UDAPose: Unsupervised Domain Adaptation for Low-Light Human Pose Estimation

Supplementary Material

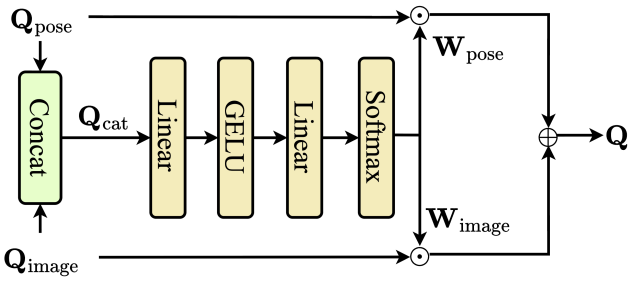


Figure 6. The architecture of our DCA module.

7. Implementation and Experimental Details

7.1. Human Pose Estimation Model

7.1.1. Architecture of DCA

We present the details of the proposed Dynamic Control of Attention (DCA) module in Fig. 6. As described in the main paper (Sec. 3 Method), DCA first concatenates $\mathbf{Q}_{\text{pose}}$ with $\mathbf{Q}_{\text{image}}$ as $\mathbf{Q}_{\text{cat}}$. Then $\mathbf{Q}_{\text{cat}}$ goes through two-layer MLP ended with Softmax to acquire $\mathbf{W}_{\text{pose}}$ and $\mathbf{W}_{\text{image}}$, adaptive weights for pose priors $\mathbf{Q}_{\text{pose}}$ and image cues $\mathbf{Q}_{\text{image}}$, respectively. At last, DCA fuses weighted sum of pose prior $\mathbf{Q}_{\text{pose}}$ and image cues $\mathbf{Q}_{\text{image}}$ as output $\mathbf{Q}$ for subsequent FFN and following decoder layers. As shown in Fig. 3 of the main paper, DCA is placed after deformable cross-attention within each decoder layer to substitute original direct sum of residual connection.

7.1.2. Loss Functions

The overall loss functions of human pose estimation model can be formulated as:

$$\mathcal{L} = \mathcal{L}_h + \mathcal{L}_c + \mathcal{L}_k \tag{11}$$

$$\mathcal{L}_h = \mu \left| H - \hat{H} \right| + \beta(1 - \text{GIoU}) \tag{12}$$

$$\begin{aligned} &\mathcal{L}_c = -\lambda\alpha(1-p_t)^\gamma \log(p_t), \\ &\text{where } p_t = p \text{ if } y = 1, p_t = 1 - p \text{ if } y \neq 1 \end{aligned} \tag{13}$$

$$\begin{aligned} \mathcal{L}_k = \omega &\left| P - \hat{P} \right| \\ &+ \theta \frac{\sum_i^K \exp\left(-\left|P_i - \hat{P}_i\right| / 2s^2k_i^2\right) \delta(v_i > 0)}{\sum_i^K \delta(v_i > 0)} \end{aligned} \tag{14}$$

where $\mathcal{L}_h$ is for human box regression that contains L1 loss and GIoU [52] loss, $\mathcal{L}_c$ is for human classification, which is a focal loss [37] with $\alpha = 0.25, \gamma = 2$, and $\mathcal{L}_k$ is for keypoint regression that includes L1 loss and the constrained L1 loss-OKS loss [56]. $|H - \hat{H}|$ is the L1 distance between the predicted human boxes and the ground-truth ones. $y \in \pm 1$ specifies the ground-truth class, and $p \in [0, 1]$ is the estimated probability for the class with label $y = 1$. $|P - \hat{P}|$ is the L1 distance between predicted keypoints inside a human and the ground-truth ones. $|P_i - \hat{P}_i|$ is the L1 distance between the i-th predicted keypoint and ground-truth one, v_i is the visibility flag of the ground truth, s is the object scale, and k_i is the per-keypoint constant that controls falloff. The loss coefficients $\mu, \beta, \lambda, \omega, \theta$ are 5, 2, 2, 10, 4.

7.2. Datasets

As introduced in the main paper, we evaluate UDAPose on the ExLPose dataset [30], which is specifically for benchmarking 2D human pose estimation in extremely low-light conditions. The ExLPose training set consists of 2,065 well-lit and optically filtered low-light image pairs, with pose annotations following the CrowdPose format [34]. These images span 251 indoor and outdoor scenes, with low-light versions generated using a dual-camera system under varying conditions to simulate diverse low-light scenarios. ExLPose provides two test sets: ExLPose-test and ExLPose-OCN. ExLPose-test, also referred to as Low-Light All (LL-A), is further divided into three difficulty levels: Low-Light Normal (LL-N), Low-Light Hard (LL-H), and Low-Light Extreme (LL-E). To validate our method's generalization ability, we also performed cross-dataset validation on EHPT-XC [8]. EHPT-XC is a dataset combining RGB and event camera data for human pose estimation and tracking in challenging low-light and motion blur conditions. It encompasses RGB video frames from 158 diverse sequences, along with pixel-wise aligned and temporally synchronized event streams, and annotations containing 38K 2D keypoints and bounding boxes with track IDs. To focus on low-light conditions, we combined the train and test split of EHPT-XC and constructed a specific subset of 12 scenes (1200 images) for cross-dataset validation.

7.3. Discussion on Dual-Camera Data Usage

The dual-camera setup in ExLPose [30] is an effective design that enables paired data collection by transferring annotations from well-lit images to their low-light counterparts. However, this setup relies on hardware-specific ac-

	AP↑@0.5:0.95						
	WL	LL-N	LL-H	LL-E	A7 M3	RIC OH3	EHPT -XC
4,000	66.1	34.7	22.4	5.4	50.0	45.1	25.4
8,000	66.3	35.4	24.8	7.8	53.2	46.8	27.9
12,000	66.9	36.6	26.2	10.4	54.7	47.2	29.1
16,000	66.8	37.5	27.3	11.3	55.0	47.8	30.5
20,000	67.3	38.7	28.0	11.7	55.0	47.9	31.0

Table 7. Performance vs. amount of synthetic training data.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	KL↓
CycleGAN [80]	36.56	0.76	0.26	50.20	0.028
UNIT [39]	33.90	0.66	0.29	45.70	0.104
UNSB [26]	34.19	0.74	0.30	96.42	0.062
EnCo [2]	35.99	0.75	0.28	48.71	0.031
Ours	**41.13**	**0.91**	**0.20**	**11.17**	**0.008**

Table 8. Evaluation for human pose anatomical consistency of our method and learning-based baselines.

quisition and cannot be applied to synthesize low-light images from existing well-lit human pose datasets. In addition, it is not easily scalable for collecting new data, as it requires a specialized camera system rather than standard cameras. In contrast, our method allows leveraging existing well-lit human pose datasets with available annotations, enabling flexible and scalable low-light data generation without requiring specialized camera systems.

In our experiment, we use the dual-camera low-light images from ExLPose as style references. While such images may not fully represent real-world low-light conditions for supervised learning (e.g., due to optical filtering), they still capture useful characteristics such as illumination patterns and noise distributions. This usage is consistent with our goal of synthesizing low-light images from well-lit data, rather than directly training on limited low-light datasets. Another reason is to ensure fair comparison with prior work [1, 30]. We use the same dataset as the source of style references, avoiding performance gains from larger or more diverse real nighttime datasets. Otherwise, improvements could be attributed to data scale instead of the proposed method. By using the same dataset, we isolate performance gains to the proposed method.

7.4. Implementation Details

For each well-lit image in ExLPose, we randomly select one low-light image from ExLPose or ExLPose-OCN as its style reference, and repeat this process 10 times, yielding 20k synthetic low-light images for training. We further analyze data scaling in Tab. 7, where performance improves with more synthetic training images (with larger gains below 12k), and use 20k images in the main experiments. The synthesis cost is approximately 263 ms per image on an RTX 4090.

Following the pipeline of ED-Pose [74], we adopt the overall pose-estimation framework and focus our contributions on the proposed DCA module. We utilize Swin-Transformer [40] (Swin-T) pretrained on ImageNet-22k [11] as the multi-scale image feature extraction backbone. During training, we apply data augmentations including random crop, random flip, and random resize (shorter side in [480, 800], longer side ≤ 1333), following DETR [5] and PETR [56]. To accelerate the early-stage training, we adopt the human query denoising training strategy from DN-DETR [33]. We use the AdamW [29, 41] optimizer with weight decay of 1×10^{-4} and train our pose model on 2 NVIDIA RTX PRO 6000 GPUs with batch size 16 for 120 epochs on ExLPose [30]. The initial learning rate is 1×10^{-4} and is decayed at the 100th epoch by a factor of 0.1. The channel dimension of the Transformer layers is set to 256. At test time, we resize each input image so that its shorter side is 800 pixels while keeping the longer side no more than 1333 pixels. DCA introduces only 4.1% inference overhead to pose model (39 ms vs. 37.4 ms per image) on an RTX PRO 6000.

7.5. Experiment Settings

For image enhancement methods [3, 13, 19, 65], we directly use the official checkpoints released by the authors to ensure a fair comparison. When applying the image enhancement models for human pose estimation evaluation, we first convert low-light images from ExLPose-test, ExLPose-OCN and EHPT-XC using their models. After that, we apply human pose estimation model trained on ExLPose well-lit images on these enhanced images to test performance.

For domain adaptive methods [1, 2, 26, 27, 39, 80], we follow the same procedure as used in existing methods [1]. The human pose estimation model is first trained on ExLPose well-lit dataset and then finetuned on augmented low-light images. At test time, we directly input low-light images from ExLPose-test, ExLPose-OCN and EHPT-XC to test their performance.

8. Anatomical Consistency

We evaluate the anatomical consistency of our generated low-light images, an important factor for reusing human pose annotations from well-lit datasets. This evaluation also helps verify that our method preserves structural consistency and avoids unintended structure leakage from the style reference. To this end, we synthesize low-light images from the well-lit inputs in ExLPose and evaluate them against the paired low-light images (i.e., ExLPose includes paired well-lit and low-light images.) using a comprehensive set of metrics. We assess image fidelity at the pixel level with peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), and at the feature level

	AP↑@0.5:0.95					
	WL	LL-N	LL-H	LL-E	A7 M3	RIC OH3
Main (ELLA)	**62.1**	29.4	13.6	1.6	35.0	27.2
Main (Ours)	61.5	32.3	23.2	8.3	37.2	35.0
Comp. (ELLA)	60.3	27.8	11.9	0.8	33.9	26.5
Comp. (Ours)	60.8	31.7	22.4	6.7	36.8	33.9
Student (ELLA)	60.8	35.6	18.6	5.0	39.1	36.2
Student (Ours)	61.1	**39.4**	**27.4**	**9.4**	**41.3**	**39.3**
LSBN+LUPI [30]	61.1	33.7	14.7	3.4	35.3	35.1

Table 9. Full comparison results of ELLA [1] and our method. "Main" refers to "main teacher". "Comp." refers to "complementary teacher". And "student" refers to "student" distillation model, which is the full model of ELLA. The best is **bold**.

	AP↑@0.5:0.95					
	WL	LL-N	LL-H	LL-E	A7 M3	RIC OH3
Direct input	60.8	23.7	7.3	0.0	27.3	24.3
Z-score-based norm.	60.9	25.4	13.2	2.2	28.4	25.0
Fixed factor	58.1	29.0	20.8	6.4	33.2	31.0
ImageNet-based	61.3	31.7	22.4	7.4	36.5	33.8
Ours	**61.5**	**32.3**	**23.2**	**8.3**	**37.2**	**35.0**

Table 10. Evaluation of the AIN module on ExLPose-test and ExLPose-OCN. Direct input refers to feeding low-light images into SD without AIN. Experiments are conducted using the DEKR pose model [14], with DHF and LCIM enabled for all normalization approaches. The best is **bold**.

with learned perceptual image patch similarity (LPIPS) and Fréchet Inception Distance (FID). More importantly, to directly quantify anatomical integrity, we compute the Kullback–Leibler (KL) divergence between predicted heatmaps on our synthetic low-light images and on the paired low-light images from ExLPose. A pose estimator (DEKR [14]) trained on the low-light data from ExLPose is used to predict the heatmaps in this experiment.

Learning-based adaptation methods (e.g., unpaired image-to-image translation or style transfer) with competitive performance are used here as baselines. As shown in Tab. 8, our method significantly outperforms across every metric. Our method achieves a PSNR of 41.13 and an SSIM of 0.91, indicating superior pixel-level accuracy. Furthermore, our method obtains the lowest LPIPS (0.20) and a remarkably low FID of 11.17, confirming that the generated images have higher perceptual quality and a feature distribution much closer to that of real images. Importantly, our method obtains a KL divergence of only 0.008, which is much lower than the best-performing baseline (CycleGAN: 0.028). These results provide solid evidence that our generation process preserves the underlying human anatomical structure faithfully, which facilitates downstream human pose estimation tasks using our synthetic data.

9. Comparison with ELLA and Supervised Low-Light Training

ELLA [1] is based on DEKR [14], which utilizes different type of loss (e.g. center-offset, joints-tags) for dual-teacher design, while our backbone, ED-Pose [74], directly regresses to 2D coordinate for each keypoint. Therefore, we cannot directly use ELLA's dual-teacher in our framework. In this case, we evaluate our low-light image synthesis in ELLA's dual-teacher pipeline without our proposed DCA. In particular, we integrate our synthetic data into the dual-teacher-student distillation framework proposed by ELLA [1]. Tab. 9 shows the detailed results at each stage of the ELLA framework, including both teacher models and the final student model.

The comparison reveals that while ELLA's main teacher achieves a slightly higher performance on well-lit (WL) images, our main and complementary teachers consistently and significantly outperform their counterparts across all low-light conditions (LL-N, LL-H, and LL-E). For instance, our main teacher improves performance on the challenging LL-H and LL-E sets by +9.6 AP and +6.7 AP, respectively. These results demonstrate that our synthetic data more effectively captures low-light characteristics than ELLA's handcrafted augmentation, leading to improved teacher models.

Consequently, the stronger teacher models using our synthetic data lead to a more effective student model. Our final distilled student surpasses the ELLA student by a substantial margin across all low-light subsets on both ExLPose-test and ExLPose-OCN. Notably, our student achieves remarkable improvements of +8.8 AP on LL-H, +4.4 AP on LL-E, and +5.1 AP on A7M3. These results demonstrate the effectiveness of our approach in generating low-light images for human pose estimation, resulting in a stronger student model within the ELLA framework.

We also include a baseline (LSBN+LUPI [30]) trained directly with labeled low-light data from ExLPose (dual-camera) as shown in the last row of Tab. 9. Despite using supervised low-light annotations, this baseline is outperformed by our method, indicating that training on synthesized low-light data can generalize better than relying on limited paired low-light data.

10. Evaluation of the AIN Module

Low-light images often contain extremely low intensity values, which can cause the VAE encoder in the SD model to produce corrupted latent codes. To address this, we introduce Adaptive Intensity Normalization to the real low-light reference images I_{LL} right before feeding it into the SD-

	AP↑ @0.5:0.95						
	WL	LL-N	LL-H	LL-E	A7 M3	RIC OH3	EHPT -XC
z_0	66.8	29.7	12.1	0.1	40.0	36.8	11.3
z_0, z_1	66.8	31.4	19.2	2.4	41.4	37.4	15.4
z_0, z_1, z_2	67.2	35.3	23.2	5.8	43.8	39.9	26.2
z_0, z_1, z_2, z_3	**67.4**	37.7	26.5	7.7	47.9	43.7	29.7
z_0, z_1, z_2, z_3, z_4	67.3	**38.7**	**28.0**	**11.7**	**55.0**	**47.9**	**31.0**

Table 11. Ablation study of LCIM on ExLPose-test, ExLPose-OCN, and EHPT-XC. z_0 refers to baseline SD without any extra intermediate features. z_1 to z_4 represent low-to-high-frequency information fused in a coarse-to-fine integration strategy. Results are reported with AIN, DHF and DCA. The best is **bold**.

VAE encoder. This process can be formulated as:

$$I_{\text{LL}} \leftarrow I_{\text{LL}} \times \frac{\delta}{\mu_{I_{\text{LL}}}} \tag{15}$$

where $\delta = 0.449$, which is the mean intensity of ImageNet [11] across all channels, and $\mu_{I_{\text{LL}}}$ represents average intensity of I_{LL} across all channels. We conduct a comprehensive ablation study to validate the AIN module with DEKR [14] as the pose estimation model. As shown in Tab. 10, we compare our method against several alternative normalization strategies.

First, we establish a baseline by feeding low-light images directly into the network without any normalization ("Direct input"). This approach yields poor performance, with Average Precision (AP) scores dropping to a mere 7.3 on the LL-H set and 0.0 on the LL-E set. This result underscores the critical need for an effective input normalization technique to handle the challenges of low-light conditions.

Next, we evaluate several alternative normalization strategies. Applying z-score-based normalization offers only a marginal improvement, which is formulated as

$$I'_{\text{LL}} = \frac{\sigma_{\text{ImageNet}}}{\sigma_{\text{LL}}}(I_{\text{LL}} - \mu_{\text{LL}}) + \mu_{\text{ImageNet}} \tag{16}$$

This approach is not suitable for low-light images, where pixel values are highly concentrated near zero. The mean-subtraction operation introduces numerous negative values, which can disrupt the original signal distribution and discard subtle but important low-light noise characteristics. Using a fixed scaling factor for the whole dataset is another option but not optimal as well, which is formulated as

$$I'_{\text{LL}} = I_{\text{LL}} \times k \tag{17}$$

Since low-light scenes exhibit diverse illumination levels, a single fixed factor can cause over-exposure in relatively brighter images and insufficient enhancement in darker ones, failing to produce a consistently normalized input. A third option, per-channel scaling (e.g., using ImageNet's standard "[0.485, 0.456, 0.406]" values), provides a slightly better result. However, this approach distorts the intrinsic color balance by altering the relative strengths of the R, G, and B channels. This can cause an undesirable color shift and prevent the model from learning to handle realistic low-light color noise faithfully.

In contrast, our proposed AIN, which adaptively rescales each image using a single, content-aware factor, achieves superior performance across all evaluated scenarios as shown in Table 10. AIN improves the AP to 32.3, 23.2, and 8.3 on LL-N, LL-H, and LL-E, respectively, outperforming all other variants. By preserving the inter-channel ratios, our method avoids color distortion. By adapting the scaling factor to each image's mean intensity, it effectively normalizes brightness without introducing clipping artifacts. This process provides a stable and informative input for the downstream network, leading to significantly improved human pose estimation accuracy. These results validate our design choices and demonstrate the effectiveness of AIN for low-light human pose estimation.

11. Analysis of LCIM

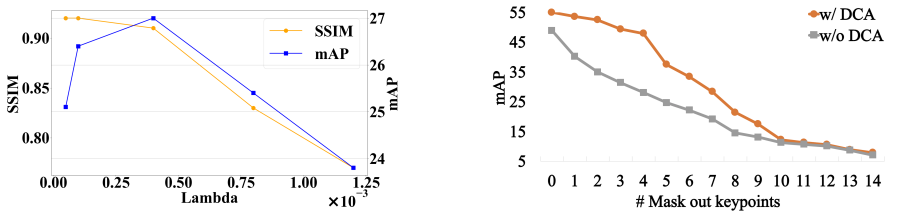
We now analyze the core of our LCIM module: the multi-scale intermediate features. As detailed in Tab. 11, we start with a baseline model (z_0) that omits all intermediate features, then progressively integrate features from coarse to fine levels (z_1 to z_4). The z_0 model, achieves a modest 40.0 AP on A7M3 and 36.8 AP on RICOH3.

These results show that fusing multi-scale features is important. Each added intermediate feature brings a consistent performance improvement. Adding all four feature levels ($+z_1 + z_2 + z_3 + z_4$) results in our strongest model, improving the AP by +15.0 on A7M3 (40.0→55.0) and +11.1 on RICOH3 (36.8→47.9) compared to the z_0 baseline. This analysis shows that our coarse-to-fine fusion strategy effectively uses multi-scale latent features from the SD encoder, which is important for robust pose estimation under challenging low-light conditions.

12. Evaluation of λ and DCA

12.1. Sensitivity of λ

As defined in Eq. (5) of the main paper, λ controls the relative weight between the pixel-level MSE loss $\mathcal{L}_{\text{MSE}}$ and the frequency-domain loss $\mathcal{L}_{\text{freq}}$ during LCIM training. We analyze its effect by varying λ and measuring both image-level quality (SSIM between synthesized and real low-light images) and downstream pose estimation performance (mAP on ExLPose-OCN). As shown in Fig. 7a, increasing λ places more emphasis on high-frequency detail preservation, which improves the fidelity of low-light noise patterns in the synthesized images and leads to higher mAP. However, beyond a certain point, an overly large λ degrades con-



(a) (b)

Figure 7. (a) Effect of λ. (b) Masking evaluation w/ and w/o DCA.

	AP↑@0.5:0.95						
	WL	LL-N	LL-H	LL-E	A7 M3	RIC OH3	EHPT -XC
SE-Block [16]	62.4	36.7	26.3	9.5	50.3	46.5	26.7
CBAM [68]	62.5	37.0	26.2	9.8	51.1	46.2	27.0
Ours (DCA)	**67.3**	**38.7**	**28.0**	**11.7**	**55.0**	**47.9**	**31.0**

Table 12. Comparison of SE-Block [16], CBAM [68], and our DCA gating mechanism. The best is **bold**.

	AP↑@0.5:0.95						
	WL	LL-N	LL-H	LL-E	A7 M3	RIC OH3	EHPT -XC
ControlNet [77]	66.3	31.7	16.4	2.7	47.6	43.7	22.4
IP-Adapter [76]	65.8	31.5	17.1	3.5	48.4	43.4	24.1
Ours	**67.3**	**38.7**	**28.0**	**11.7**	**55.0**	**47.9**	**31.0**

Table 13. Comparison of ControlNet [77], IP-Adapter [76], and our method. The best is **bold**.

tent consistency, as indicated by a drop in SSIM. This occurs because the frequency loss begins to dominate, causing the decoder to prioritize noise texture over structural content from the well-lit source image. Conversely, a small λ underweights the frequency loss, producing synthesized images that lack realistic low-light noise and thus provide insufficient training signal for the pose estimator. Based on this tradeoff, we set $\lambda = 4 \times 10^{-4}$ in all experiments.

12.2. Robustness Analysis of DCA

To further evaluate DCA beyond the ablation study in the main paper, we design a masking experiment that probes DCA's ability to use pose priors when visual information is missing. Specifically, we evaluate on ExLPose-OCN (A7M3) by progressively masking a random subset of ground-truth keypoints in each test image, simulating scenarios where varying numbers of joints are occluded or invisible. As shown in Fig. 7b, DCA consistently outperforms the baseline (without DCA) when a small number of keypoints are masked. This is because DCA detects unreliable image cues for the masked keypoints and shifts its reliance toward learned pose priors, leading to more reliable predictions for these keypoints compared to relying on noisy visual cues alone. As the number of masked keypoints increases, the gap between DCA and the baseline narrows. This is expected: when the majority of keypoints are invisible, even pose priors offer limited information, as the model has fewer visible joints to anchor its structural reasoning. This also indicates a limitation of DCA under extreme conditions, where very limited visual evidence constrains the effectiveness of pose priors.

13. Ablation Study of DCA

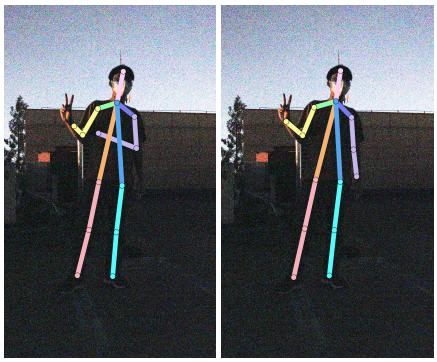
We compare DCA against two general-purpose attention gating mechanisms: SE-Block [16] and CBAM [68]. Each replaces DCA at the same position in the decoder layer, fusing $\mathbf{Q}_{\text{pose}}$ and $\mathbf{Q}_{\text{image}}$ before the FFN. All three variants use the same synthesized low-light training data, with DHF and LCIM enabled for all variants. As shown in Tab. 12, SE-Block and CBAM both improve over the no-gating baseline (Table 5 in the main paper, "+ DHF" row). However, DCA outperforms both. On ExLPose-test, DCA leads SE-Block by 1.7–2.2 AP across the low-light subsets and by 4.9 AP on well-lit images. The gap is larger on cross-dataset evaluation, where DCA exceeds SE-Block and CBAM by 4.3 and 4.0 AP on EHPT-XC, respectively. This is likely due to the explicit softmax competition in DCA between pose-prior and image-cue channels: the two weights sum to one per keypoint, forcing the model to make a binary-like choice for each joint. SE-Block and CBAM instead learn generic channel or spatial reweighting without this structural constraint, so they lack the inductive bias to suppress unreliable image cues for specific keypoints.

14. Comparison to ControlNet and IP-Adapter

To further evaluate the quality of our synthesized low-light data, we compare against two commonly used diffusion-based conditioning methods: ControlNet [77] and IP-Adapter [76]. Both methods are trained on paired well-lit and low-light images from the ExLPose dual-camera system, providing them with direct pixel-level supervision that our method does not require. ControlNet adds spatial conditioning to the diffusion model, while IP-Adapter injects reference image features through a decoupled cross-

	Before	After
L. shoulder	1.73	1.74
R. shoulder	1.81	1.78
L. elbow	1.72	1.41
R. elbow	1.77	1.47
L. wrist	1.69	1.24
R. wrist	1.74	1.49
L. hip	1.70	1.72
R. hip	1.80	1.81
L. knee	1.85	1.55
R. knee	1.84	1.57
L. ankle	1.76	1.51
R. ankle	1.81	1.52
Head	1.89	1.93
Neck	1.82	1.89

Before After



	Before	After
L. shoulder	1.74	1.71
R. shoulder	1.71	1.74
L. elbow	1.82	1.73
R. elbow	1.81	1.72
L. wrist	1.85	1.74
R. wrist	1.84	1.93
L. hip	1.91	1.75
R. hip	1.93	1.73
L. knee	1.83	1.13
R. knee	1.86	1.17
L. ankle	1.88	1.31
R. ankle	1.87	1.35
Head	1.79	1.84
Neck	1.82	1.70

Before After

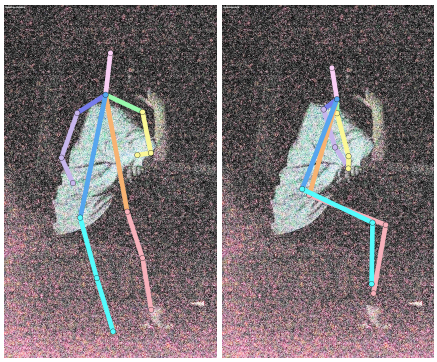


Figure 8. Qualitative ablation of our DCA module. L. represents left, R. represents right.

attention mechanism. All experiments are conducted using ED-Pose [74] with the DCA module enabled, and our method also includes the proposed DHF and LCIM.

As shown in Tab. 13, our method outperforms both baselines across all evaluation sets without relying on paired training data. On ExLPose-test, the performance gap widens as conditions become more challenging: our method leads by 7.0 AP on LL-N, 10.9 AP on LL-H, and 8.2 AP on LL-E compared to the best-performing baseline. On ExLPose-OCN, we observe gains of 6.6 AP on A7M3 and 4.2 AP on RICOH3. On the cross-dataset EHPT-XC benchmark, our method achieves 31.0 AP, surpassing IP-Adapter by 6.9 AP. This advantage mainly comes from our DHF and LCIM modules, which extract and inject high-frequency low-light characteristics at multiple scales in the decoder. In contrast, ControlNet and IP-Adapter use general-purpose conditioning mechanisms that are not designed for modeling low-light noise patterns. These results suggest that task-specific characteristic injection can be more effective than general diffusion-based conditioning for low-light data synthesis, even when the latter has access to paired supervision.

15. Qualitative results

15.1. Comparison of Pose Prediction

We present a qualitative comparison of pose prediction results between our proposed UDAPose and related methods including DarkIR [13], QuadPrior [65], CycleGAN [80], UNSB [26], EnCo [2], ELLA [1] in Fig. 9. The qualitative results clearly demonstrate the superior capability of our approach in predicting human pose under low-light conditions. We first observe the limitations of enhancement-based methods. Both DarkIR [13] and QuadPrior [65] rely on a pre-processing step to enhance the image. However, this enhancement procedure is often ill-posed in extreme darkness and can introduce visual artifacts that mislead the subsequent pose estimation model. This is evident as they produce a biologically implausible pose, or fail to detect the person altogether. In contrast, domain adaptation methods such as CycleGAN [80], UNSB [26], EnCo [2], and ELLA [1] show improved performance by training on synthetic data. Nevertheless, they still produce inaccurate joint locations. We attribute this to the limited fidelity of their synthetic data, which fails to fully capture the complex degradations of real-world low-light imagery. Our method overcomes these limitations and yields a substantially more accurate result. This superior performance stems from two key factors: (1) our high-fidelity data synthesis pipeline, which provides training examples that reflect low-light characteristics, and (2) our DCA module, which adaptively balances unreliable visual cues from the noisy image with robust, learned anatomical priors. This allows our model to maintain structural coherence and precision even in extreme conditions.

15.2. Comparison of Synthesized Images

We present a qualitative comparison between our proposed UDAPose and related methods including QuadPrior [65], CycleGAN [80], UNIT [39], UNSB [26], StyleID [9] in Fig. 10. The qualitative results clearly demonstrate the superior capability of our approach in generating low-light images that capture the characteristics of low-light images.

Among the image enhancement-based methods, QuadPrior [65] attempts to brighten low-light images but tends to produce over-smoothed results with significant loss of texture and detail. For the image-to-image translation methods, CycleGAN [80] produces images with excessive color shifts and unrealistic noise patterns. UNIT [39] and UNSB [26] generate images with irregular noise distributions that significantly differ from real low-light conditions, making them less effective for training human pose estimation models. StyleID [9], while better at preserving the overall scene structure, still struggles to accurately capture the complex noise patterns of the low-light images.

Figure 9. Pose predictions of UDAPose compared against competing methods. The first two columns show results from enhancement-based methods; all other columns display results on the original low-light images. The low-light images are scaled for visualization only.

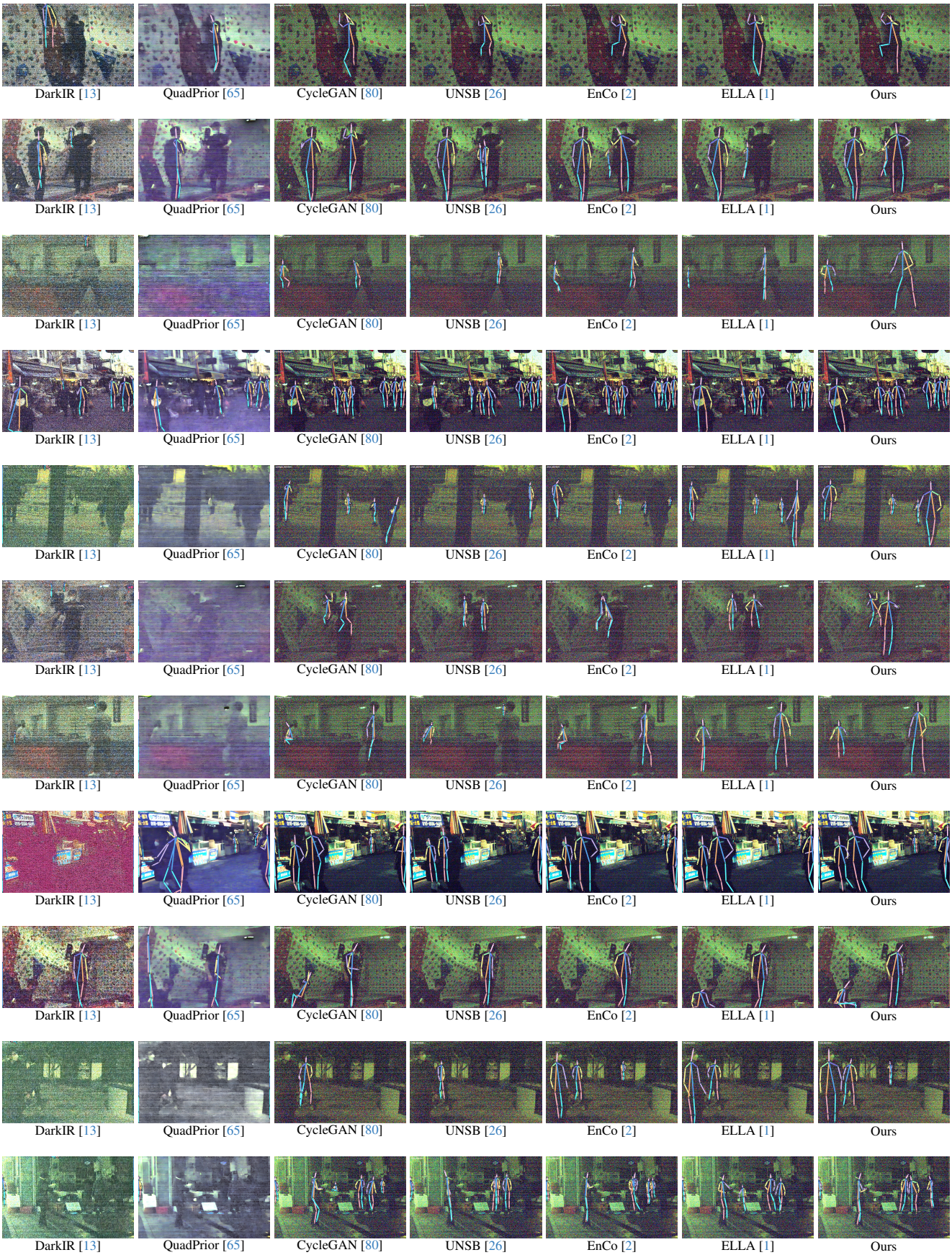
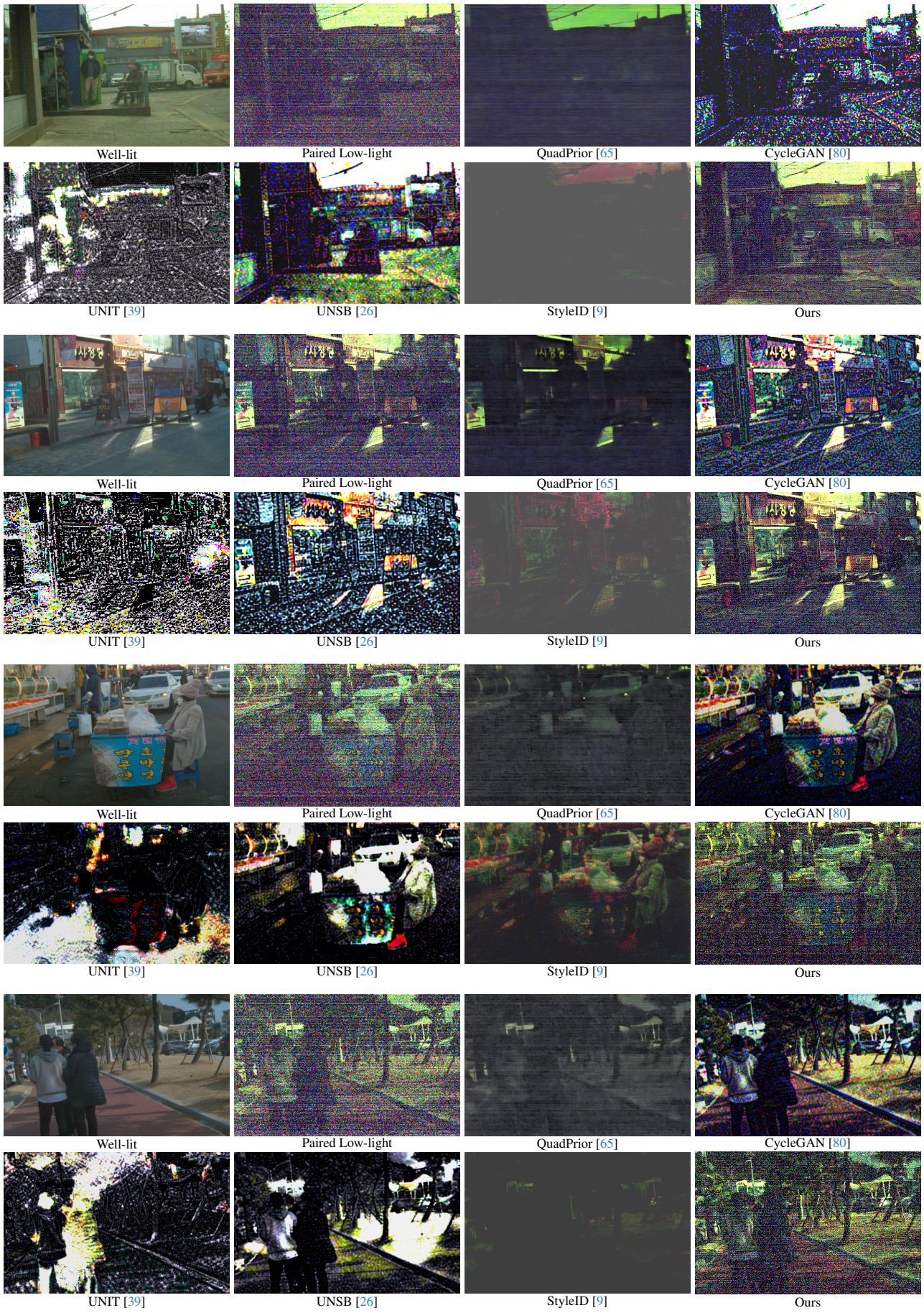
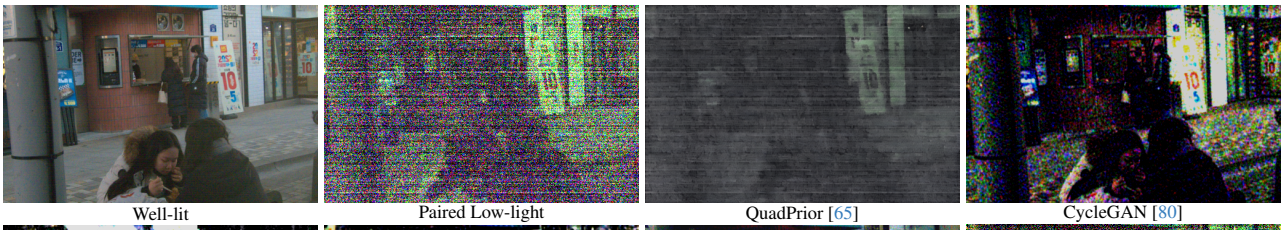


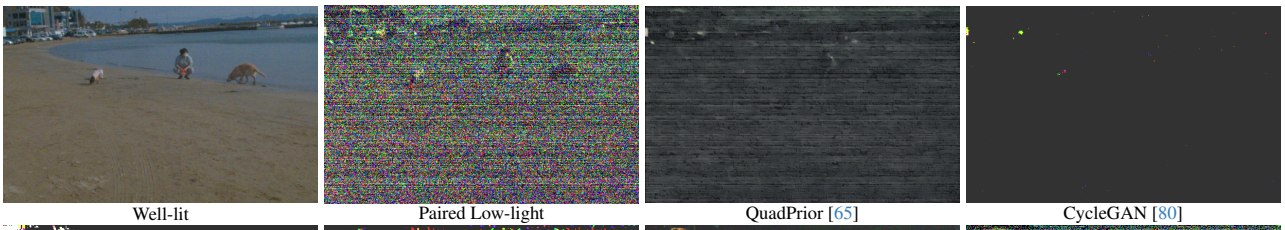
Figure 10. Qualitative comparison of our data synthesis method with baselines.





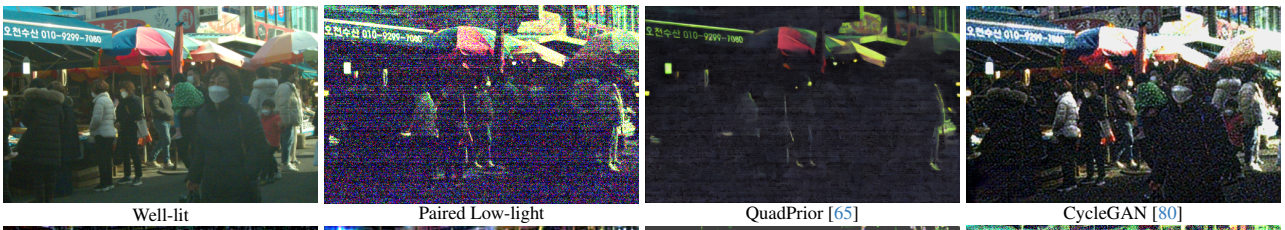
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours



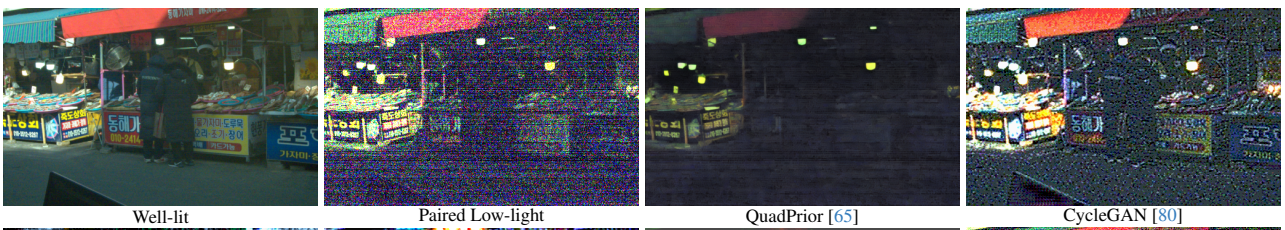
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours



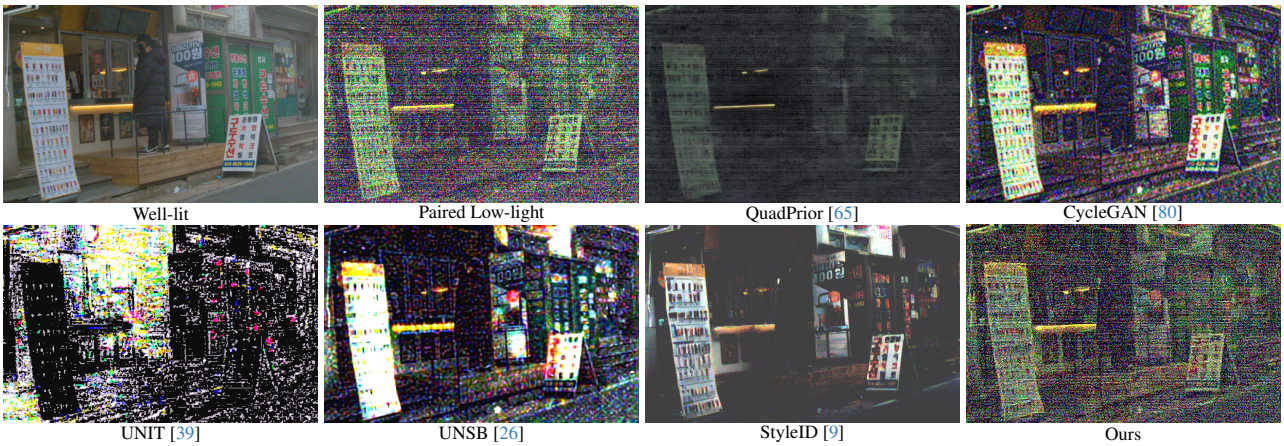
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours

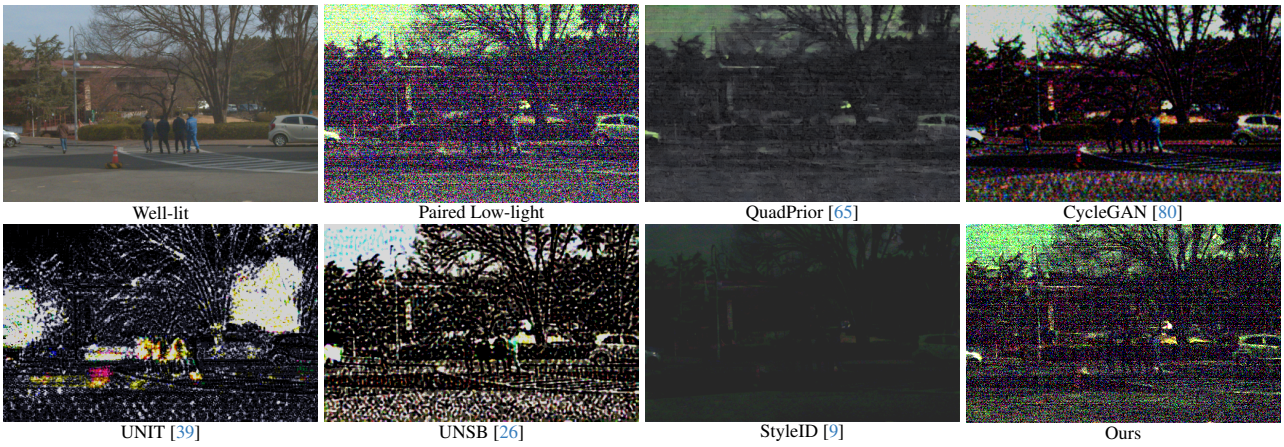


Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

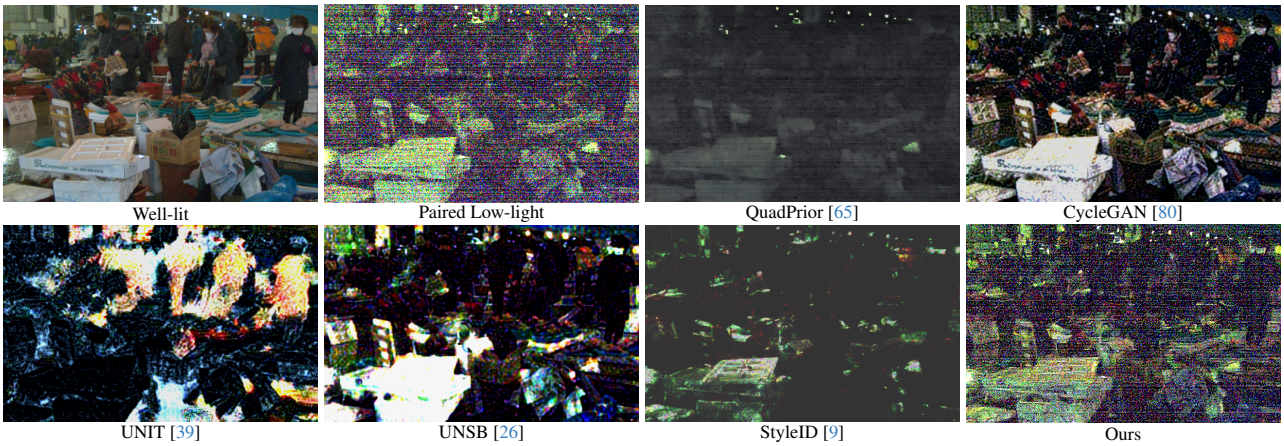
UNIT [39] | UNSB [26] | StyleID [9] | Ours



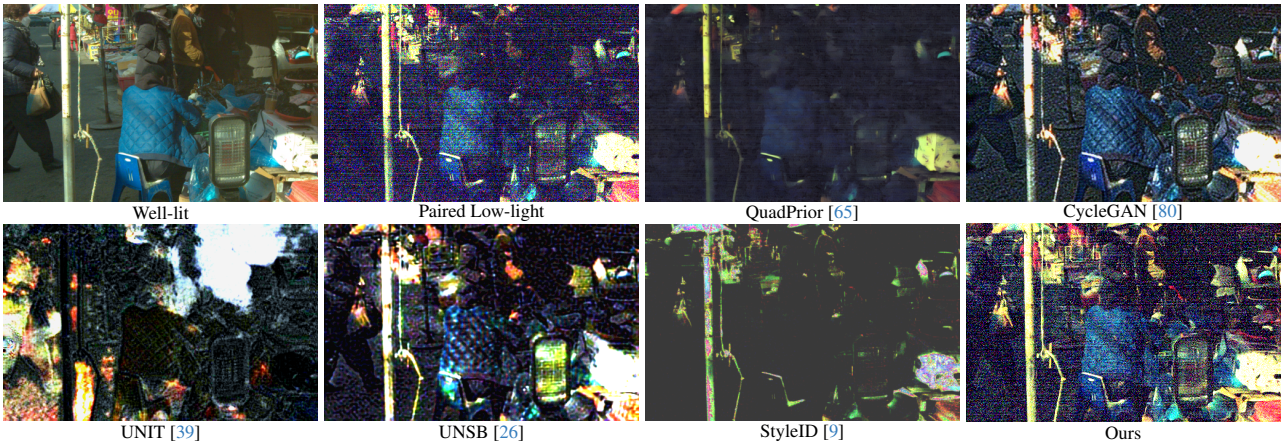
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]
UNIT [39] | UNSB [26] | StyleID [9] | Ours



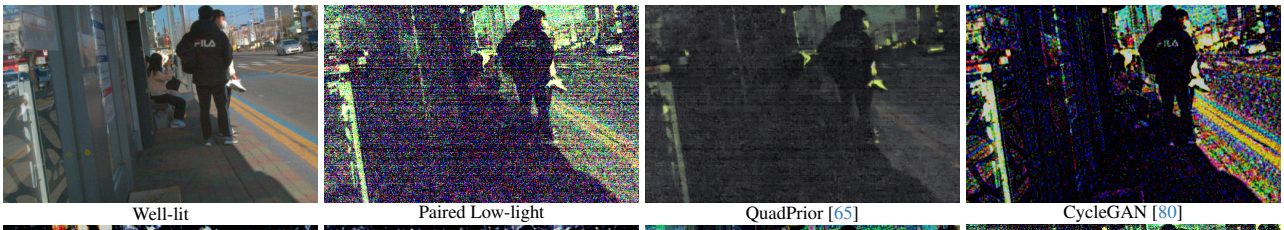
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]
UNIT [39] | UNSB [26] | StyleID [9] | Ours



Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]
UNIT [39] | UNSB [26] | StyleID [9] | Ours

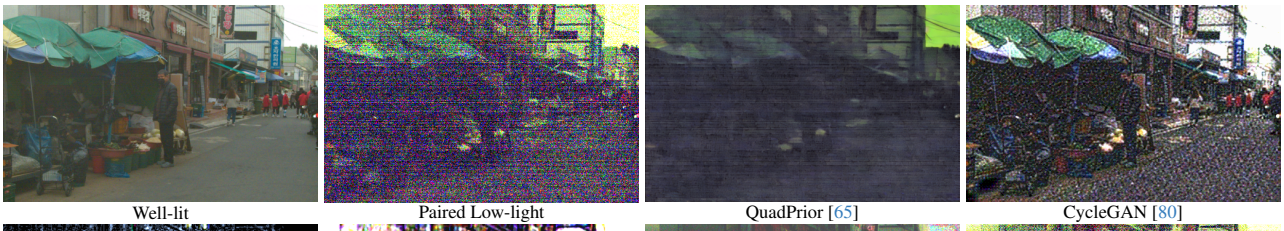


Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]
UNIT [39] | UNSB [26] | StyleID [9] | Ours



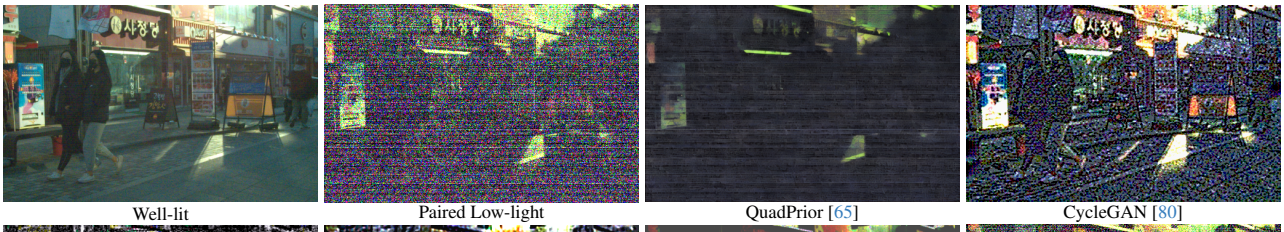
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours



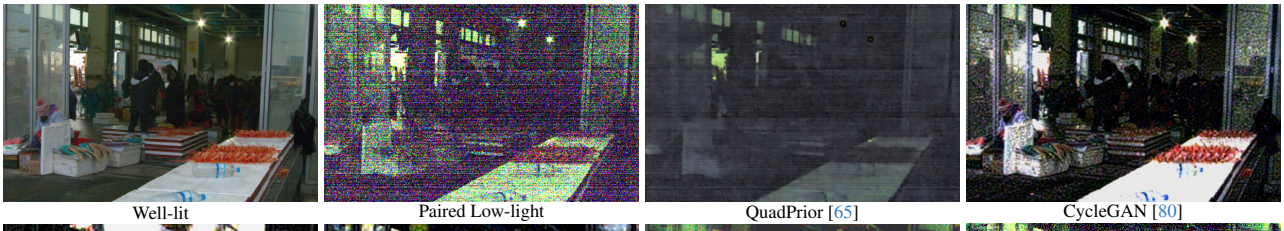
Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours



Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours



Well-lit | Paired Low-light | QuadPrior [65] | CycleGAN [80]

UNIT [39] | UNSB [26] | StyleID [9] | Ours

In contrast, our method generates low-light images that exhibit low-light noise characteristics. The synthetic images produced by our method preserve high-frequency details while modeling complex noise patterns observed in low-light conditions. The LCIM module is key to the superior quality of our synthetic low-light images, as it effectively captures and transfers complex low-light characteristics from unpaired real low-light images to well-lit inputs. As a result, our UDAPose overcomes the limitations of existing approaches, generating more effective training data that better prepares the pose estimation model for low-light scenarios.

15.3. Comparison of DCA

We provide a qualitative comparison to demonstrate the effectiveness of our DCA module. Without DCA, the model tends to assign uniformly high importance to image cues for all keypoints, as indicated by the consistently high values in the "Before" columns. This forces the model to overly rely on visual evidence, even when it is corrupted by noise or low visibility. Consequently, this leads to erroneous human pose predictions as shown in Fig. 8. Our DCA module effectively resolves this issue by learning to dynamically balance the influence of image cues and pose priors. As shown in the "After" columns, DCA significantly reduces the cue weights for keypoints with low visibility. By down-weighting these unreliable signals, the model can leverage its learned pose priors for improved pose estimation. This results in substantially more accurate and coherent poses, correcting the initial errors and demonstrating that DCA is crucial for achieving robustness in challenging, low-visibility conditions.

16. Limitations

While our results are promising, there are still opportunities to build on this work in future research. The current framework, including the proposed LCIM, DHF, and DCA modules, is specifically tailored to model degradations from insufficient illumination, primarily by transferring noise characteristics and balancing unreliable visual cues with learned pose priors. A promising future direction is to extend this generative approach to handle a broader spectrum of low-visibility scenarios, such as dense fog, heavy rain, or severe motion blur. This would likely require designing new modules capable of synthesizing these more complex degradations, thereby enhancing the model's generalization to diverse and challenging real-world conditions.

The reliance on a large-scale diffusion model like SD introduces substantial computational overhead. The data synthesis pipeline is resource-intensive, requiring significant GPU memory and time for generating the training dataset. This presents a practical barrier to rapid adaptation for new, custom low-light environments. Future work could explore the use of more efficient generative models, such as consistency models or distilled diffusion models, to mitigate this cost and improve accessibility.