Human Pose Estimation in Extremely Low-Light Conditions
— *Supplementary Material* —

Sohyun Lee[1*] Jaesung Rim[1*] Boseung Jeong[2] Geonu Kim[2] Byungju Woo[2]
Haechan Lee[1] Sunghyun Cho[1,2†] Suha Kwak[1,2†]

[1]Graduate School of AI, POSTECH [2]Dept. of CSE, POSTECH

http://cg.postech.ac.kr/research/ExLPose

Contents

This supplementary material presents additional experimental details and results that are omitted from the main paper due to the space limit. Section A explains extremely low-light images of the ExLPose dataset. Section B shows a comparison of dataset scale with other human pose datasets. Section C presents a detail of the geometric alignment of our dual-camera system. Section D describes experimental details about implementation (Sec. D.1) and network architectures (Sec. D.2). Section E shows an in-depth analysis of our method, including lighting condition analysis (Sec. E.1), lighting condition insensitive features analysis (Sec. E.2), and additional analysis for our method (Sec. E.3). Section F shows detailed results of the various combinations of the existing enhancement methods and the pose estimation network. Section G gives extensive experimental results to investigate the components of our method, such as LSBN (Sec. G.1), LUPI (Sec. G.2, G.3, G.4), and intensity scaling (Sec. G.5). Section H presents experimental results of person detection in extremely low-light conditions. Finally, Section I shows additional qualitative results.

A. Explanation on Extremely Low-light Images

Low-light conditions are important as they are prevalent in many scenarios with limited illumination, such as nighttime and low-light indoor environments. Capturing images in such conditions is challenging due to the requirement of a long exposure time, which can cause motion blur, a thorny problem to solve. To avoid blur, a common practice is to utilize extremely low-light images captured with a short exposure time for various tasks in low-light environments, such as image enhancement [3, 12, 24, 25], action recognition [27], image matching [19], optical flow estima-

* Equal contribution. † Corresponding authors.

tion [29]. In the case of human pose estimation, capturing images with a long exposure time is impossible due to the movement of humans. Thus, we study pose estimation under extremely low-light conditions using an extremely low-light image rather than images captured with long exposure. The low-light images can be amplified to improve visibility using global scaling, but noise is also amplified, resulting in extremely noisy images. Fig. a1 shows low-light and scaled images in the SID [3] and ExLPose datasets. The scaled images are extremely noisy since the scaling amplifies noise; the degree of noise depends on that of darkness.

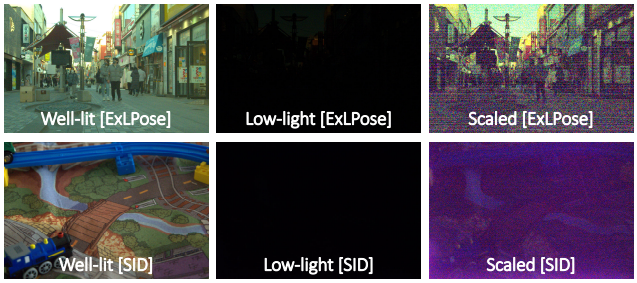

Figure a1. Examples of well-lit, extremely low-light, and scaled low-light images in the SID [3] and ExLPose datasets.

B. Scale Comparison to Other Datasets

We compare our dataset to existing human pose estimation datasets. As shown in the Table a1, the scale of our dataset is large enough since it is comparable to common datasets in terms of the number of annotated people as shown in the table. Note that the unit of supervision in pose estimation is a human instance, not an image.

Dataset	Paired	#Poses	Multi-person
LSP [13]		2,000	
LSP Extended [14]		10,000	
MPII Single-person [1]		26,429	
FLIC [18]		5,003	
We are family [6]		3,131	✓
OCHuman [28]		8,110	✓
MPII Multi-person [1]		14,993	✓
CrowdPose [16]		80,000	✓
ExLPose (Ours)	✓	14,214	✓

Table a1. Statistics of the ExLPose and other datasets.

C. Geometric Alignment of Our Camera System

We physically align two camera modules of the dedicated dual-camera system as much as possible to capture images that are aligned with each other. Nevertheless, there may exist a small amount of geometric misalignment between the cameras as discussed in [17]. Moreover, while moving around the camera system collecting the dataset, the movement of the camera system may introduce additional geometric misalignment.

To resolve this, we captured a reference image pair of a static scene before collecting data every time we moved the camera system, and estimated a homography between them. We set the exposure time of the low-light camera module 100 times longer so that pairs of reference images have the same brightness for accurate homography estimation. Then, we estimated a homography matrix between them using the method of [7] and aligned the collected well-lit images using the estimated homography matrix. Fig. a2(a)-(b) visualize the effect of geometric alignment using stereo-anaglyph images where a pair of well-lit and scaled low-light images from the camera modules are visualized in red and cyan. As the figure shows, even before the geometric alignment, images from our camera system have only a small amount of misalignment. Nevertheless, the geometric alignment can successfully resolve the remaining misalignment.

Figure a2. Stereo-anaglyph images (a) before and (b) after geometric alignment.

D. Experimental Details

In this section, we present the implementation settings and detailed network architectures that are omitted from the main paper due to the space limit.

D.1. Implementation Details

Pose Estimation Network. The pose estimation network is trained by the Adam optimizer [15] with a weight decay of $1e{-}5$ and a learning rate set initially to $5e{-}4$ and decreased by a factor of 2 every six epochs. Each mini-batch consists of 32 images from each lighting condition. Following Cascaded Pyramid Network (CPN) [4], each human box image

is cropped and resized to a fixed size, 256×192. Then we apply random scale ($0.7 \sim 1.35$) augmentation. During training, we multiply a weight of $1e{-}3$ to $\mathcal{L}_{\text{LUPI}}$.

Person Detection Network. The person detection network, *i.e.* Cascade R-CNN [2] with ResNeXt101 [26] pre-trained on ImageNet [5], is optimized by the SGD with a momentum of 0.9, and a weight decay of $1e{-}4$ within 12 epochs. The initial learning rate is set to $2e{-}3$ for the detection network, and it is decayed by a factor of 0.1 for the eight and eleven epochs. For each lighting condition, the mini-batch size is set to 2. All input images are resized to a fixed size, 1333×800. Similar to the pose estimation network, we multiply a weight of $1e{-}3$ to $\mathcal{L}_{\text{LUPI}}$ during training. In testing, duplicated detected boxes are filtered out by Non-Maximum Suppression (NMS) with an IoU threshold of 0.7.

D.2. Network Architectures

Fig. a3 and Fig. a4 depict the detailed architecture of human pose estimation and person detection, respectively.

Pose Estimation Network. For the human pose estimation network, as shown in Fig. a3, we adopt CPN [4], involving two sub-networks of GlobalNet and RefineNet. GlobalNet is similar to the feature pyramid structure for key point estimation. In GlobalNet, each feature from four residual blocks (*i.e.*, R1-R4) is applied 1×1 convolutional kernel and then element-wise summed. Then, RefineNet concatenates all the features from GlobalNet. For GlobalNet and RefineNet, the pose estimation losses are applied to the output feature of each sub-network, where we denote $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{refine}}$ for them, respectively. As mentioned Sec. 5.1 in the main paper, LUPI is applied to the feature maps of the first convolutional layer and the following four residual blocks of a ResNet backbone. For more details on the pose estimation network, please refer to the CPN [4] paper.

Person Detection Network. We utilize Cascade R-CNN [2] as the person detection network. In Fig. a4, teacher and student networks are trained by the common detection losses for classification and regression, denoted as $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{reg}}$, respectively. We first try applying LUPI on the feature maps of the first convolutional layer and the four residual blocks following our experimental setting in human pose estimation. However, we found that applying LUPI on the output features from the feature pyramid network is more effective than applying the loss on the features from residual blocks. Therefore, as presented in the figure, we finally apply LUPI on the feature maps of the first convolutional layer and the following feature pyramid network. More detailed information for the person detection network is described in the Cascade R-CNN [2] paper.

E. Empirical Analysis

E.1. Analysis on Lighting Conditions

Our method is based on the assumption that the neural style encodes the lighting condition. In this section, we empirically verify this assumption by visualizing the distributions of Gram matrices computed from low-light and well-lit images. Their Gram matrices are computed at 1st, 2nd, 3rd, and 4th Res Blocks of ResNet-50 [10] pre-trained on the ImageNet dataset [5]. Fig. a5 shows t-SNE [20] visualization of Gram matrices. In the figure, we can observe that images of the same lighting condition are grouped together in the style spaces, which validates our assumption.

E.2. Qualitative Analysis on Lighting Condition Insensitive Features

We conduct an experiment to investigate the impact of lighting condition insensitive features learned by our method. To this end, we train an image reconstruction model composed of a backbone of CPN as an encoder and a decoder which consists of transposed convolution layers on the well-lit image of the ExLPose dataset. Then, we replace the encoder of the reconstruction model with that of the pose estimation model trained by our method. Fig. a6 shows the reconstructed results on well-lit and low-light images. The figure demonstrates that our model learns features insensitive to lighting conditions in that reconstructed images from well-lit and low-light are consistent results.

E.3. Additional Quantitative Analysis on Our Method

In the main paper, we compare the average Hausdorff distance [11] between sets of Gram matrices under different lighting conditions. In this section, we show additional results to investigate the style gaps between well-lit and paired low-light images. To this end, we compute the mean squared error (MSE) [23] distance on gram matrices between low-light and well-lit images of a pair before and after applying LUPI. Then, we report the average values for the overall ExLPose dataset. Fig. a7 presents that the style gaps between low-light and well-lit conditions are reduced by LUPI.

F. Results of Various Applying Enhancement Methods

F.1. Results on the ExLPose Dataset

In this section, we report the detailed results of adopting existing enhancement methods. To this end, we first apply enhancement methods on low-light images of the ExLPose dataset as pre-processing. For the learning-based enhancement method (*i.e.*, LLFlow), we train the method using paired low-light and well-lit images of the ExLPose dataset.

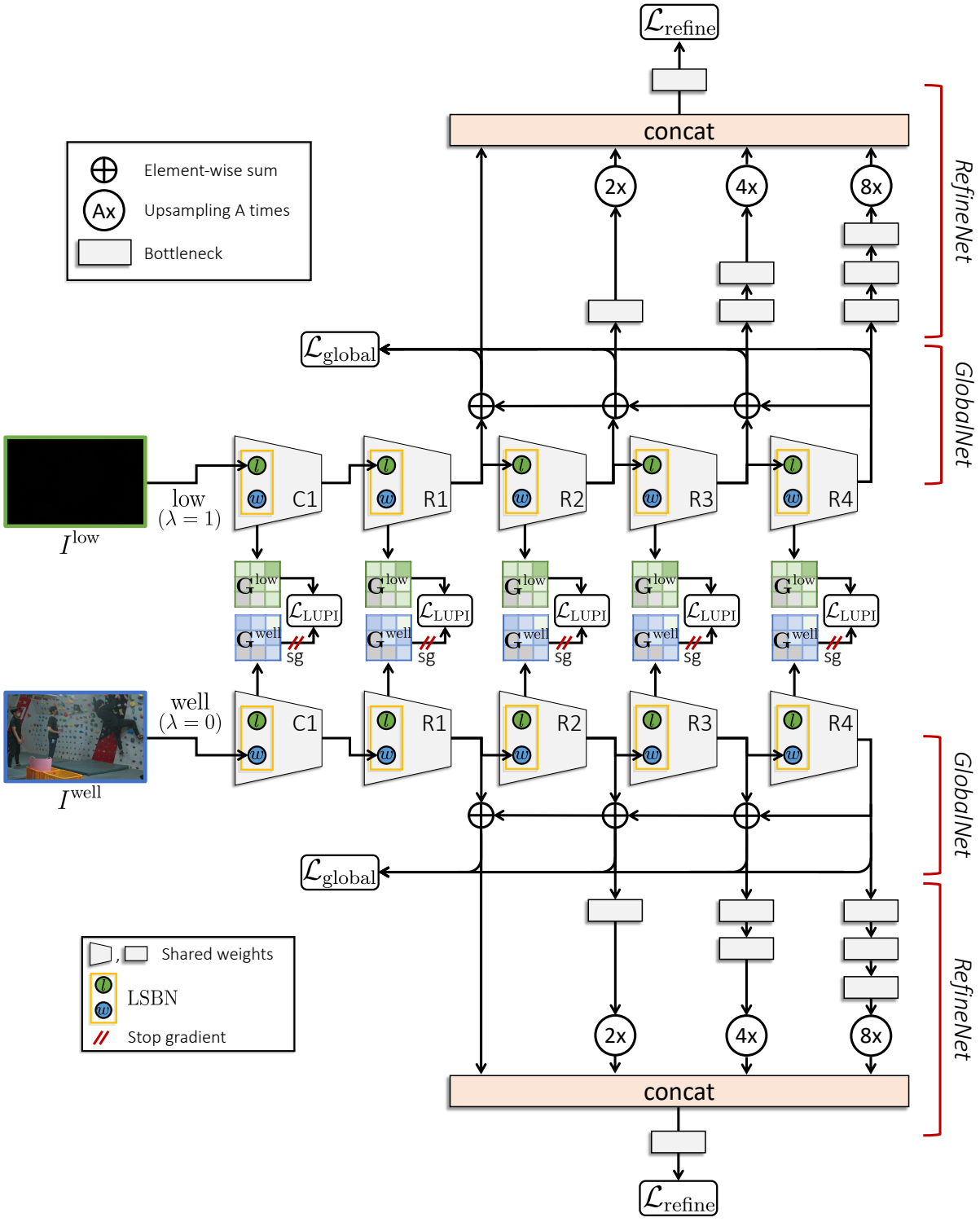

Figure a3. Illustration of the detailed human pose estimation network architecture. Both of teacher (*bottom*) and student (*top*) are trained by the same pose estimation loss of $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{refine}}$ from GlobalNet and RefineNet of CPN [4]. LUPI is applied to the feature maps of the first convolutional layer (*i.e.*, C1) and four residual blocks (*i.e.*, R1-R4) of a ResNet backbone. Teacher and student share all the parameters except LSBNs.

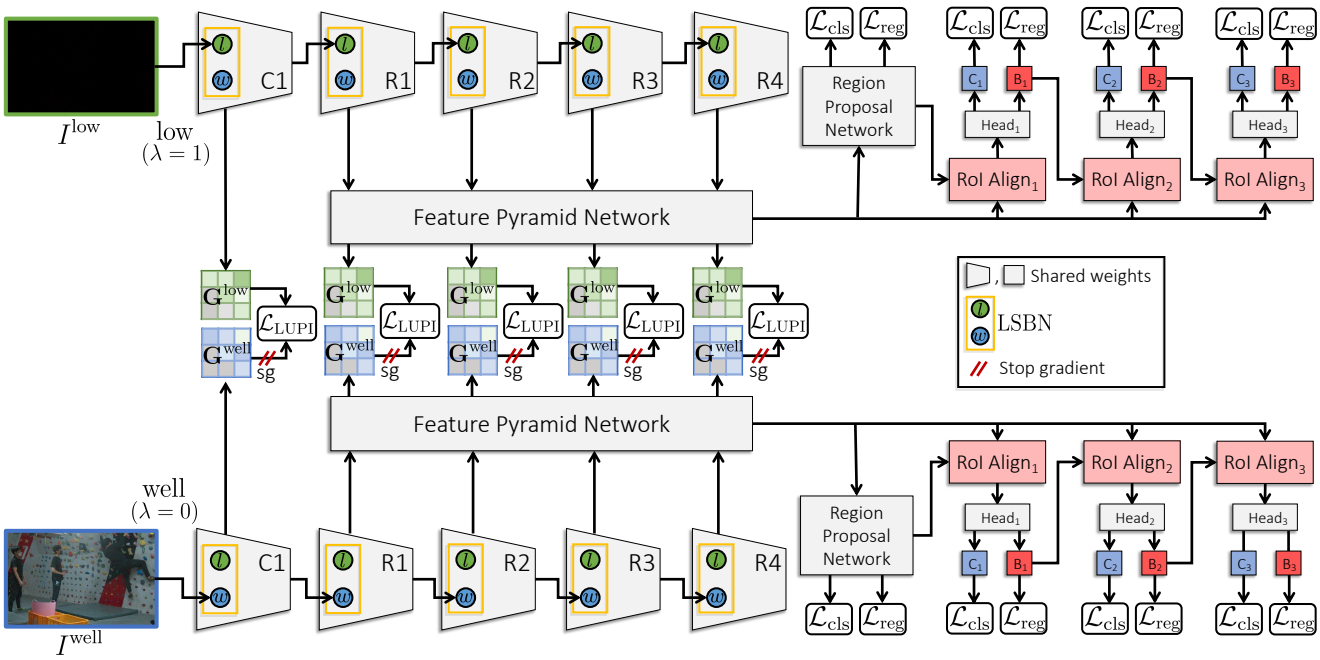

Figure a4. The overall architecture of the person detection network and training strategy. Both of teacher (*bottom*) and student (*top*) are trained by the same detection loss functions ($\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{reg}}$), and student takes additional supervision from teacher through LUPI. The loss for LUPI is applied to the feature maps of the first convolutional layer (*i.e.*, C1) and the following feature pyramid network (FPN) which takes feature maps of four residual blocks (*i.e.*, R1-R4) of a ResNet backbone as inputs. Teacher and student share all the parameters except LSBNs. "Head", "C", and "B" denote roi head, object classification, and bounding box, respectively.

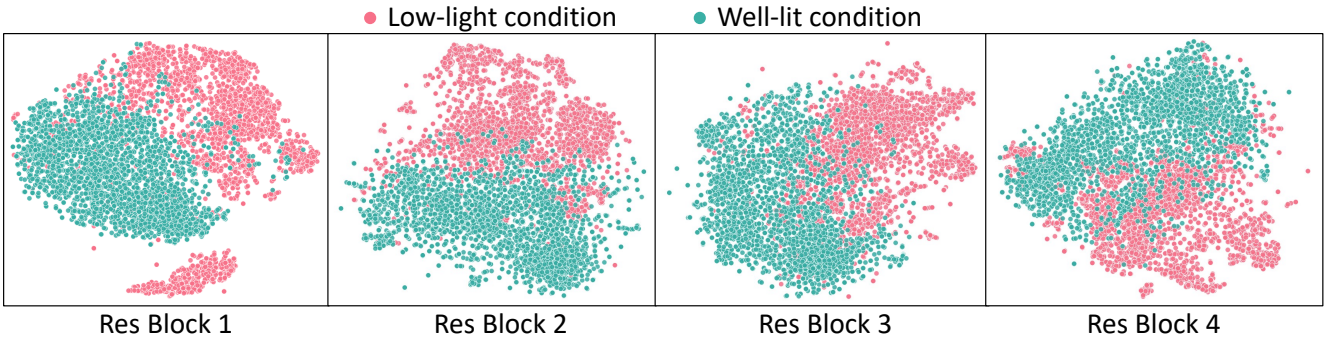

Figure a5. t-SNE visualization of the distributions of Gram matrices computed from low-light and well-lit images. The Gram matrices are computed from feature maps of the 1st, 2nd, 3rd, and 4th Res Blocks of ResNet-50 [10] pre-trained on the ImageNet dataset [5]. In all the visualizations, images of the same lighting condition are clustered together, suggesting that neural styles in the form of Gram matrices encode lighting conditions.

For the traditional enhancement method (*i.e.*, LIME), we apply the method on the low-light images of the ExL-Pose dataset without the training process. The straightforward way to incorporate these enhancement modules with a pose estimation network is to exploit the enhanced low-light images as inputs to evaluate Baseline-well, which is the pose estimation network trained using well-lit images only. In Table a2, LLFlow + Baseline-well and LIME + Baseline-well show the performance of pose estimation using LLFlow [22] and LIME [9] as pre-processing, respectively. They achieve inferior performance since the enhancement module is optimized to enhance the quality of low-light images but not to improve performance on the downstream recognition models. To address this issue, we train the pose estimation network using enhanced low-light images, denoted by LLFlow + Baseline-low† and LIME + Baseline-low†. The table shows that the pose estimation network trained using enhanced low-light images performs better. We also found that training both enhanced low-light and well-lit images significantly improves the perfor-

Figure a6. Input images and reconstructed images by the proposed method.

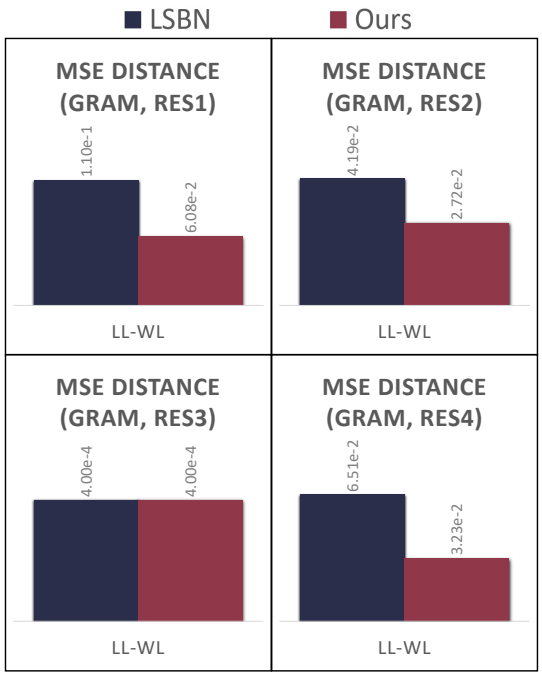

Figure a7. Style gaps between low-light (LL) and well-lit (WL) conditions. The gap is computed by averaging the mean squared error (MSE) distance between low-light and paired well-lit images on the ExLPose dataset.

mance. As presented in the table, LLFlow + Baseline-all† and LIME + Baseline-all† show the pose estimation network trained using both enhanced low-light and well-lit images achieves the best performance among other variants. Our method outperforms all variants of adopting enhancement methods regardless of their training strategy.

AP@0.5:0.95	LL-N	LL-H	LL-E	LL-A	WL
Baseline-well	23.5	7.5	1.1	11.5	**68.8**
Baseline-low	32.6	25.1	13.8	24.6	1.6
Baseline-all	33.8	25.4	14.3	25.4	57.9
LLFlow + Baseline-well	22.9	11.7	2.4	12.8	-
LLFlow + Baseline-low†	30.7	18.4	8.0	19.6	40.9
LLFlow + Baseline-all†	35.2	20.1	8.3	22.1	65.1
LIME + Baseline-well	23.1	5.8	1.0	10.8	-
LIME + Baseline-low†	31.6	24.3	12.6	23.6	36.2
LIME + Baseline-all†	40.6	27.1	13.4	28.3	63.2
Ours	**42.3**	**34.0**	**18.5**	**32.7**	68.5

Table a2. Quantitative results in AP@0.5:0.95 on the ExLPose dataset; Low-light-normal, Low-light-hard, Low-light-extreme, Low-light-all, Well-lit splits. Baseline-low† is a model trained using enhanced low-light images, and Baseline-all† denotes a model trained using both enhanced low-light and well-lit images.

F.2. Results on the ExLPose-OCN Dataset

We also evaluate each combination of enhancement and pose estimation methods in Sec. F.1 on the ExLPose-OCN dataset. Table a3 shows that the tendency of each model is similar to that of each model in Table a2. In detail, a trained pose estimation model using both enhanced low-light and well-lit images (*i.e.*, LLFlow + Baseline-all† and LIME + Baseline-all†) outperforms the variants of them. Our method is still superior to all variants of the combinations of enhancement and pose estimation methods.

G. Additional Ablation Studies

G.1. Effect of LSBN

This section presents extensive experiments to investigate the effect of LSBN. As mentioned Sec. 5.1.1 in the

AP@0.5:0.95	A7M3	RICOH3	Avg.
Baseline-well	23.7	23.9	23.8
Baseline-low	15.2	15.6	15.4
Baseline-all	32.8	31.7	32.2
LLFlow + Baseline-well	30.4	24.5	27.3
LLFlow + Baseline-low†	20.5	18.7	19.5
LLFlow + Baseline-all†	25.6	28.2	27.0
LIME + Baseline-well	10.9	7.9	9.3
LIME + Baseline-low†	20.7	13.4	16.8
LIME + Baseline-all†	33.2	28.4	30.7
Ours	**35.3**	**35.1**	**35.2**

Table a3. Quantitative results in AP@0.5:0.95 on the ExLPose-OC dataset; A7M3, and RICOH3 splits. Baseline-low† is a model trained using enhanced low-light images, and Baseline-all† denotes a model trained using both enhanced low-light and well-lit images.

main paper, domain adaptation (DA) methods cannot effectively reduce the large domain gap between low-light and well-lit conditions. As presented in Table a4, DANN [8], AdvEnt [21], and LUPI show inferior performance, proving the necessity of LSBN. When they are combined with LSBN, the performance of each method is improved since it successfully bridges the large domain discrepancy between low-light and well-lit conditions. When compared with 'LSBN + DANN' and 'LSBN + AdvEnt', our method (*i.e.*, 'LSBN + LUPI') outperforms them thanks to the effectiveness of our neural style-based approach, LUPI. It is worth noting that our method can be a plug-and-play to the feature extractor, while AdvEnt is hard to apply to the task where the entropy map cannot be computed.

AP@0.5:0.95	LL-N	LL-H	LL-E	LL-A	WL
DANN	34.9	24.9	13.3	25.4	58.6
AdvEnt	33.0	24.1	11.6	23.8	60.0
LUPI	34.2	23.1	11.2	24.0	61.7
LSBN + DANN	42.2	30.5	16.7	30.8	67.4
LSBN + AdvEnt	41.3	31.2	**19.0**	31.5	**68.5**
LSBN + LUPI (Ours)	**42.3**	**34.0**	18.5	**32.7**	**68.5**

Table a4. Analysis for the effect of LSBN. The results are reported in AP@0.5:0.95 on the ExLPose dataset; Low-light-normal, Low-light-hard, Low-light-extreme, Low-light-all, and Well-lit conditions.

G.2. Effect of the Neural Style of LUPI

We investigate the effect of LUPI by comparing LSBN + LUPI (*i.e.*, our method) with LSBN + LUPI-*feat* that directly approximates feature maps of the teacher instead neural styles. We conduct an in-depth analysis by comparing them in terms of the feature gaps between low-light and well-lit conditions. As presented in Fig. a8, LSBN + LUPI effectively reduces the average Hausdorff distance of feature maps between different lighting conditions, although LSBN + LUPI-*feat* directly approximates the feature maps of well-lit images. We conjecture that the advantage of using neural styles over directly using features comes from the less image-dependent characteristic of neural styles, *i.e.*, features are more image-specific information that changes with each image; thus it is more difficult for LSBN + LUPI-*feat* to reduce the difference between features. In consequence, LSBN + LUPI effectively aligns the feature distributions of different lighting conditions since they aim to approximate the styles which represent the characteristics of lighting conditions.

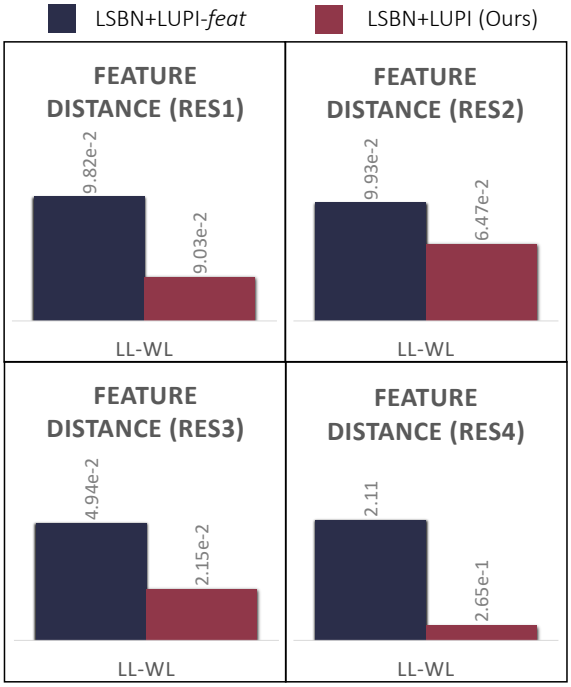

Figure a8. Quantitative analysis on feature distribution gap between low-light (LL) and well-lit (WL) conditions. The distance is measured by the average Hausdorff distance between lighting condition sets.

G.3. Effect of Layer Selection of LUPI

We investigate the impact of the selection of layers where LUPI is applied. Table a5 compares different choices for the layer selection. In the table, C1 applies LUPI to the output of the first convolutional layer, and C1:Rn applies LUPI to the 1st to the n-th residual blocks as well as the first convolution layer. As shown in the table, the performance improves as LUPI is applied to more blocks, and our final model (C1:R4) achieves the best performance.

AP@0.5:0.95	LL-N	LL-H	LL-E	LL-A	WL
C1	41.0	30.0	15.0	29.7	67.6
C1:R1	40.8	30.0	17.0	30.3	67.1
C1:R2	42.2	30.6	15.8	30.7	67.9
C1:R3	**43.1**	32.2	17.7	32.2	67.9
Ours (C1:R4)	42.3	**34.0**	**18.5**	**32.7**	**68.5**

Table a5. Analysis on layers where LUPI is applied.

G.4. Effect of the Gradient Direction of LUPI

When training with LUPI, we let the gradient from the loss flow *only* to the student in order to train the student with privileged information from the teacher, i.e., information flows in one direction from the teacher to the student. In this section, we study the effect of this one-direction strategy on LUPI. To this end, we prepare three variants of our approach: 'T $\rightarrow$ S (Ours)', 'T $\leftrightarrow$ S' and 'T $\leftarrow$ S'. 'T $\rightarrow$ S (Ours)' is our proposed approach. 'T $\leftrightarrow$ S' allows the gradient from LUPI to flow to both the teacher and student models, i.e., the teacher and student can affect each other. 'T $\leftarrow$ S', on the other hand, allows the gradient to flow only to the teacher. Table a6 compares the performance of these three variants. As shown in the table, our approach clearly outperforms the others. This result implies our one-direction strategy is essential for learning about LUPI.

AP@0.5:0.95	LL-N	LL-H	LL-E	LL-A	WL
T $\leftrightarrow$ S	41.6	32.8	16.7	31.6	67.0
T $\leftarrow$ S	39.7	29.6	15.7	29.5	68.4
T $\rightarrow$ S (Ours)	**42.3**	**34.0**	**18.5**	**32.7**	**68.5**

Table a6. Analysis on the impact of the gradient direction from LUPI.

G.5. Effect of Intensity Scaling

As described in the main paper, the average channel intensity of each low-light image is automatically scaled to 0.4 before being fed to the student network. Table a7 shows the performance of Baseline-all and the proposed method trained on original low-light images and scaled low-light images. In low-light conditions, automatically scaled low-light images significantly improve the performance of both models. However, Baseline-all trained on the scaled low-light images performs much worse in well-lit conditions.

We suspect that, in the case of using original low-light images, the Baseline-all model is biased to the well-lit condition. It is because well-lit images have large pixel intensities, so the scale of gradient of them is larger than that of original low-light images. Then, in the case of using scaled low-light images, the Baseline-all model is less biased for the well-lit condition, so the performance on the well-lit condition is decreased. However, the proposed method is less biased due to the lighting condition invariant features of LSBN and LUPI. Consequently, Table a7 demonstrates that intensity scaling of low-light images improves the performance of both Baseline-all and our method for low-light conditions.

AP@0.5:0.95	Method	LL-N	LL-H	LL-E	LL-A	WL
No scaling	Baseline-all	22.3	6.5	2.5	11.2	61.3
	Ours	25.5	10.0	5.9	14.5	67.1
Scaling	Baseline-all	33.8	25.4	14.3	25.4	57.9
	Ours	**42.3**	**34.0**	**18.5**	**32.7**	**68.5**

Table a7. Analysis on impact of scaling for low-light images.

H. Results of Person Detection

The ExLPose dataset provides human pose and bounding box labels for training and evaluation of human pose estimation methods. Moreover, human bounding boxes in the ExLPose dataset can serve as a detection dataset on low-light images.We adopt Cascade R-CNN [2] as our person detection network and compare our method with other solutions in the same way as described in the main paper.

Table a8 shows summarized the person detection performance of ours and other solutions. In the table, Baseline-all is a Cascade R-CNN trained on both low-light and well-lit images with person detection loss only, and Baseline-all still underperforms our method. For comparing other solutions, we adopt LLFlow and LIME for enhancement methods and DANN for domain adaptation. AdvEnt cannot be applied to the person detection task as the method is based on the entropy minimization of prediction. Accordingly, we did not conduct experiments about AdvEnt on person detection. As shown in the table, these methods rather degrade performance in low-light conditions as a direct adoption of enhancement and domain adaptation is not effective. On the other hand, our method outperforms by large margins in all the low-light conditions, and the results show the effectiveness of our method for person detection in the low-light condition.

AP@0.5:0.95	LL-N	LL-H	LL-E	LL-A	WL
Baseline-low	39.8	30.5	17.4	30.2	40.7
Baseline-well	22.6	3.8	0.5	9.7	53.2
Baseline-all	44.9	32.3	18.3	33.0	60.8
LLFlow + Baseline-all	38.3	29.0	15.2	28.4	61.3
LIME + Baseline-all	45.8	32.1	17.0	32.8	**63.5**
DANN	43.7	30.1	14.8	30.7	55.7
Ours	**46.2**	**34.5**	**21.0**	**34.9**	60.9

Table a8. Person detection results in AP@0.5:0.95 on the ExLPose dataset; Low-light-normal, Low-light-hard, Low-light-extreme, Low-light-all, and Well-lit conditions.

I. Additional Qualitative Results

I.1. Results on the ExLPose-OCN Dataset

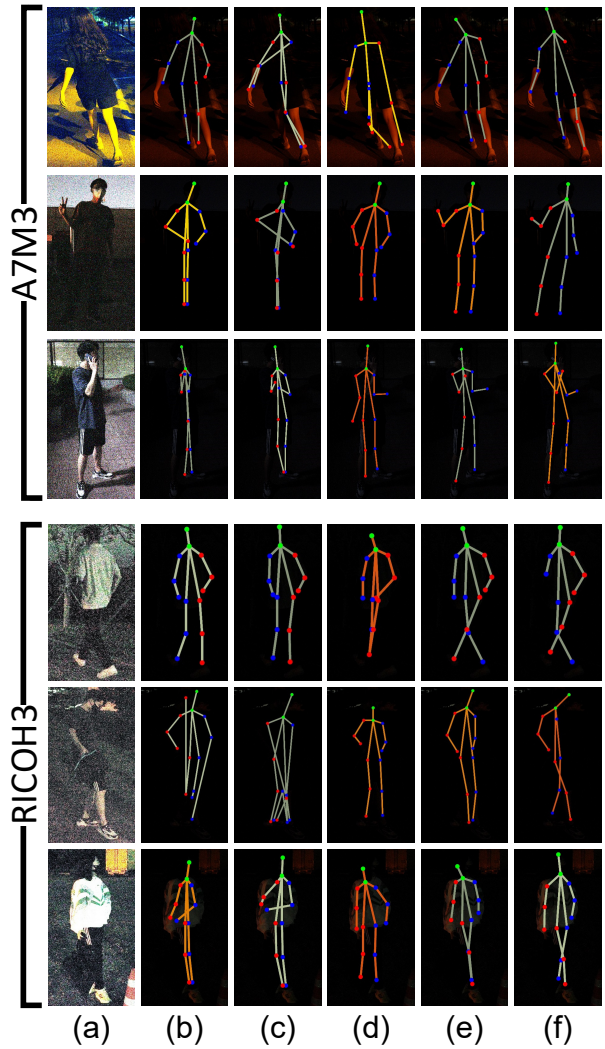

Figure a9. Qualitative results on the ExLPose-OCN dataset. Predicted poses and labels are visualized on corresponding low-light images. (a) Scaled low-light images. (b) Baseline-all. (c) DANN. (d) LIME + Baseline-all. (e) Ours. (f) Ground-truth.

We provide the ExLPose-OCN dataset to evaluate the generalization capability of pose estimation methods to unseen cameras. Fig. a9 shows qualitative comparisons of our method with Baseline-all, DANN, and LIME + Baseline-all on the ExLPose-OCN dataset. As shown in the figure, our method accurately estimates human poses, but other methods do not generalize well to unseen cameras and often fail to estimate accurate human poses.

I.2. Results on the ExLPose Dataset

Fig. a10 shows additional qualitative results of Baseline-all, DANN [8], LIME [9] + Baseline-all and our method. This again demonstrates that Baseline-all and DANN often fail to predict poses, while our method surpasses them.

I.3. Failure Cases of Our Method

We provide failure cases of our method in Fig. a11. The first row of the figure shows the results of the pose estimation network on low-light images which have little pixel information. In such images, noise components are prevalent, and the remaining pixel information is too small to estimate human poses. Our method also fails to predict human poses for occluded humans, as shown second and third rows in the figure.

I.4. Results of Enhancement Methods

Fig. a12 shows enhanced low-light images of the ExLPose and ExLPose-OCN datasets using LLFlow [22] and LIME [9]. For the ExLPose dataset, LLFlow successfully enhances low-light images and reduces the noises of low-light images. However, LLFlow cannot generalize well to the ExLPose-OCN dataset due to different image signal processors and exhibit different noise distributions. Enhanced low-light images using LIME have the remaining noise as the method does not consider noise well. These limitations may reduce the generalization capability and performance of the pose estimation when enhancement methods are combined with the pose estimation network.

I.5. Results of Multi-person Pose Estimation

Fig. a13 presents qualitative results for the person detection of Baseline-all, DANN, LIME + Baseline-all, and our method. These predicted bounding boxes of our method are exploited for multi-person pose estimation; its qualitative results are shown in Fig. a14. The figure shows that our method successfully performs multi-person pose estimation while other solutions largely fail to estimate human poses.

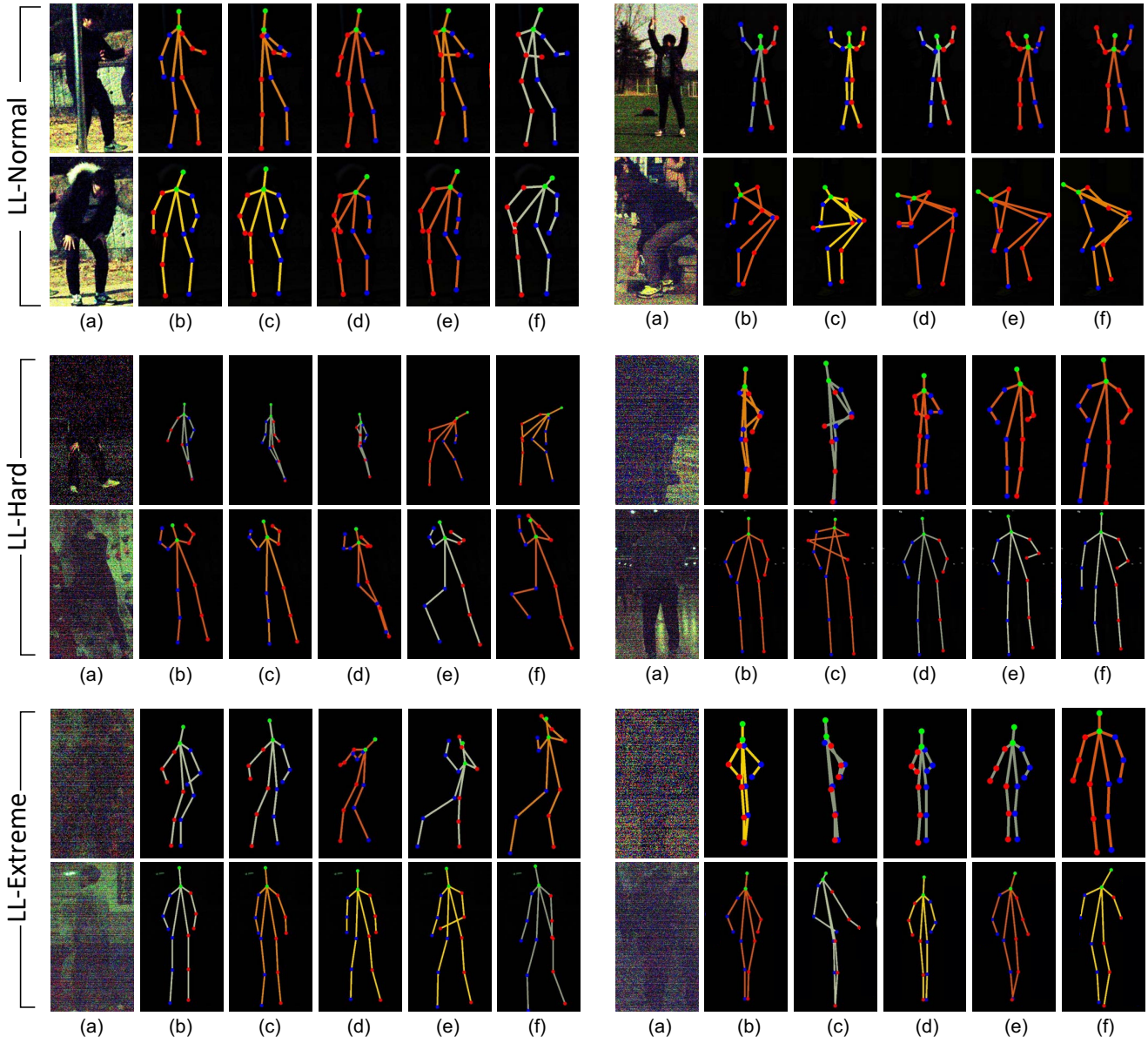

Figure a10. Qualitative results of single-person pose estimation on the ExLPose dataset. Predicted poses and labels are visualized on corresponding low-light images. (a) Scaled low-light images. (b) Baseline-all. (c) DANN. (d) LIME + Baseline-all. (e) Ours. (f) Ground-truth.

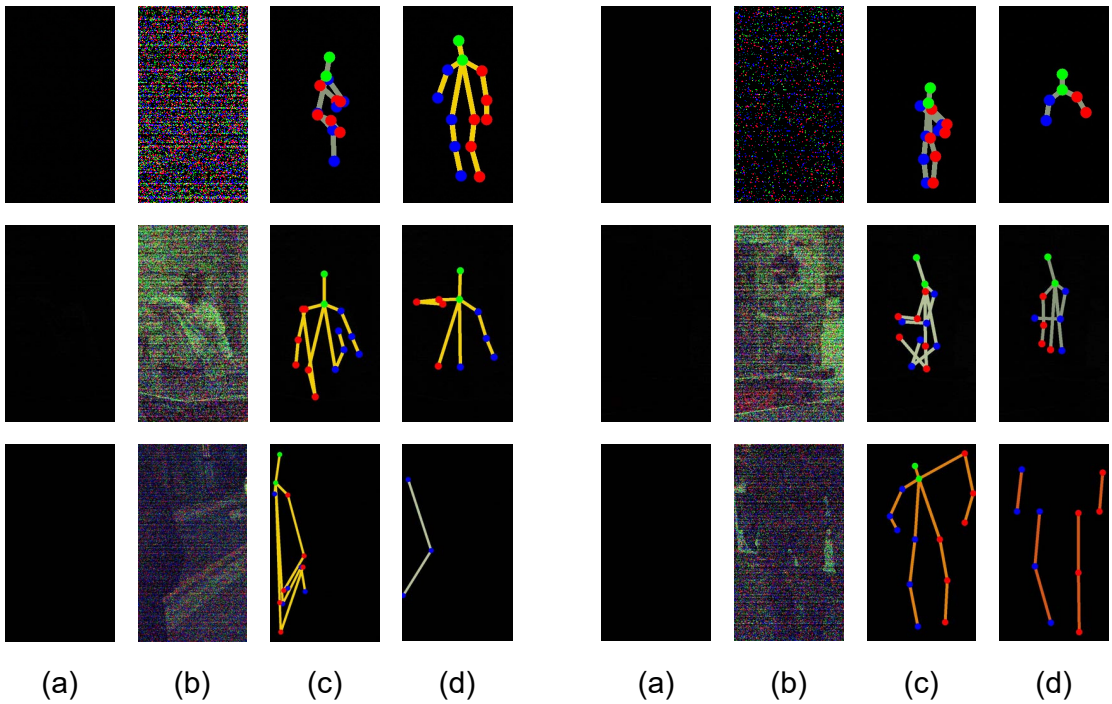

Figure a11. Failure cases of our method. (a) Low-light images. (b) Scaled low-light images. (c) Our results. (d) Ground-truth.

Figure a12. Qualitative results of enhanced low-light images on the ExLPose and ExLPose-OCN datasets. We except well-lit images of the ExLPose-OCN dataset as the dataset provides only low-light images. (a) Low-light images. (b) LLFlow. (c) LIME. (d) Well-lit images.

Figure a13. Qualitative results for the person detection on the ExLPose dataset. Predicted boxes and labels are visualized on corresponding low-light images. (a) Scaled low-light images. (b) Baseline-all. (c) DANN. (d) LIME + Baseline-all. (e) Ours. (f) Ground-truth.

Figure a14. Qualitative results for multi-person pose estimation on the ExLPose dataset. Predicted poses and labels are visualized on corresponding low-light images. (a) Scaled low-light images. (b) Baseline-all. (c) DANN. (d) LIME + Baseline-all. (e) Ours. (f) Ground-truth.

References

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

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[6] Marcin Eichner and Vittorio Ferrari. We are family: Joint pose estimation of multiple persons. In *Proc. European Conference on Computer Vision (ECCV)*, 2010.

[7] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 2016.

[9] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing (TIP)*, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1993.

[12] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[13] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. British Machine Vision Conference (BMVC)*, 2010.

[14] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[16] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[18] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[19] Wenzheng Song, Masanori Suganuma, Xing Liu, Noriyuki Shimobayashi, Daisuke Maruta, and Takayuki Okatani. Matching in the dark: A dataset for matching image pairs of low-light scenes. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[20] L.J.P van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008.

[21] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C Kot. Low-light image enhancement with normalizing flow. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[23] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine (SPM)*, 2009.

[24] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *Proc. International Workshop on Deep Learning for Human Activity Recognition*, 2021.

[28] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.