# Leveraging Thermal Imaging for Robust Human Pose Estimation in Low-Light Vision Supplemental Material

Mickael Cormier[1,3,4], Caleb Ng Zhi Yi[1], Andreas Specker[1,4], Benjamin Blaß[2], Michael Heizmann[3,1,4], and Jürgen Beyerer[3,1,4]

[1] Fraunhofer IOSB, Germany, {firstname.lastname}@iosb.fraunhofer.de
[2] Stahl-Holding-Saar, Germany, benjamin.blass@stahl-holding-saar.de
[3] Karlsruhe Institute of Technology, Germany, {firstname.lastname}@kit.edu
[4] Fraunhofer Center for Machine Learning, Germany

**Table 1:** An overview of SOTA models for benchmarking experiments. All models are trained with a batch size of 16 except for DEKR with a batch size of 10.

| Models | Backbone | Input Size | Type | Architecture | Epoch |
|---|---|---|---|---|---|
| Top-down | | | | | |
| HRNetw48-udp [8] | HRNet-w48 [8] | $256 \times 192$ | HM | CNN | 210 |
| ViTPose-h [9] | ViTAE-G [10] | $256 \times 192$ | HM | Transformer | 210 |
| DeepPose-r50 [6] | ResNet-50 [2] | $256 \times 192$ | R. | CNN | 210 |
| SimCC [3] | ResNet-50 [2] | $256 \times 192$ | CC | CNN | 210 |
| Bottom-up | | | | | |
| DEKR [1] | HRNet-w32 [8] | $512 \times 512$ | HM + R. | CNN | 140 |
| One-stage | | | | | |
| YOLOX-Pose-l [5] | CSPDarknet [7] | $600 \times 600$ | R. | CNN | 300 |
| RTMO-l [4] | CSPDarknet [7] | $600 \times 600$ | CC | CNN | 600 |

## 1 Implementation Details

For the top-down methods, we apply image-level random flipping (horizontal), random half body, random scaling ($[0.5, 1.5]$), random rotation ($[-80°, 80°]$) and random translation ($[-0.16 \cdot bbox_w, 0.16 \cdot bbox_w]$) and ($[-0.16 \cdot bbox_h, 0.16 \cdot bbox_h]$) on the groundtruth bounding boxes. DEKR, our bottom-up approach, employ random image flipping (horizontal), random image shifting($[-0.2 \cdot img_w, 0.2 \cdot img_w]$) and ($[-0.2 \cdot img_h, 0.2 \cdot img_h]$), random resizing ($[0.75, 1.5]$) and rotating ($[-40°, 40°]$). For the single-stage methods, we apply random image flipping (horizontal), random image shifting ($[-0.1 \cdot img_w, 0.1 \cdot img_w]$) and ($[-0.1 \cdot img_h, 0.1 \cdot img_h]$), random resizing ($[0.75, 1.0]$) and rotating ($[-10°, 10°]$), mixup, mosaic augmentation and sequential HSV color transformations.

**Table 2:** Experiments for augmentation

| Experiments | Grayscale | HSV<br>H:[-100,100] \| S:[-30,100] \| V:[-20,100] | Invert |
|---|---|---|---|
| RGB | ✗ | ✗ | ✗ |
| Grayscale | ✓ | ✗ | ✗ |
| RGB + Grayscale | ✓($p = 0.5$) | ✗ | ✗ |
| RGB + Augmentations | ✗ | ✓ | ✓ |
| Grayscale + Augmentations | ✓ | ✓ | ✓ |

**Table 3:** Bounding boxes size split for the LLVIP-Pose.

| Set | Bounding Boxes | | |
|---|---|---|---|
| | **Small** | **Medium** | **Large** |
| Train | 0 | 552 | 18,081 |
| Test | 0 | 38 | 7,464 |

**Table 4:** Crowd index split for LLVIP-Pose. The number of image pairs and poses are provided "easy", "medium" and "hard".

| Set | **Easy** | | **Medium** | | **Hard** | |
|---|---|---|---|---|---|---|
| | Image Pairs | Poses | Image Pairs | Poses | Image Pairs | Poses |
| Train | 5,795 | 14,328 | 992 | 4,109 | 66 | 196 |
| Test | 3,049 | 5,918 | 405 | 1,562 | 8 | 22 |

## References


1. 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)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., Xia, S.T.: Simcc: A simple coordinate classification perspective for human pose estimation. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
4. Lu, P., Jiang, T., Li, Y., Li, X., Chen, K., Yang, W.: Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. arXiv preprint arXiv:2312.07526 (2023)
5. Maji, D., Nagori, S., Mathew, M., Poddar, D.: Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2637–2646 (2022)
6. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)
7. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 390–391 (2020)
8. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)

9. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
10. Xu, Y., Zhang, Q., Zhang, J., Tao, D.: Vitae: Vision transformer advanced by exploring intrinsic inductive bias. Advances in neural information processing systems **34**, 28522–28535 (2021)