional noise with color distortion. In the deep learning era, a number of convolutional neural network (CNN)-based [35, 42, 50] and GAN-based [20, 24] methods are developed. Recently, Wang et al. introduce an invertible network to capture the distribution of the well-lit images [52]. Cai et al. propose a transformer-based approach to restore the hidden corruption in low-light images [1]. However, these image enhancement methods still face difficulties in faithfully restoring extremely low-light images. Thus, even with enhanced images of these methods, low-light pose estimation is still challenging.

3 Proposed Method

Given an image $I \in \mathbb{R}^{3\times h\times w}$ which contains persons in the low-light conditions, 2D human pose estimation is to obtain the human skeletons from an image by locating K types of human keypoints for each person. An overview of our proposed framework is shown in Fig. 2, where we present a novel domain-adaptive dual-teacher framework, utilizing both center-based and keypoint-based teachers to generate enriched pseudo labels for effective student model training, as shown in the yellow boxes.

3.1 Main Teacher and Complementary Teacher

Main Teacher We build our main teacher based on DKER [14], where a person's pose is represented as the center and offset in addition to the heatmap for each keypoint. In particular, the center location of the i-th person is $\mathbf{c}^i = \frac{1}{K}\sum_{k=1}^{K}\mathbf{p}_k^i$

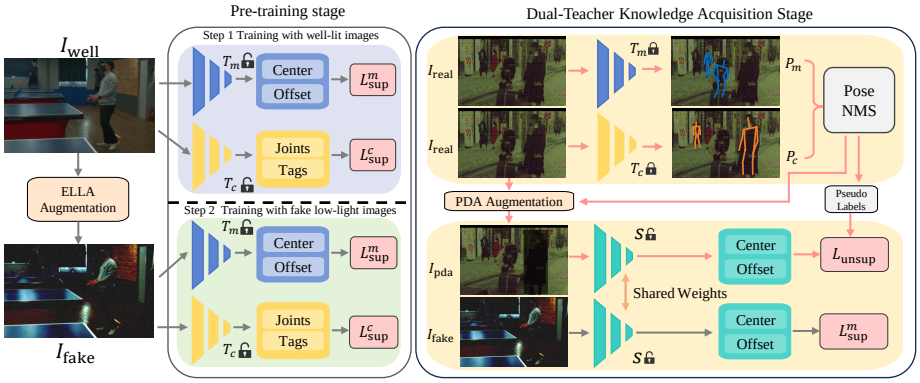


Fig. 2: The overview of our method. The grey uni-directional arrows indicate the supervised workflow in the pre-training stage. The grey and pink arrow indicates the workflow of the supervised and unsupervised losses in the knowledge acquisition stage. The low-light images, I_{fake}, I_{real} and I_{pda} are all brightened for improved visibility. NMS stands for non-maximum suppression.

and the offsets from the human center to each keypoints are $\mathbf{o}^i = \{\mathbf{p}^i_1 - \mathbf{c}^i, \mathbf{p}^i_2 - \mathbf{c}^i, ..., \mathbf{p}^i_k - \mathbf{c}^i\}$, where $\mathbf{p}^i_k \in \mathbb{R}^2$ indicates the 2D location of the k-th keypoints for the i-th person. The model is supervised by both heatmap and offset losses:

$$L^m_{\text{sup}} = L^m_H + \lambda_m L^m_O, \tag{1}$$

where L^m_H denotes a MSE loss of the main teacher's heatmaps, and L^m_O is a smooth L1 loss of the offset map.

Complementary Teacher The main drawback of the center-based representation is that the pose estimation fully relies on the center's predictions. If the center of a person is failed to be detected, the whole person is missed, especially for the low-contrast and partially visible persons in the low-light conditions. Therefore, we build a complementary teacher to predict human poses for partially visible persons following HigherHRNet-style design [6]. Specifically, the network is trained to estimate the keypoints' heatmap $\hat{H}_c$ and the corresponding tag map $\hat{T}_c$. A person's pose is grouped based on all the identity information via Hungarian algorithm [28]. The method is supervised by following losses:

$$L^c_{\text{sup}} = L^c_H + \lambda_c L^c_{\text{tag}}, \tag{2}$$

where, L^c_H is a MSE loss of the complementary teacher's heatmaps, and L^c_{tag} consists of push and pull losses of the tag maps. More details are provided in the supplementary.

3.2 Pre-Training Stage

Firstly, the two teachers are trained on the well-lit dataset [29] to gain the common knowledge of the human bodies. They are supervised by the losses

defined in Eq. (1) and Eq. (2) respectively. After the supervised training with the labeled well-lit data, teachers can predict human poses on well-lit images, but they cannot process low-light images due to the substantial domain gap between well-lit and low-light conditions. As our framework relies solely on well-lit ground-truth data, we propose Extreme Low-Light Augmentation (ELLA) to augment low-light images from well-lit ones. This process involves incorporating a set of low-light characteristics into the well-lit images.

The distinguishing characteristics of low-light images include extreme darkness, high levels of noise, and low contrast. To simulate these low-light specific features, ELLA applies a series of random augmentations sequentially to well-lit images, including gamma correction, brightness adjustment, contrast reduction, and Gaussian noise. The first two augmentations aim to intensify the darkness of the images, while contrast reduction is employed to decrease image contrast, and Gaussian noise is added to introduce noise into the images. It is important to note that in ELLA, we do not explicitly simulate light sources. This is because our primary focus is on low-light conditions, where no light source is typically available. An example of the augmented images using ELLA is shown in Fig. 3. The following equations, Eq. (3), Eq. (4), Eq. (5), and Eq. (6), mathematically represent the four ELLA augmentations:

$$I_{\text{out}} = 255 \cdot (I_{in}/255)^{\gamma}, \quad (3) \qquad I_{\text{out}} = cI_{\text{in}} + (1-c)I_{\text{grey}}, \quad (5)$$

$$I_{\text{out}} = bI_{\text{in}}, \quad (4) \qquad I_{\text{out}} = I_{\text{in}} + \text{Gaussian}(0, var), \quad (6)$$

where I_{in} is the input image, I_{grey} is the greyscale image of I_{in}, and I_{out} stands for the output image. $\gamma \in [2, 5]$, $b \in [0.01, 0.05]$, $c \in [0.2, 1.0]$ and $var \in [0, 40]$ are the parameters uniformly sampled from their respective ranges.

The augmented well-lit images are named fake low-light images, denoted as I_{fake}. The augmentation is added randomly so that both teachers are trained with a mixture of well-lit images and fake low-light images and supervised by the losses in Eq. (1) and Eq. (2) respectively (each of the ELLA augmentations has a 0.5 probability to be added). It is important to use both well-lit images and fake low-light images in the training because it is difficult for the model to transfer the knowledge to low-light domain if all inputs suddenly change to the fake low-light images.

3.3 Dual-Teacher Knowledge Acquisition Stage

After the pre-training stage, the teachers possess the preliminary capability to perform human pose estimation for extremely low-light images. In this stage, our primary objective is to enhance the student model's ability for detecting and localizing human keypoints by leveraging real low-light images and guiding the student to surpass both teachers. In particular, given a non-paired fake low-light image I_{fake}, a real low-light image I_{real}, and the pre-trained two teachers T_m, T_c, we train the student S, who shares the same architecture of the main teacher at the beginning, by a supervised loss and an unsupervised loss. The



(a) Gamma **(b)** Brightness **(c)** Contrast **(d)** Noise **(e)** All Aug. **(f)** Bri. All

Fig. 3: Examples of our ELLA. (a)-(d) are the results of the individual augmentation in ELLA. (e) is the image with all the augmentations, and Aug. stands for augmentations. (e) is brightened up for improved visibility as shown in (f), where Bri. All denotes brightened images with all the augmentations.

workflow of the supervised and unsupervised losses is illustrated in Fig. 2 with grey and pink arrows respectively.

Supervised Loss The supervised loss is a continuation of the pre-training stage. To strengthen the student's knowledge on the low-light domain, a larger proportion of fake low-light images is expected to be added in this stage. Ideally, all well-lit images that are used in the student's training should be augmented in the low-light fashion by ELLA. However, this would trigger the knowledge forgetting problem [33, 49] which is an inherent problem in domain adaptation methods. To strengthen the learning in the target domain while keeping the source knowledge, we make a few adjustment to the ELLA as follows.

Adjust ELLA On the one hand, we set ELLA to do gamma correction and brightness adjustment as Eq. (4) and Eq. (3) to every input well-lit images. This aims to augment more samples in the low-light domain. On the other hand, to avoid the forgetting problem, we randomly crop a few image patches in an augmented image and restore it to the well-lit domain. The cropping probability for an image is 0.15 to guarantee the majority of the inputs are still fully-augmented images. The supervised loss is computed by the following equations (shown in Fig. 2 with grey arrows). First, well-lit images I_{well} are augmented to fake low-light images I_{fake} with adjusted ELLA as:

$$I_{\text{fake}} = \text{AdjustELLA}(I_{\text{well}}). \tag{7}$$

Second, student S produces the prediction for I_{fake}:

$$\hat{H}_f, \hat{O}_f = S(I_{\text{fake}}), \tag{8}$$

where $\hat{H}$ and $\hat{O}$ refer to the heatmaps and offset maps, which are used in the calculation of the supervised loss following Eq. (2).

Unsupervised Loss Apart from the supervised loss, we teach the student with the pseudo labels selected and fused by the predictions from both teachers. As shown in the Fig. 2, the two teachers firstly generate their own pose predictions P_m and P_c by $P_m = T_m(I_{\text{real}})$ and $P_c = T_c(I_{\text{real}})$. Since the two teachers score the poses in different manner (the main teacher directly uses the center's activation,

C_m, while the complementary teacher averages the heatmap activation of all keypoints as the confidence score, C_c), we use different thresholds s_m and s_c to select out the high-confidence poses. Then the poses from both teachers are concatenated by and applied non-maximum suppression (NMS) [18] to rule out the redundant poses predictions to obtain the pseudo labels P_{all} defined as:

$$P_{\text{all}} = \text{NMS}(P_m[C_m > s_m] \oplus P_c[C_c > s_c]), \tag{9}$$

where $\oplus$ indicates concatenation operation.

Person-specific Degradation Augmentation The student needs to gain low-light domain-specific knowledge in order to outperform the teachers in challenging scenarios under low illumination. For example, the right person with black shirt in the last row of Fig. 1 is largely indistinguishable from the dark background, which demonstrates that low contrast is a major challenge for low-light human pose estimation. Therefore, the network has to learn the specific human structure in a texture-less condition. To this end, we propose a Person-specific Degradation Augmentation (PDA) to realistically mimic the extremely low-light conditions at individual-level. Specifically, after getting the selected poses P_{all} from the teachers, a set of bounding boxes is generated to crop the individuals. The cropped image patches are augmented based on Eq. (4), which are then used to replace the original ones from input I_{real} to generate output I_{pda}.

The last step of the unsupervised loss computation is to generate the pseudo labels based on P_{all} and to supervise the output of the student:

$$\hat{H}_r, \hat{O}_r = S(\text{PDA}(I_{\text{real}}, P_{\text{all}})). \tag{10}$$

The unsupervised loss $L^m_{\text{unsup}} = L^m_H + \lambda_m L^m_O$ is similar to Eq. (1). Combining both supervised and unsupervised losses, the loss of the student S is defined as follows:

$$Ls = \lambda_{\text{sup}} L^m_{\text{sup}} + \lambda_{\text{unsup}} L_{\text{unsup}}. \tag{11}$$

4 Experiments

4.1 Implementations and Dataset

Dataset ExLPose datasets [29] are used in validating our method, which is a new dataset that is specifically designed for evaluating 2D human pose estimation under extremely low-light conditions. The training set of ExLPose provides $2,065$ pairs of well-lit and low-light images and the ground-truth definition follows the CrowdPose [31] format. These images were taken from 251 indoor and outdoor scenes in the daytime and the paired low-light images are artificially produced by using a dual-camera system with different settings to ensure a wide coverage of diverse low-light conditions.

There are two different testing sets in ExLPose [29]. One is ExLPose-OCN , in which 360 extremely low-light images are captured at night with two different

Table 1: Evaluation on ExLPose-OCN with top-down baselines (lower part of the table) and the bottom-up baselines (upper part of the table) tested in our paper. Our method is a bottom-up method. The reported top-down baselines are all fully-supervised methods. LL indicates low-light labels. WL indicates well-lit labels. The best is **bold**. The second best is underlined.

Methods	Training Labels		$AP^{\uparrow}$@0.5:0.95		
	LL	WL	A7M3	RICOH3	Avg.
Base-low [14]	✓		27.1	15.9	21.5
Base-well [14]		✓	5.3	7.4	6.3
RFormer [1]		✓	18.9	17.7	18.3
LLFlow [52]		✓	23.7	19.1	21.4
UDA-HE [26]		✓	6.5	7.2	6.9
AdvEnt [46]		✓	9.1	11.2	10.1
CPN-low [5]		✓	23.7	23.9	23.8
CPN-well [5]	✓		15.2	15.6	15.4
CPN-all [5]	✓	✓	32.8	31.7	32.2
LLFlow [52]	✓	✓	25.6	28.2	27.0
LIME [15]	✓	✓	33.2	28.4	30.7
DANN [13]	✓	✓	27.9	30.6	29.3
AdvEnt [46]	✓	✓	28.2	29.0	28.6
LSBN+LUPI [29]	✓	✓	35.3	35.1	35.2
Ours		✓	**39.1**	**36.2**	**37.6**

cameras, A7M3 and RICOH3. Another one is ExLPose-test, which consists of 491 man-made low-light images captured and produced similarly as those in the training set. Low-light All (LL-A) is used to denote the whole ExLPose-test, which is further divided into three subsets according to their difficulty level including Low-Light Normal (LL-N), Hard (LL-H), and Extreme (LL-E).

Among the two testing sets, ExLPose-OCN is captured in real low-light environments, while the low-light images in the ExLPose-test are obtained by artificially reducing the amount of light by 100 times [29]. In addition, ExLPose-OCN is a test-only dataset, which is used to evaluate the generalization capability of a model trained on ExLPose-test. Therefore, the evaluation on ExLPose-OCN provides a more objective measurement of the performance of low-light human pose estimation. In our method, we only use the unpaired labeled well-lit images and unlabeled low-light images in the training set to train our models. Low-light ground-truth data is never used in training our model.

Evaluation Protocol We follow the COCO protocol [34], as used in LSBN+LUPI [29], to validate on ExLPose-test and ExLPose-OCN. Average precision at various thresholds (i.e., $AP@0.5:0.95$) is reported for every subset.

Implementation Details DEKR-W32 [14] is adopted as our backbone network for the main teacher and the student. HigherHRNet [6] is modified to be the complementary teacher. In the pre-training stage, the input image of our framework is cropped to 512×512. The data augmentations follow the DEKR paper setting, which include random rotation $[-30°, 30°]$, random scaling $[0.75, 1.5]$, random translation $[-40, 40]$ and random flipping. The training strategy of both teachers

is the same with DEKR. The only difference is the loss weighting factors, the $\lambda_m = 0.03$ and $\lambda_c = 0.001$.

In the knowledge acquisition stage, the confidence score thresholds are set to be $s_m = 0.9$ and $s_c = 0.5$ respectively for dual teachers. In training, we use Adam optimizer. The learning rate is set as $1e{-}4$. λ_{sup} and λ_{unsup} are set to be 1.0. Our framework is trained on 4 RTX3090 GPUs (24GB of VRAM each) with a per-GPU batch size of 9 for the pre-training and 8 for the dual teacher knowledge acquisition stage. The training time for the dual-teacher setup is 139 seconds per epoch, which is comparable to the single-teacher setting at 103 seconds per epoch. Additional information is provided in supplementary material.

Baselines To compare with existing image enhancement and domain adaptation methods, two SOTA image enhancement methods (RFormer, LLFlow) [1, 52] and two SOTA domain adaptation methods (UDA-HE, AdvEnt) [26, 46] are used as baselines. To ensure a fair comparison, we employ the same base human pose estimator (DEKR) [14] for all the aforementioned methods, as well as for our method, utilizing only well-lit ground-truth data during training. Following the setup in ExLPose [29], we replace AdvEnt's semantic segmentation backbone with the same 2D pose estimator used in our method. UDA-HE is a model-agnostic domain-adaptive human pose estimation method. We upgrade the backbone network (ResNet) in UDA-HE by HRNet, the same backbone used by our method and re-train AdaIN to ensure a fair comparison. Note, UDA-HE is configured under a bottom-up paradigm, which is included as a bottom-up baseline in Tab. 2. For a comprehensive comparison, we add the top-down baselines reported in LSBN+LUPI [29], which are trained with both well-lit and low-light ground-truth data (unpaired). CPN [5] is used as the base human pose estimator for these top-down baselines. Among the additional baselines [29], there is an additional domain adaptation method (DANN) [13] and another image enhancement method (LIME) [15]. Note that the training data of LSBN+LUPI [29] is different from all the baselines because it relies on paired well-lit and low-light images with ground truths. All top-down baselines employ the same top-down method (CPN) [5] as the base human pose estimator.

4.2 Performance on ExLPose-OCN

Quantitative comparison on ExLPose-OCN is provided in Tab. 1. The upper part of the table is the bottom-up baselines and the lower part is top-down. Note that our method is a bottom-up one. The top-down baselines include CPN fully supervised on low-light images (CPN-low), well-lit images (CPN-well), or both of them (CPN-all), CPN integrated with low-light image enhancement methods [15, 52], and domain adaptation methods [13, 46]. It is worth mentioning all the top-down baselines (except the CPN-well) are fully supervised by both the well-lit and low-light ground truths and the reported performance, all using ground-truth person detections. The bottom-up baselines include DEKR trained on labeled low-light or well-lit images, denoted as Base-low and Base-well, respectively. Also, we integrate SOTA image enhancement methods, including RFormer [1]

Table 2: Evaluation on ExLPose-test with a bottom-up baseline [14], image enhancement [1, 52], and domain adaptation [26, 46] methods. All the methods are fully-supervised by well-lit labels. The best is **bold**. The second best is underlined.

Methods	AP†@0.5:0.95				
	LL-N	LL-H	LL-E	LL-A	WL
Base-well [14]	1.3	0.0	0.0	0.8	60.3
RFormer [1]	16.9	2.3	0.1	5.7	60.0
LLFlow [52]	11.0	0.8	0.9	4.5	60.1
UDA-HE [26]	16.4	3.6	0.1	7.4	53.7
AdvEnt [46]	4.6	0.7	0.0	2.0	56.4
Ours	**35.6**	**18.6**	**5.0**	**21.1**	**61.6**

Table 3: Evaluation on ExLPose-test with top-down methods trained on both low-light and well-lit labels (denoted by †). Differently, our method is a bottom-up method trained with well-lit labels alone. The best is **bold**. The second best is underlined.

Methods	AP†@0.5:0.95				
	LL-N	LL-H	LL-E	LL-A	WL
CPN-low† [5]	25.4	18.2	5.0	17.2	1.2
CPN-well† [5]	15.3	2.7	0.4	6.7	59.9
CPN-all† [5]	26.4	18.2	6.1	17.6	52.3
LLFlow† [52]	28.7	15.7	5.3	17.4	60.7
LIME† [15]	31.9	**21.2**	**7.6**	**21.1**	57.7
DANN† [13]	28.0	17.5	5.3	17.8	52.0
Ours	**35.6**	18.6	5.0	**21.1**	**61.6**

and LLFlow [52], as well as domain adaptation methods, including output-level method UDA-HE [26] and feature-level method AdvEnt [46]. The performance of the image enhancement methods is tested by applying the Base-well to the enhanced images. Therefore, all the bottom-up baselines including ours are trained with the well-lit ground-truth data alone with the training set of ExLPose and we only use ExLPose-OCN for evaluation.

Our method outperforms the SOTA method [29] by 6.8% (2.4 AP) on average. As ExLPose-OCN captures actual low-light images, the superior performance validates the effectiveness of our method in addressing the low-light challenges, where our method surpasses all the bottom-up and top-down baselines without using low-light ground truths or paired well-lit and low-light data. This highlights the potential of our method in real-world applications, where we achieve performance surpassing the SOTA methods without using low-light ground truths.

4.3 Performance on ExLPose-test

We compare our method with various bottom-up baselines on ExLPose-test, including a fully-supervised baseline [14] and four baselines using well-lit ground-truth data only for training based on image enhancement [1,52] or domain adaptation [26, 46], as shown in Tab. 2. As expected, the fully-supervised baseline, Base-well, is heavily biased towards their training data due to the lack of cross-domain generalization. Its poor performance clearly demonstrates the substantial gap between well-lit and extremely low-light domains. Our method outperforms the baselines using image enhancement or domain adaptation. In particular, our performance beats the second best method on the LL-N subset by 18.7 AP which almost doubled the second-best. For LL-H and LL-E subsets, all methods except ours struggled to achieve reasonable performance. Our method also maintains the best performance on the WL (well-lit) subset compared with DEKR-well.

To make a comprehensive comparison, we further compare with top-down baselines incorporated with image enhancement and domain adaptation meth-

Table 4: Ablation Study of the proposed modules in our method over ExLPose-test. PT stands for Pre-Training stage. Step 1 is the fully-supervised training on well-lit images. Step 2 is the fully supervised training with fake low-lit images after adding ELLA. KA indicates the Knowledge Acquisition stage. Single indicates the training with single teacher. Dual indicates our dual-teacher framework. The best is **bold**. The second best is underlined.

	Our Design		**AP$^{\uparrow}$@0.5:0.95**				
	ELLA	PDA	LL-N	LL-H	LL-E	LL-A	WL
PT Step 1			1.3	0.0	0.0	0.8	60.3
PT Step 2	✓		29.4	13.6	1.6	15.8	**62.1**
KA	✓		29.5	14.6	2.9	16.8	60.5
Single	✓	✓	33.4	16.6	3.1	19.0	59.4
Dual	✓	✓	**35.6**	**18.6**	**5.0**	**21.1**	61.6

Table 5: Number of the additional pseudo labels provided by the complementary teacher in ExLPose training set. OKS Thres., referring to OKS threshold, is used to determine the validity of pseudo labels by comparing the OKS between pseudo labels and the ground truth with them. PL indicates pseudo labels. Main indicates the mean teacher. Comp. indicates complementary teacher.

OKS Thres.	# of PL in Main	# of PL in Comp.	# of additional PL
0.3	10203	12155	4870 (+47.7%)
0.5	5056	8735	2136 (+49.8%)
0.7	3293	5435	1180 (+35.8%)
0.9	1463	1574	501 (+34.2%)

ods in Tab. 3. Note that the comparison is not fair to our approach as the top-down baselines reported in [29] were all trained using both low-light and well-lit ground truths. Furthermore, top-down human pose estimation methods generally outperform their bottom-up counterparts due to their utilization of human detection, as well as the reduction in human scale variation achieved through detection cropping and human patch normalization [8].

However, we still achieve the highest performance on the LL-N, LL-A, and WL subsets in Tab. 3, while obtaining comparable performance on other subsets without utilizing the low-light ground-truth data. Note LSBN+LUPI [29] is not included because its training data is different from the baselines in Tab. 3. LSBN+LUPI utilizes paired well-lit and low-light data and the ground truths, while the baselines included in the table use unpaired well-lit and low-light datasets in training. With the extra paired training data and ground truths, LSBN+LUPI manages to attain competitive performance on ExLPose-test, where its performance on the WL subset (61.5) is still worse than ours, but it is better than ours in LL-A (25.0). However, when evaluating the generalization ability of each method on the real low-light dataset (ExLPose-OCN), where there is no training and only testing, our method surpasses all the baselines, including LSBN+LUPI, as demonstrated in Tab. 1.

4.4 Ablation Studies

Ablation study is performed to validate the effectiveness of the individual modules in our method is shown in Tab. 4. We use our framework in the pre-training stage without any newly proposed components as the baseline, and subsequently add each new module in separate experiments.

ELLA We evaluate our method's performance with and without the supervision of fake low-light images generated from WL. The results are in the first two rows of Tab. 4. The inclusion of ELLA significantly boosts the performance on LL-N, LL-H, and LL-A test subsets from almost zero to a reasonable level.

PDA We investigate our method's performance with and without PDA in Stage 2 training. When comparing the third row (single teacher without PDA) to the fourth row (single teacher with PDA), we observe a 20.2% (3.4 AP) improvement on LL-A, accompanied by a marginal 0.4 AP drop on WL. This highlights the effectiveness of PDA in enabling the student to surpass the teachers.

Dual Teacher Comparing the last two rows, we observe that the model's performance with dual teachers (last row) surpasses that of a single teacher (second last row) in every split. Notably, The complementary teacher achieves a 11% (2.1 AP) improvement, which is comparable to the 13% (2.2 AP) improvement obtained by the main teacher on LL-A (all subsets), underscoring the effectiveness of our dual-teacher framework. Additionally, with the help of the dual teachers, the student model exhibits enhanced performance on the well-lit split.

To validate that the complementary teacher provides more valid pseudo labels, quantitative evaluation is provided in Table 5. We measure the validity of pseudo labels by computing their OKS with corresponding ground truths and count the number of non-overlapping pseudo labels by thresholding the OKS at 0.5 between the dual teachers, following [31, 34]. The evaluation demonstrates that the complementary teacher consistently produces additional valid pseudo labels across different thresholds of quality, providing an average of 41.8% more valid pseudo labels compared to using the main teacher alone. The increased number of valid pseudo labels leads to further improvement in the student's training, highlighting the effectiveness of our dual-teacher framework.

5 Conclusion

We introduce the first human pose estimation method for extremely low-light conditions, utilizing well-lit ground-truth data exclusively. Our novelty lies in leveraging center-based and keypoint-based teachers to generate enriched pseudo labels, enabling the student model to achieve competitive performance on extremely low-light images without the need for training with low-light ground truths. To our knowledge, this represents a significant advancement in the field of low-light human pose estimation.

Acknowledgements

We thank the anonymous reviewers and meta-reviewers for their valuable feedback.

References

1. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. arXiv preprint arXiv:2303.06705 (2023)
2. Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W.: Cross-domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9498–9507 (2019)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
4. Celik, T., Tjahjadi, T.: Contextual and variational contrast enhancement. IEEE Transactions on Image Processing **20**(12), 3431–3441 (2011)
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
6. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5386–5395 (2020)
7. Cheng, H.D., Shi, X.: A simple and effective histogram equalization approach to image enhancement. Digital signal processing **14**(2), 158–170 (2004)
8. Cheng, Y., Ai, Y., Wang, B., Wang, X., Tan, R.T.: Bottom-up 2d pose estimation via dual anatomical centers for small-scale persons. Pattern Recognition **139**, 109403 (2023)
9. Crescitelli, V., Kosuge, A., Oshima, T.: Poison: Human pose estimation in insufficient lighting conditions using sensor fusion. IEEE Transactions on Instrumentation and Measurement **70**, 1–8 (2020)
10. Crescitelli, V., Kosuge, A., Oshima, T.: An rgb/infra-red camera fusion approach for multi-person pose estimation in low light environments. In: 2020 IEEE Sensors Applications Symposium (SAS). pp. 1–6. IEEE (2020)
11. Desmarais, Y., Mottet, D., Slangen, P., Montesinos, P.: A review of 3d human pose estimation algorithms for markerless motion capture. Computer Vision and Image Understanding p. 103275 (2021)
12. Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., Yin, J.: Towards multi-pose guided virtual try-on network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9026–9035 (2019)
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
14. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14676–14686 (2021)

15. Guo, X.: Lime: A method for low-light image enhancement. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 87–91 (2016)
16. Han, Z., Sun, H., Yin, Y.: Learning transferable parameters for unsupervised domain adaptation. IEEE Transactions on Image Processing **31**, 6424–6439 (2022)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
18. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4507–4515 (2017)
19. Huang, L., Li, Y., Tian, H., Yang, Y., Li, X., Deng, W., Ye, J.: Semi-supervised 2d human pose estimation driven by position inconsistency pseudo label correction module. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 693–703 (2023)
20. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
21. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European Conference on Computer Vision. pp. 443–460. Springer (2022)
22. Jiang, J., Ji, Y., Wang, X., Liu, Y., Wang, J., Long, M.: Regressive domain adaptation for unsupervised keypoint detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6780–6789 (2021)
23. Jin, R., Zhang, J., Yang, J., Tao, D.: Multibranch adversarial regression for domain adaptative hand pose estimation. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6125–6136 (2022)
24. Jin, Y., Yang, W., Tan, R.T.: Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In: European Conference on Computer Vision. pp. 404–421. Springer (2022)
25. Kennerley, M., Wang, J.G., Veeravalli, B., Tan, R.T.: 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11484–11493 (2023)
26. Kim, D., Wang, K., Saenko, K., Betke, M., Sclaroff, S.: A unified framework for domain adaptive pose estimation. In: European Conference on Computer Vision. pp. 603–620. Springer (2022)
27. Kocabas, M., Karagoz, S., Akbas, E.: Multiposenet: Fast multi-person pose estimation using pose residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 417–433 (2018)
28. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
29. Lee, S., Rim, J., Jeong, B., Geonu Kim, B.W., Lee, H., Cho, S., Kwak, S.: Human pose estimation in extremely low-light conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
30. Li, C., Lee, G.H.: From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1482–1491 (2021)
31. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10863–10872 (2019)

32. Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. IEEE Transactions on Image Processing **27**(6), 2828–2841 (2018)
33. Lin, H., Zhang, Y., Qiu, Z., Niu, S., Gan, C., Liu, Y., Tan, M.: Prototype-guided continual adaptation for class-incremental unsupervised domain adaptation. In: European Conference on Computer Vision. pp. 351–368. Springer (2022)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
35. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: Deeplpf: Deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12826–12835 (2020)
36. Mu, J., Qiu, W., Hager, G.D., Yuille, A.L.: Learning from synthetic animals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12386–12395 (2020)
37. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. Advances in Neural Information Processing Systems **30** (2017)
38. Peng, Q., Zheng, C., Chen, C.: Source-free domain adaptive human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4826–4836 (2023)
39. Punnappurath, A., Abuolaim, A., Abdelhamed, A., Levinshtein, A., Brown, M.S.: Day-to-night image synthesis for training nighttime neural isps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10769–10778 (June 2022)
40. Rahman, S., Rahman, M.M., Abdullah-Al-Wadud, M., Al-Quaderi, G.D., Shoyaib, M.: An adaptive gamma correction for image enhancement. EURASIP Journal on Image and Video Processing **2016**(1), 1–13 (2016)
41. Raychaudhuri, D.S., Ta, C.K., Dutta, A., Lal, R., Roy-Chowdhury, A.K.: Prior-guided source-free domain adaptation for human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14996–15006 (2023)
42. Sharma, A., Tan, R.T.: Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11977–11986 (2021)
43. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
44. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
45. Tian, Z., Chen, H., Shen, C.: Directpose: Direct end-to-end multi-person pose estimation. arXiv preprint arXiv:1911.07451 (2019)
46. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2517–2526 (2019)
47. Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11060–11068 (2022)

48. Wang, D., Zhang, S., Hua, G.: Robust pose estimation in crowded scenes with direct pose-level inference. Advances in Neural Information Processing Systems **34** (2021)
49. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
50. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6849–6857 (2019)
51. Wang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE transactions on image processing **22**(9), 3538–3548 (2013)
52. Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.P., Kot, A.: Low-light image enhancement with normalizing flow. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2604–2612 (2022)
53. Wei, K., Fu, Y., Yang, J., Huang, H.: A physics-based noise formation model for extreme low-light raw denoising. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
54. Wei, K., Fu, Y., Zheng, Y., Yang, J.: Physics-based noise modeling for extreme low-light photography. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8520–8537 (2021)
55. Weng, C.Y., Curless, B., Kemelmacher-Shlizerman, I.: Photo wake-up: 3d character animation from a single photo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5908–5917 (2019)
56. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 466–481 (2018)
57. Xie, R., Wang, C., Zeng, W., Wang, Y.: An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11240–11249 (2021)
58. Xue, N., Wu, T., Xia, G.S., Zhang, L.: Learning local-global contextual adaptation for multi-person pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
59. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
60. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 889–898 (2019)