Supplementary Materials for: Sparse and Privacy-enhanced Representation for Human Pose Estimation

Ting-Ying Lin[1*]
tingyinglin@m111.nthu.edu.tw
Lin-Yung Hsieh[1*]
linyunghsieh@gapp.nthu.edu.tw
Fu-En Wang[1]
fulton84717@gapp.nthu.edu.tw
Wen-Shen Wuen[2]
vincent_wuen@novatek.com.tw
Min Sun[1]
sunmin@ee.nthu.edu.tw

[1] Vision Science Lab
National Tsing Hua University
Hsinchu, Taiwan

[2] Novatek Microelectronics Corp.
Hsinchu, Taiwan

1 Further information on Human Pose Estimation

1.1 Implementation Details

As mentioned in Section 5.1 of the main paper, we conduct tests using three CNN backbones: the DHP19 [1] proposed model (218K), U-Net-Small (1.9M), and U-Net-Large (7.7M). The DHP19 [1] model incorporates a CNN architecture comprising 17 convolutional layers. Besides, the two U-Net models are constructed based on the architecture proposed by [2], which integrates three downsampling and upsampling operations, as depicted in Figure 1. Additionally, we adjust the channel depths of the 3×3 convolutional layers to accommodate the varying model sizes. Specifically, in Figure 1, the channel depth N is set to 32 for U-Net-Small and 64 for U-Net-Large. The input frames are resized to 288×384. For each joint, the model outputs a heatmap that indicates the likelihood of the joint position at each pixel. To generate a target heatmap for a joint, we initialize an all-zero map of the same size as the input frame and set a value of 1 to the pixel corresponding to the annotated joint position. The heatmap is then smoothed using Gaussian blur with $\sigma = 4$. Mean Squared Error (MSE) is employed as the loss function, and RMSProp is used as the optimizer.

1.2 Performance Analysis of Joint Speed

To validate the ability of motion vectors (MV) in rectifying inaccuracies in predicted joint positions using edge images, we evaluate the Mean Per Joint Position Error (MPJPE) for

* indicates equal contribution

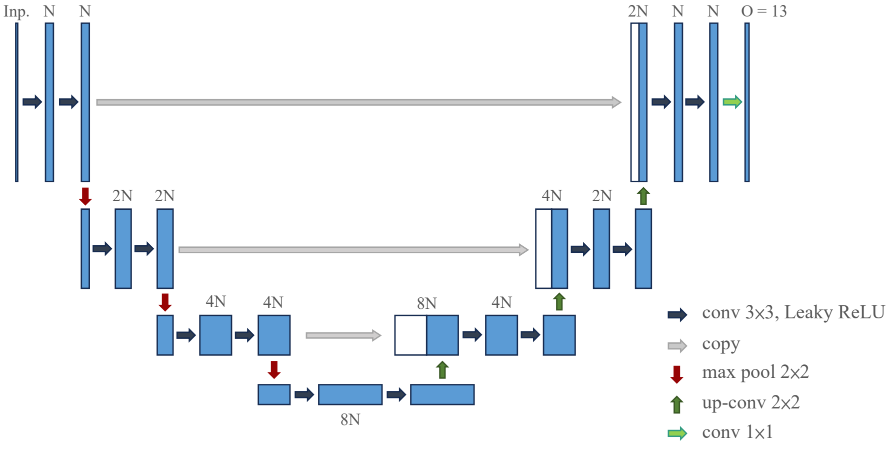


Figure 1: The architecture of U-Net-Small, and U-Net-Large. They are constructed based on U-Net [2] and incorporate three downsampling and upsampling operations. The output channel size of the convolution layers in the figure is a multiple of N, where N is set to 32 for U-Net-Small and 64 for U-Net-Large.

each of the 13 joints separately, based on different speed levels. As stated in the main paper, we categorize all individual joints in our SPHP dataset into three speed levels: slow, medium, and fast. For a 640×480 image, joints that move less than 4 pixels compared to the previous frame are classified as slow, those moving between 4-6 pixels are considered medium, and joints moving more than 6 pixels are categorized as fast.

The number of joints at different speed levels is shown in Table 1. It reveals that over three-quarters of joint speeds fall into the slow category, while approximately 13.5% and 11% are classified as medium and fast, respectively. We separately use edge and fusion (edge and motion vectors) as inputs and utilize two different backbone models as examples, namely DHP19 [1] and U-Net-Small [2]. The results, presented in Table 2, indicate that the fusion modality generally outperforms the edge modality when evaluating fast joints, as

	slow	medium	fast
Nose	151909	24880	15211
ShoulderR	153705	22587	15708
ShoulderL	154281	21815	15904
ElbowR	130450	33854	27696
ElbowL	142771	25689	23540
HipR	154352	22757	14891
HipL	154678	22501	14821
HandR	109474	39823	42703
HandL	134175	26796	31029
KneeR	143297	29046	19657
KneeL	148686	25024	18290
FootR	147437	24091	20472
FootL	155822	19214	16964
Total	1881037	338077	276886
	75.36%	13.54%	11.09%

Table 1: The number of joints at different speed levels. We categorize each joint within each frame into three levels (slow, medium, and fast) and then analyze their distribution across these speed categories. "Total" denotes the aggregated sum of all joints.

Table 2: Comparison of MPJPE for each joint at different speed levels. The experiments are conducted using respectively "edge" and "fusion" inputs, with two models: (a) DHP19 [1] and (b) U-Net-Small.

(a) DHP19 [1]

	slow		medium		fast	
keypoint	edge	fusion	edge	fusion	edge	fusion
Nose	1.60	**1.49**	2.09	**1.88**	2.00	2.02
ShoulderR	2.87	**2.83**	3.36	**3.34**	2.93	**2.70**
ShoulderL	2.73	3.16	2.85	3.02	2.38	2.88
ElbowR	6.96	7.64	8.12	8.32	9.16	**8.58**
ElbowL	5.88	7.92	8.07	9.03	9.05	**8.58**
HipR	6.54	6.76	7.08	7.63	7.28	7.80
HipL	6.66	7.79	7.08	7.58	7.52	8.15
HandR	6.31	**6.30**	6.08	**5.81**	6.41	**5.52**
HandL	5.82	5.88	7.14	**6.90**	7.65	**6.02**
KneeR	11.74	**8.90**	11.29	**10.07**	13.28	**10.24**
KneeL	8.38	**7.87**	10.42	**9.22**	14.38	**11.81**
FootR	5.20	**4.64**	6.72	**6.71**	8.75	**8.70**
FootL	3.78	**3.68**	8.13	**7.29**	15.33	**12.88**
Total	5.66	5.69	6.90	**6.76**	8.22	**7.34**

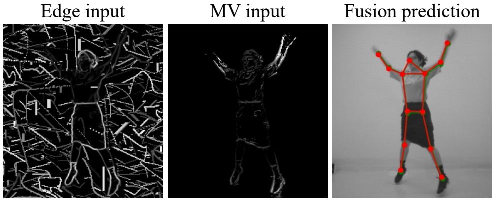
(b) U-Net-Small

	slow		medium		fast	
keypoint	edge	fusion	edge	fusion	edge	fusion
Nose	1.01	1.02	1.16	**1.11**	1.05	**1.00**
ShoulderR	2.00	**1.92**	1.99	**1.76**	2.04	**1.71**
ShoulderL	2.16	**2.11**	2.29	**2.09**	2.34	**1.86**
ElbowR	3.10	**2.93**	3.34	**3.11**	3.48	**3.19**
ElbowL	3.23	**2.86**	4.15	**3.35**	3.72	**3.20**
HipR	5.47	5.50	5.66	**5.64**	6.11	**5.72**
HipL	5.41	5.42	5.66	**5.31**	6.35	**5.93**
HandR	3.62	4.33	3.44	3.66	3.52	**3.47**
HandL	3.53	**3.19**	3.69	**3.32**	3.49	**3.23**
KneeR	5.38	**5.29**	5.78	**5.03**	6.07	**5.13**
KneeL	6.25	6.81	6.38	**6.05**	8.88	**7.70**
FootR	2.66	**2.24**	4.25	**3.10**	5.23	**3.67**
FootL	1.86	**1.79**	3.67	**3.35**	5.67	**5.53**
Total	3.51	**3.48**	3.95	**3.62**	4.33	**3.85**

highlighted in **bold**. Specifically, when using U-Net-Small in Table 2(b), the fusion modality consistently outperforms the edge modality for all keypoints at the fast level. This finding reaffirms the significant role of motion vectors in enhancing performance, especially during fast movements.

2 Robustness in cluttered backgrounds

To showcase the robustness of our approach in cluttered backgrounds, we incorporated randomly simulated background edges into the edge images of both training and testing data. Then, we retrain the model and achieve the MPJPE of 3.26 (close to 3.07 with a clear background, as stated in the main paper). Qualitative results are depicted in Figure 2. Our method is robust to background edges since MV selectively detects the areas with motion changes and filters out the static clutter.


Edge input

MV input

Fusion prediction

Figure 2: Qualitative result in simulated cluttered backgrounds.

3 Additional experiments on HumanEVA Dataset

We also evaluate our method on HumanEVA [3] dataset. The results are shown in Table 3. Our fusion methods surpasses the single modality on both traditional and sparse convolutions. This shows the generalizability of our method across various datasets.

	Gray	Edge	MV	Fusion
Conv.	4.03	4.78	9.58	4.42
Sparse Conv.	-	5.70	20.01	5.37

Table 3: MPJPE on HumanEVA [3] dataset with different input modalities. "Conv." stands for convolutions.

Model	Params#		FPS		MPJPE	
	EF	LF	EF	LF	EF	LF
DHP19	218K	655K	38.88	9.32	3.56	3.89
U-Net-Small	1.9M	5.8M	36.13	7.22	3.07	2.83
U-Net-Large	7.7M	23.1M	13.89	7.09	2.90	2.82

Table 4: Comparing early fusion (EF) with late fusion (LF).

4 Fusion Comparison

Compared to our Early Fusion (EF) method in Section 5.1, the Late Fusion (LF) model results in 3 times model size, about 2 $\sim$ 5 times lower FPS, and limited to no MPJPE improvement (see Table 4). Therefore, we select early fusion in our fusion method.

5 User Study

To evaluate the cross-modality face-matching ability of humans on our **SPHP** dataset, we conduct a survey involving 100 participants, including 59 males and 41 females. The survey consists of two parts, each containing ten questions. In the first part, participants were asked to match the edge face to a provided grayscale reference face from 10 edge images, as shown in Figure 3(a). The average accuracy for this task is 18.8%. In the second part, participants selected the corresponding face from 10 grayscale images, given a grayscale reference face, as illustrated in Figure 3(b). For this task, the participants achieve an average accuracy of 87.3%. Based on these results, we can conclude that individuals possess a limited ability to recognize daily faces when presented with leaked edge images.

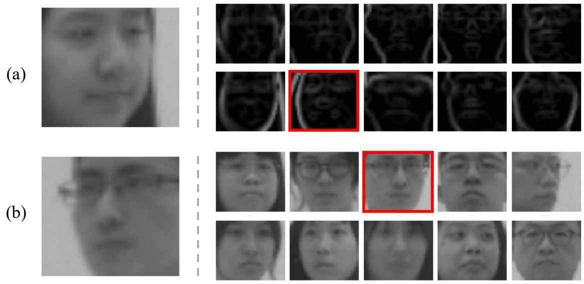


Figure 3: Examples from our human survey. In part (a), participants were asked to match the edge face to a grayscale reference face from 10 edge images. In part (b), they selected the corresponding face from 10 grayscale images, given a grayscale reference face. The red boxes represent the answer choices for the examples.

References

[1] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[3] Leonid Sigal, Alexandru Balan, and Michael Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 03 2010. doi: 10.1007/s11263-009-0273-6.