DarkAct: A RGB-Thermal Dataset and Fusion Framework for Multimodal Low-Light Action Recognition

Supplementary Material

7. Technical Details of DarkAct-Net

We provide additional technical details of DarkAct-Net that were omitted from the main manuscript due to space limitations. Please refer to Figure 3 of the main manuscript for better clarity.

7.1. Modal Embedding

The Modal Embedding module processes raw RGB or thermal video inputs and extracts modality-specific spatiotemporal features for downstream processing. Given an input video tensor $X \in \mathbb{R}^{T\times C\times H\times W}$, where T denotes the number of frames, C is the channel dimension ($C = 3$ for RGB and $C = 1$ for thermal), and H, W denote spatial resolution, we apply three parallel convolutional branches with different receptive fields:

$$F_1 = \mathrm{Conv}_{3\times3}^{\mathrm{ReLU+BN}}(X),$$

$$F_2 = \mathrm{Conv}_{5\times5}^{\mathrm{ReLU+BN}}(X),$$

$$F_3 = \mathrm{Conv}_{7\times7}^{\mathrm{ReLU+BN}}(X),$$

These branches capture complementary multi-scale spatial cues. We then construct two pairwise concatenations:

$$F_{cat}^1 = [F_1; F_2] \in \mathbb{R}^{T\times2C\times H\times W},$$

$$F_{cat}^2 = [F_2; F_3] \in \mathbb{R}^{T\times2C\times H\times W}.$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation.

Afterwards, each concatenated feature tensor is fed into a lightweight MLP-style refinement block consisting of a 3×3 convolution with Sigmoid activation, followed by a dilated 7×7 convolution with ReLU activation and batch normalization. Formally,

$$F^1 = \mathrm{DilConv}_{7\times7}^{\mathrm{ReLU+BN}}\big(\sigma(\mathrm{Conv}_{3\times3}(F_{\mathrm{cat}}^1))\big),$$

$$F^2 = \mathrm{DilConv}_{7\times7}^{\mathrm{ReLU+BN}}\big(\sigma(\mathrm{Conv}_{3\times3}(F_{\mathrm{cat}}^2))\big),$$

where $\sigma(\cdot)$ denotes the Sigmoid activation.

The resulting multi-scale features are aggregated via channel concatenation:

$$F_{\mathrm{multi}} = [F^1; F^2] \in \mathbb{R}^{T\times2C\times H\times W},$$

A 1×1 convolution with ReLU+BN is applied to fuse the features and reduce the channel dimension:

$$F_{\mathrm{refined}} = \mathrm{Conv}_{1\times1}^{\mathrm{ReLU+BN}}(F_{\mathrm{multi}}) \in \mathbb{R}^{T\times C\times H\times W}$$

To incorporate temporal ordering and modality-aware contextual cues, we introduce a learnable spatiotemporal positional embedding: $P \in \mathbb{R}^{T\times C\times H\times W}$. The final modal-specific representation is obtained via element-wise addition:

$$G = F_{\mathrm{refined}} + P,$$

where $G \in \mathbb{R}^{T\times C\times H\times W}$.

This representation integrates multi-scale spatial cues, temporal positional information, and modality awareness, serving as the unified input to subsequent modules (MAA and LAF) within DarkAct-Net.

7.2. Light-Adaptive Query

The Light-Adaptive Query module produces the query vector defined in Eq. (5) of the main manuscript. It is computed as:

$$F^Q = \mathrm{FC}\left(\mathrm{AP}\left(\mathrm{DilC}_{3,7\times7}^{\mathrm{PReLu}}\big([\,Y^r \odot F^r\,;\, Y^t \odot F^t\,]\big)\right)\right),$$

where the operations are defined as follows: $F^r \in \mathbb{R}^{C\times H\times W}$ and $F^t \in \mathbb{R}^{C\times H\times W}$ are the RGB and thermal features produced at each Transformer stage; Y^r and Y^t are the outputs of the MAA module for RGB and thermal modalities, respectively. Both are processed by a downsampling convolution to match the spatial resolution of F^r and F^t; $\odot$ denotes element-wise multiplication, enforcing modality reliability weighting guided by illumination-aware saliency; $\mathrm{DilC}_{3,7\times7}^{P}(\cdot)$ is a dilated 7×7 convolution with dilation rate 3, followed by a PReLU activation; $[\cdot;\cdot]$ represents channel-wise concatenation, resulting in a feature of size $[F^r; F^t] \in \mathbb{R}^{2C\times H\times W}$; $\mathrm{AP}(\cdot)$ denotes global average pooling, reducing the spatial dimensions to produce a compact descriptor in $\mathbb{R}^{2C}$; $\mathrm{FC}(\cdot)$ is a fully connected layer that maps the pooled feature to $F^Q \in \mathbb{R}^{C\times N}$, which serves as the illumination-adaptive query for cross-attention within the LAF module.

8. Prompts for Testing VLMs

We adopts following user prompts for testing zero-shot performance of VLMs as shown in Table 3 of the manuscript.

- Single-modal setting for either RGB and thermal: *"There is a video with one person in scene, please analyse the human action in video and choose exactly one action category from the following 27 classes:[calling, climbing stairs, crouching, closing door, opening door, drinking,*

holding umbrella, jumping, lifting, photographing, picking, pouring, pushing, taking off coat, putting on coat, running, searching, sitting, slapping, sleeping, squatting, standing, turning, typing, using smartphone, walking, waving]. Your answer must contain only one category name, without explanations or punctuation".

- Multimodal setting: *"There are RGB and thermal videos, please analyse the human action in these video and choose exactly one action category from the following 27 classes:[calling, climbing stairs, crouching, closing door, opening door, drinking, holding umbrella, jumping, lifting, photographing, picking, pouring, pushing, taking off coat, putting on coat, running, searching, sitting, slapping, sleeping, squatting, standing, turning, typing, using smartphone, walking, waving]. Your answer must contain only one category name, without explanations or punctuation."*

9. Additional Experimental Analysis

9.1. Class-wise Performance

Table A1 reports the per-class Top-1 and Top-5 accuracies of DarkAct-Net, corresponding to the results summarized in Table 2 of the main manuscript. While DarkAct-Net achieves the best overall performance among all evaluated models, its recognition accuracy exhibits noticeable variation across action categories. This variability arises from several factors: *First*, class imbalance in DarkAct (as described in Section 3.2 of the main manuscript), where frequent actions such as *calling* contain substantially more samples than infrequent ones like *taking off coat*. *Second*, intrinsic difficulty of certain actions under low-light conditions, where subtle or fast motion (e.g., *jumping*, *typing*, *photographing*) leads to weaker appearance cues, especially for small or distant subjects. *Third*, high inter-class similarity, such as between *opening/closing door* or *pushing/picking*, which becomes even more challenging under illumination degradation.

Despite these challenges, DarkAct-Net maintains strong Top-5 accuracy across almost all categories, demonstrating its robustness in capturing motion and cross-modal cues. These observations highlight opportunities for future work, such as class-balanced training strategies, improved temporal modeling, or advanced cross-modal enhancement tailored for low-light environments.

Classes	DarkAct-Net	
	Top-1	Top-5
Calling	90.4	100.0
Climbing stairs	100.0	100.0
Crouching	79.2	96.1
Closing door	86.3	100.0
Opening door	89.9	100.0
Drinking	55.3	71.1
Holding umbrella	98.2	100.0
Jumping	39.3	60.1
Lifting	72.7	91.9
Photographing	48.1	55.8
Picking	93.4	98.4
Pouring	96.6	99.0
Pushing	66.7	91.3
Taking off coat	28.6	92.0
Putting on coat	90.0	98.0
Running	75.1	98.9
Searching	95.4	98.7
Sitting	73.6	93.4
Slapping	100.0	100.0
Sleeping	100.0	100.0
Squatting	85.5	99.2
Standing	95.7	100.0
Turning	92.1	100.0
Typing	49.7	61.9
Using smartphone	97.4	100.0
Walking	78.2	81.9
Waving	76.1	88.1

Table A1. Per-class Top-1 and Top-5 accuracy of DarkAct-Net on the DarkAct test set. The results correspond to the overall performance reported in Table 2 of the main manuscript.

9.2. Visualization of MAA Module

To better understand the behavior of the Motion-Aware Attention (MAA) module, we visualize its intermediate representations for paired RGB and thermal videos. Figure A1 presents examples from the DarkAct dataset. For each modality, we show (from left to right): the raw video frame, the computed temporal saliency map, and the motion-enhanced features Y^r or Y^t.

The visualizations reveal that the saliency maps for both modalities successfully highlight foreground human motion, allowing the model to suppress background noise that is common in low-light environments. Due to the inherent sensing differences between RGB and thermal cameras, the saliency maps exhibit distinct patterns: RGB emphasizes texture and edges, while thermal accentuates body heat contours. These differences provide cross-spectral complementarity, enabling the model to capture motion cues that are difficult to observe in one modality alone.

Consequently, the motion-enhanced features Y^r and Y^t encode consistent foreground motion while preserving modality-specific strengths. This facilitates more effective fusion in subsequent stages, ultimately leading to improved robustness and recognition accuracy under varying illumination conditions.

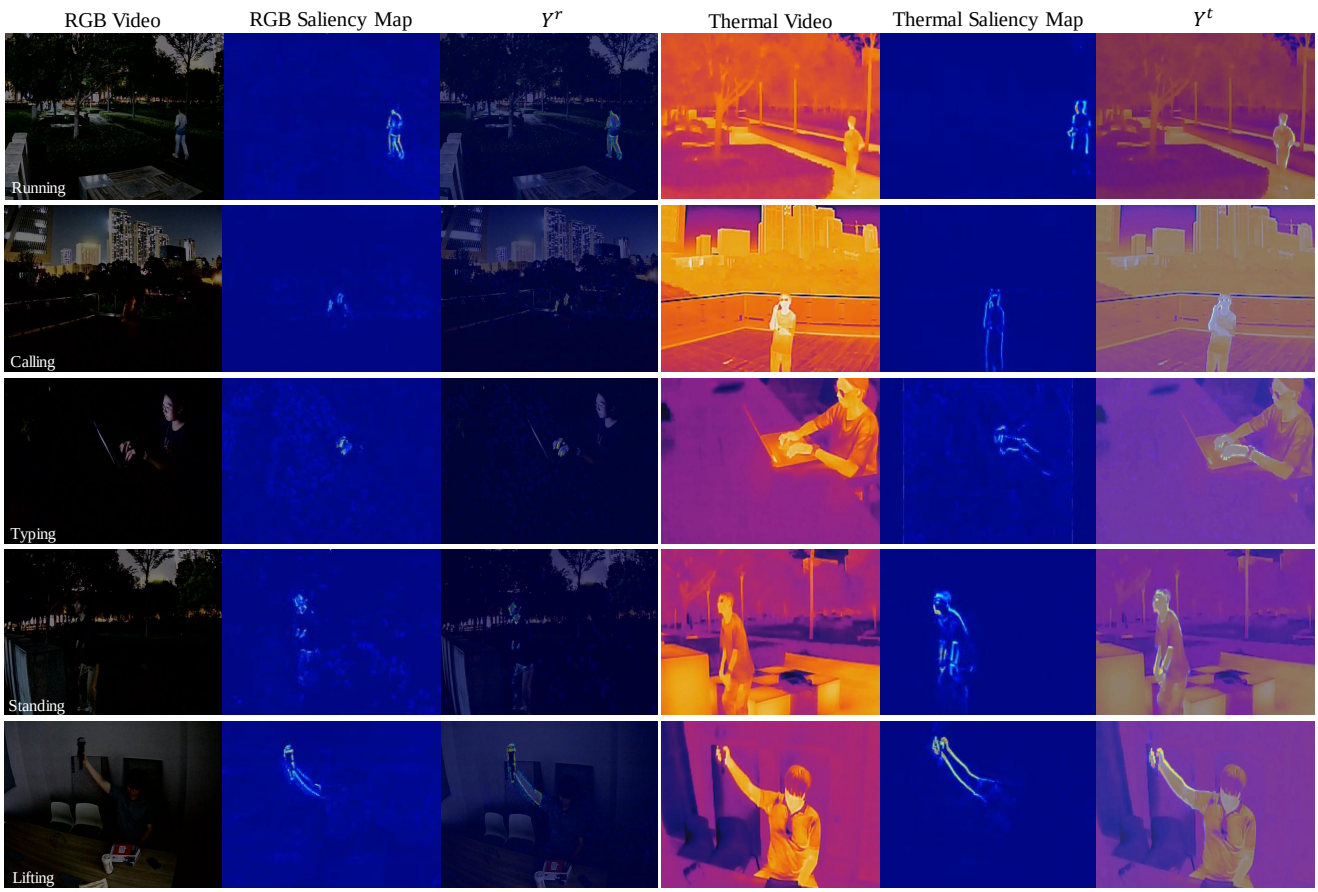


Figure A1. Visualization of the MAA module on paired RGB–thermal videos. For each modality, we show (left to right): the input frame, the temporal saliency map, and the motion-enhanced feature Y^r or Y^t.

9.3. Confusion matrix analysis

The confusion matrix provides a detailed view of how different models misclassify visually similar or motion-related action categories, offering deeper insight beyond aggregate Top-1/Top-5 metrics.

To further analyze recognition behavior on the 27 DarkAct classes, we present confusion matrices for several state-of-the-art methods. Specifically, we select the best-performing models from Table 2 of the main manuscript, including: (1) multimodal methods including DarkAct-Net, MRFS, and CMX; and (2) unimodal methods including MViT, OverLoCK, and ConvNeXt, evaluated separately on RGB and thermal inputs.

Figure A2 compares the confusion matrices of these models. DarkAct-Net exhibits a substantially stronger diagonal response, indicating more confident and accurate predictions across most categories. In the off-diagonal regions, DarkAct-Net consistently produces lower values, demonstrating fewer cross-class confusions. These improvements are particularly notable for actions that are highly ambiguous in low-light environments (e.g.,*taking off coat*), where RGB-only or thermal-only models struggle. Overall, the confusion matrices highlight the effectiveness of DarkAct-Net's multimodal design for more reliable recognition under challenging low-light conditions.

10. More Examples of DarkAct

To further illustrate the diversity and visual characteristics of the DarkAct dataset, we present additional video frame samples for all 27 action categories at the end of this appendix. These examples highlight variations in illumination levels, human–camera distances, viewpoints, and scene contexts across both RGB and thermal modalities. The expanded set of visual examples provides a more comprehensive overview of the challenges presented by DarkAct and demonstrates the richness of cross-spectral cues available for low-light action recognition.

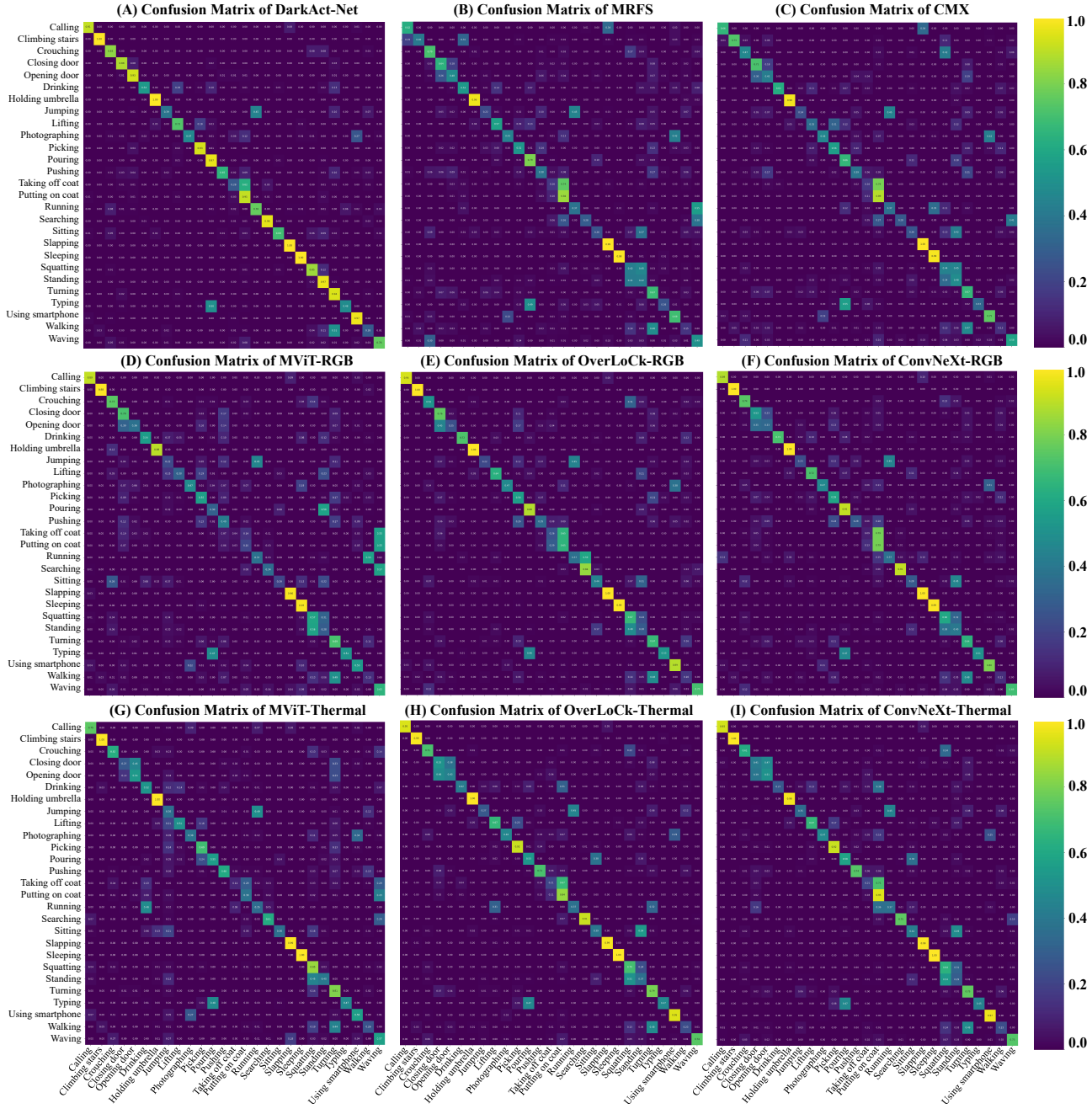
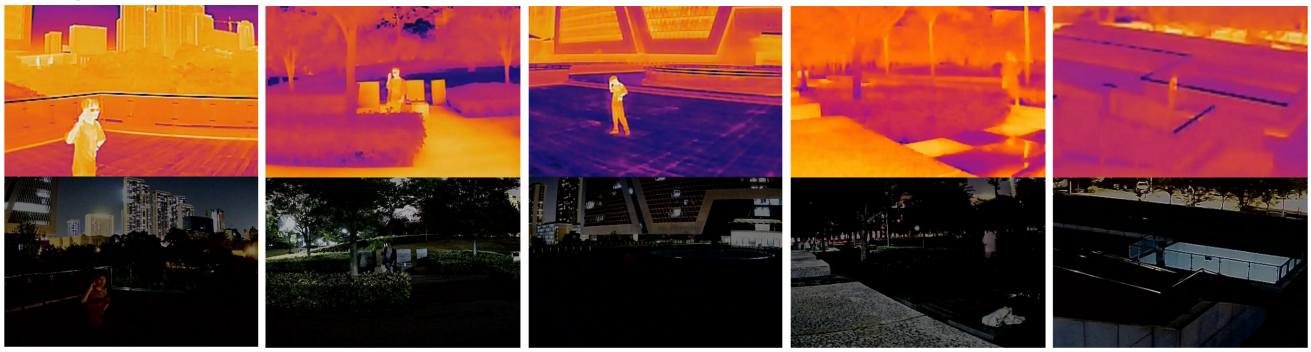
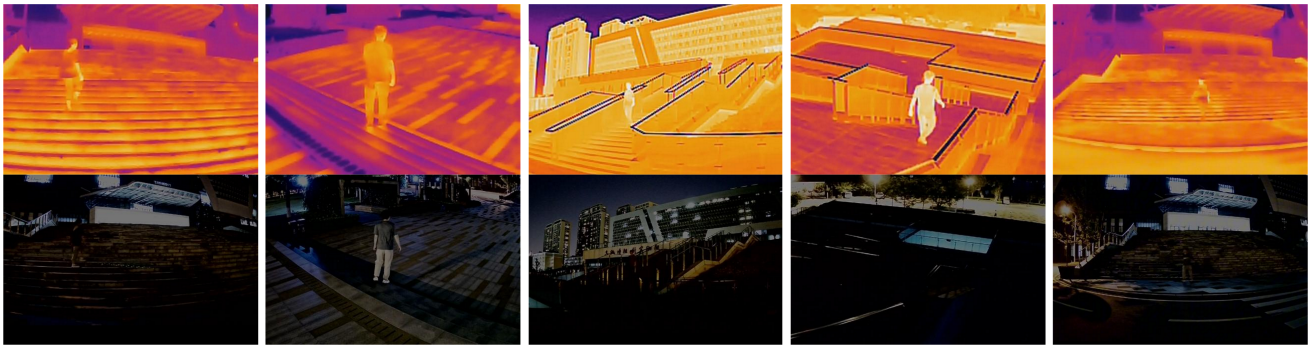


Figure A2. Comparison of confusion matrices for state-of-the-art unimodal and multimodal models on the 27-class DarkAct dataset.

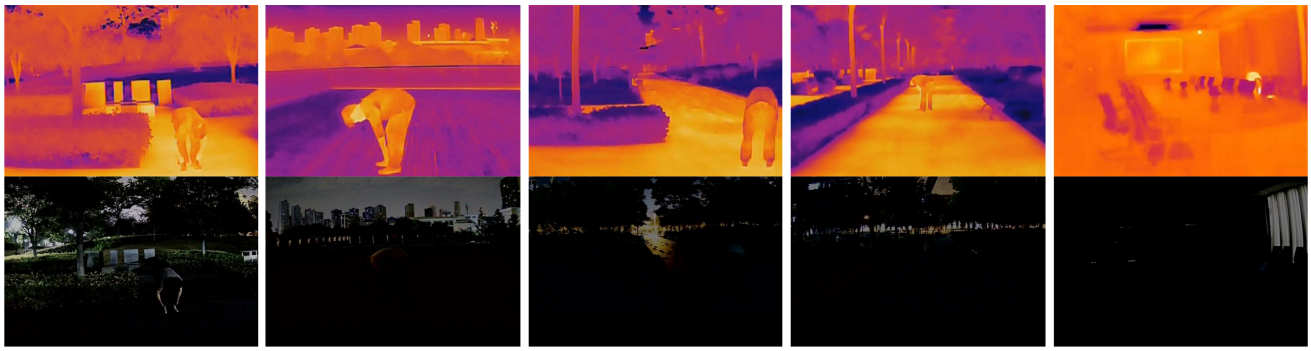
Calling



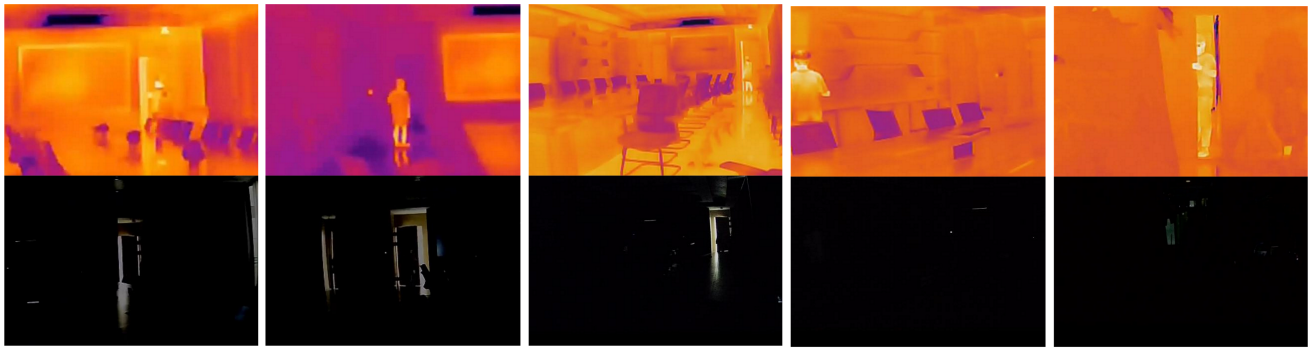
Climbing stairs



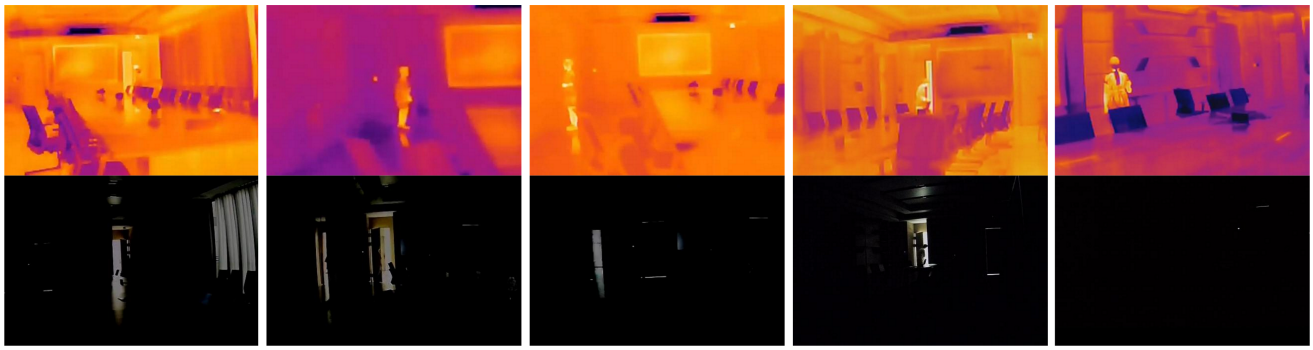
Crouching



Closing door

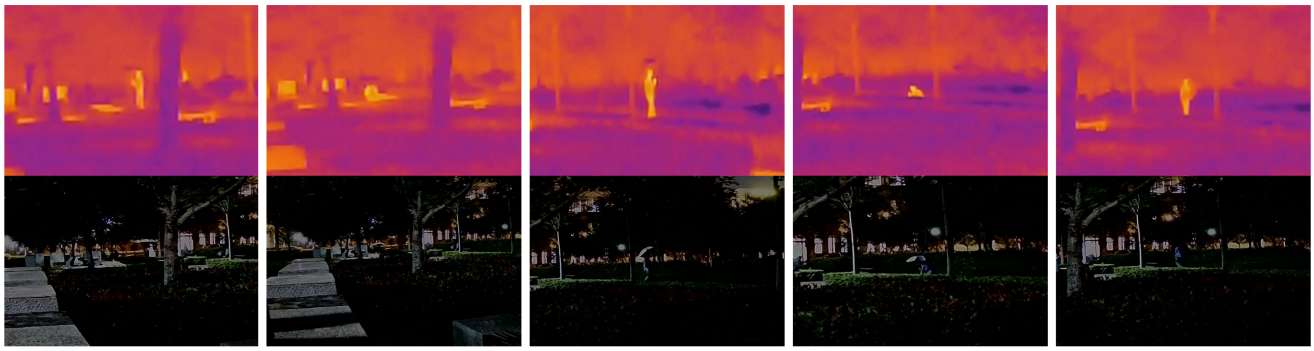


Opening door

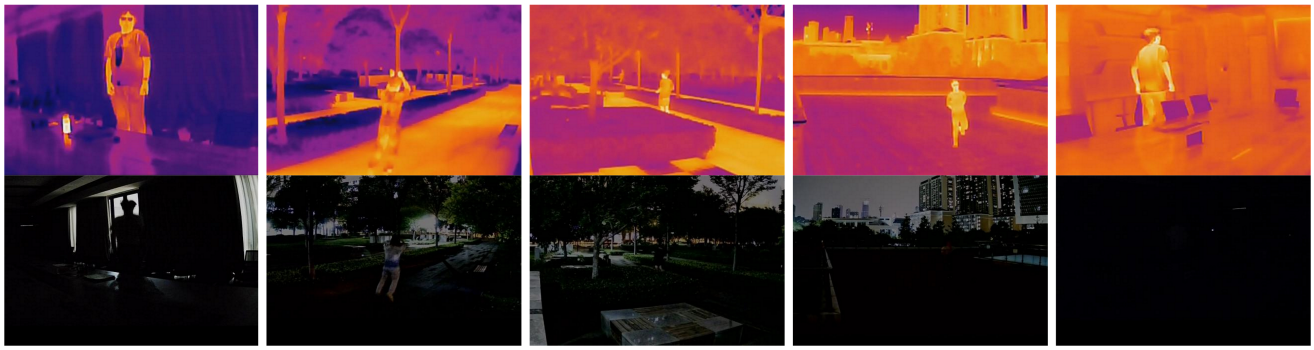


Drinking

Holding umbrella

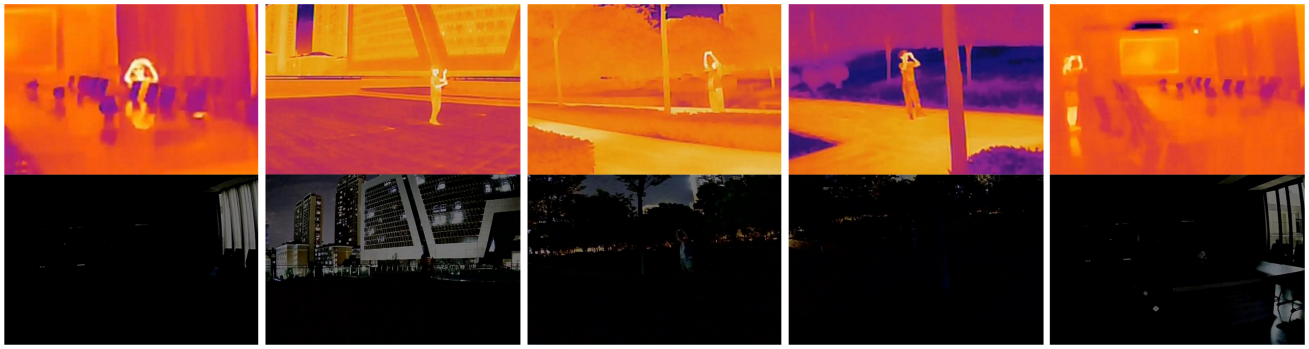


Jumping

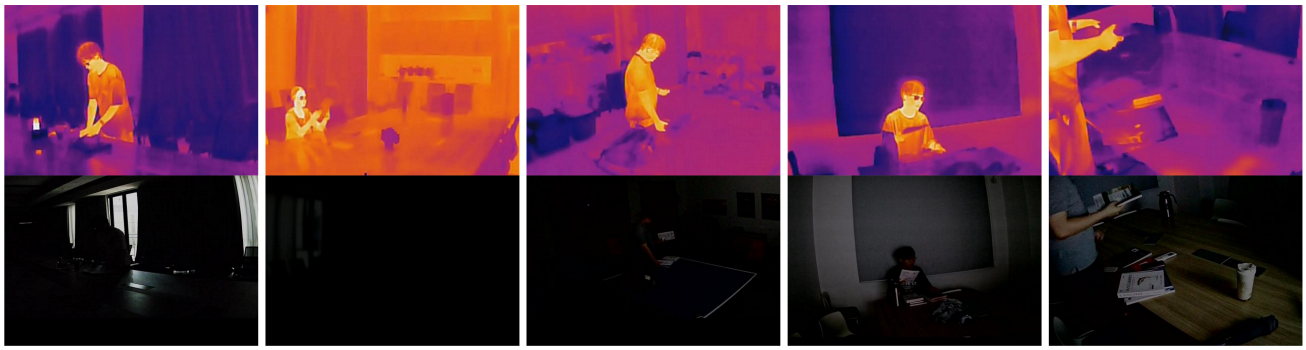


Lifting

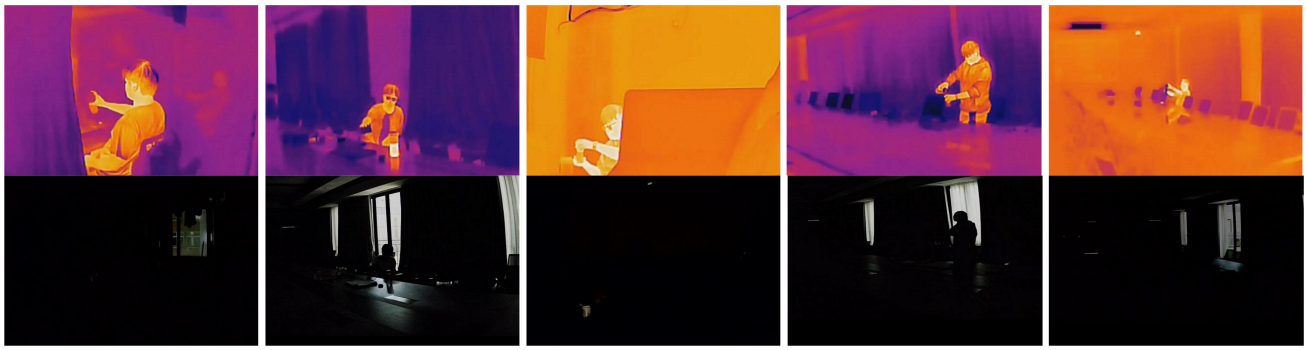
Photographing



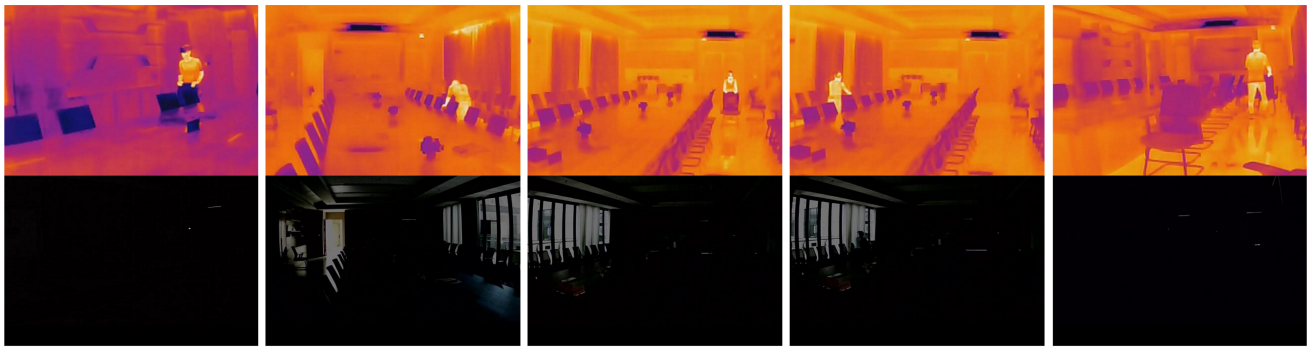
Picking



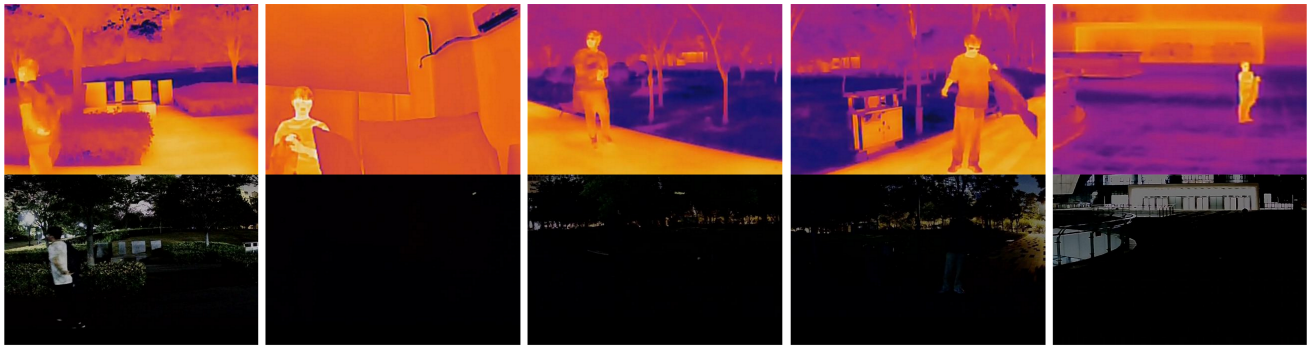
Pouring



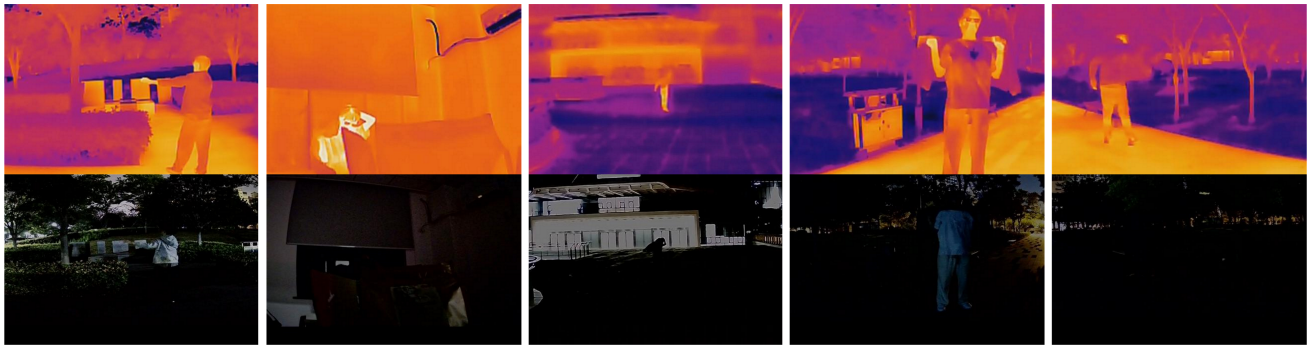
Pushing



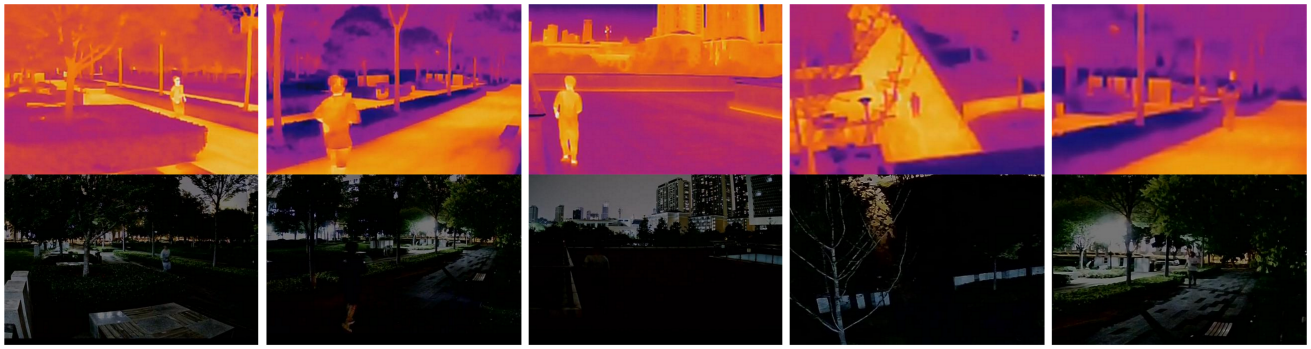
Taking off coat



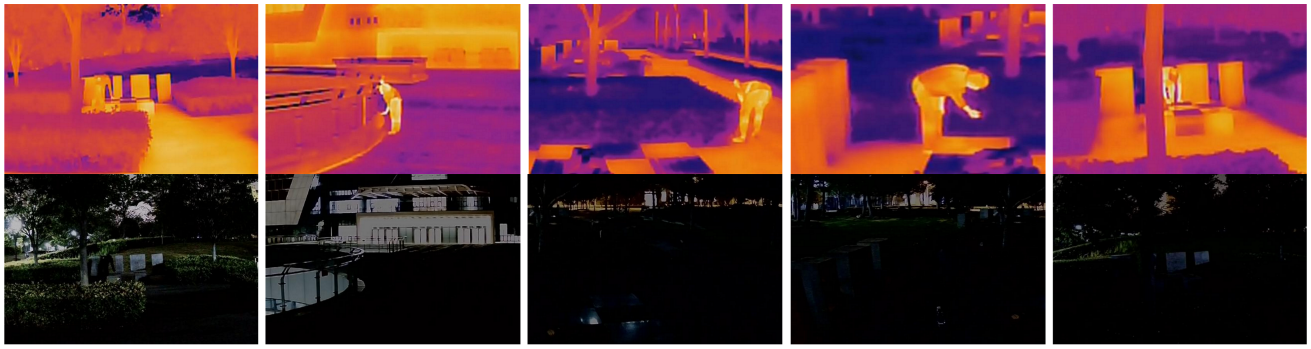
Putting on coat



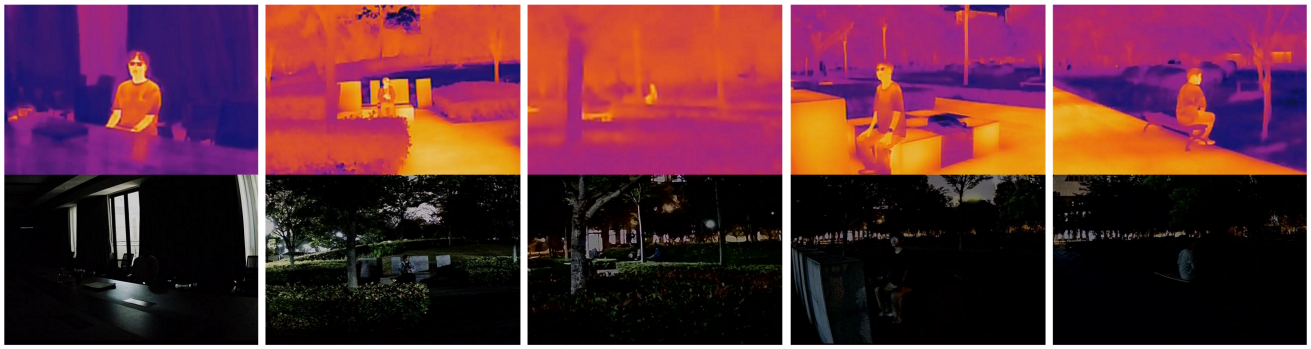
Running



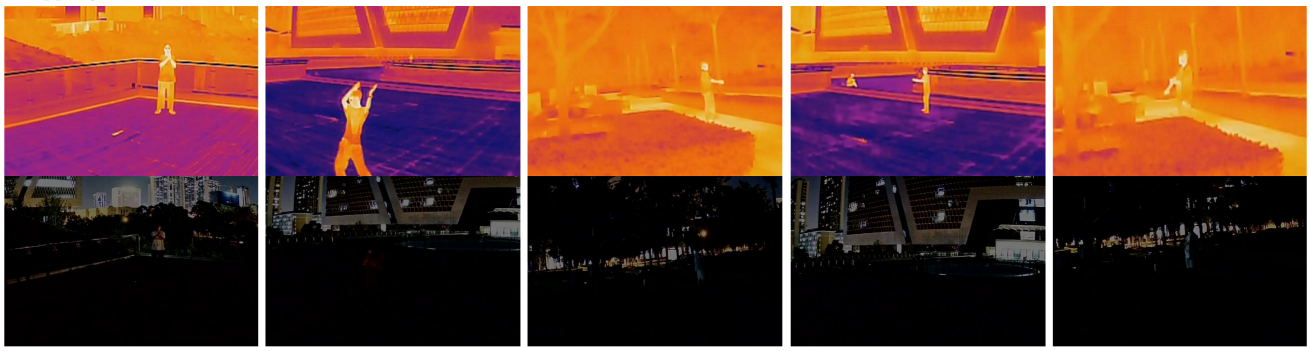
Searching



Sitting

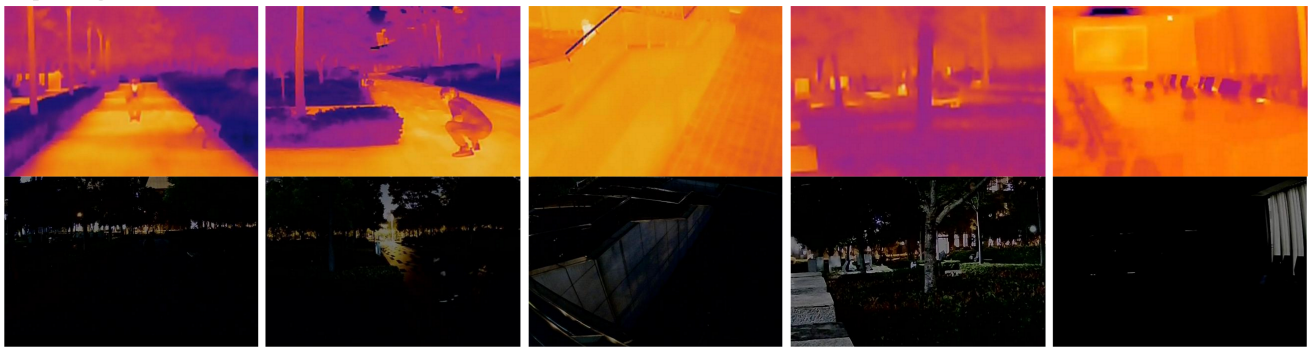


Slapping



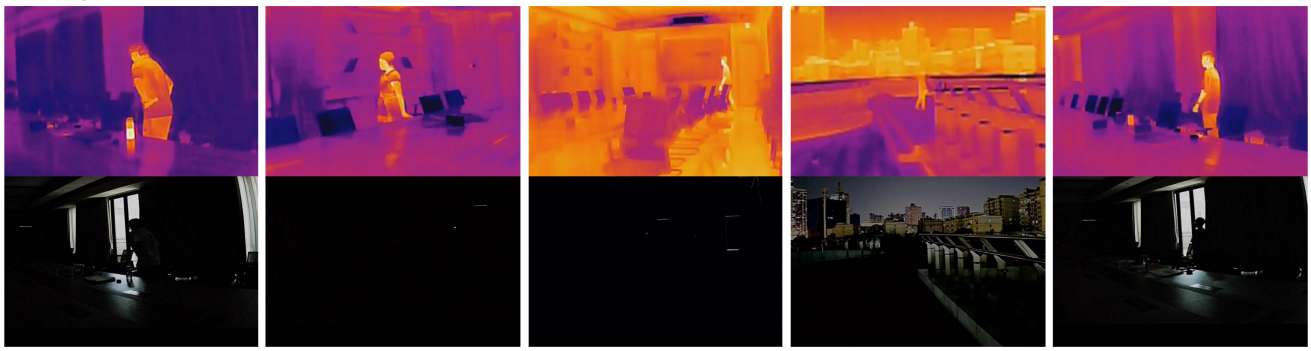
Sleeping

Squatting



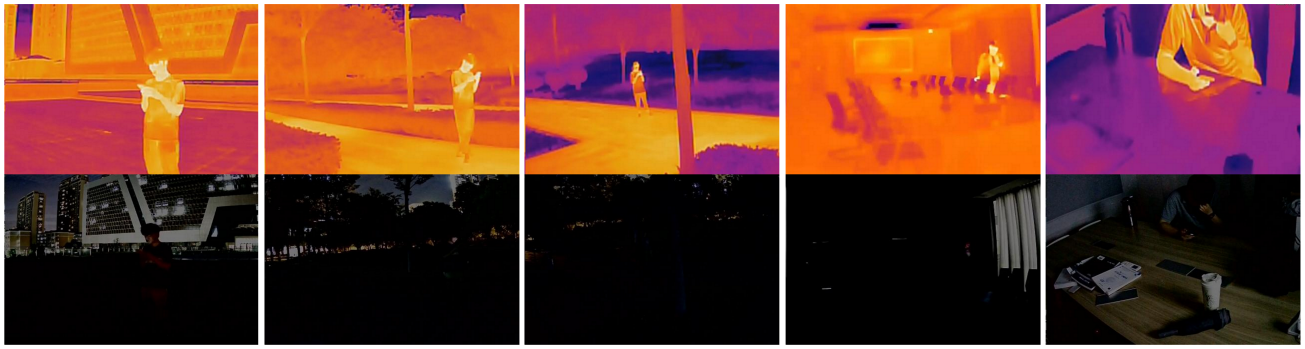
Standing

Turning

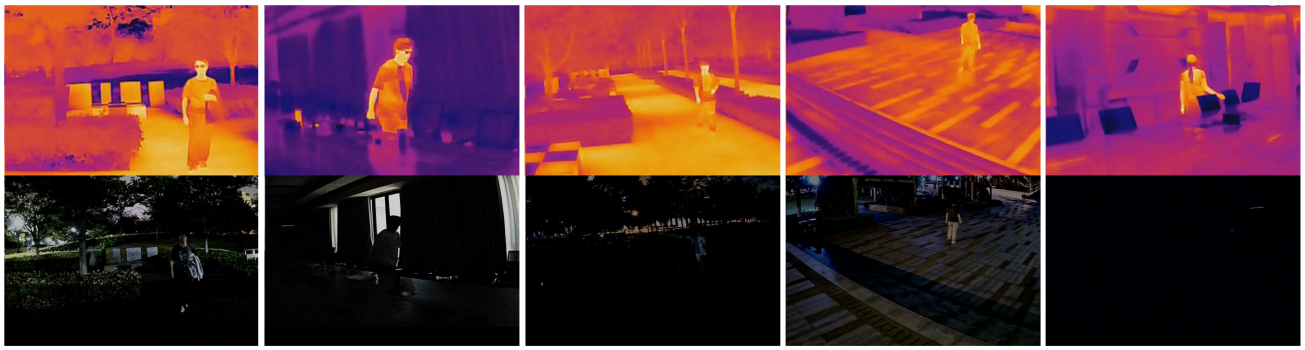


Typing

Using smartphone



Walking



Waving