LwPosr: Lightweight Efficient Fine Grained Head Pose Estimation (Supplementary Material)

Naina Dhingra
ETH Zurich
ndhingra@ethz.ch

In this supplementary material accompanying the paper "LwPosr: Lightweight Efficient Fine Grained Head Pose Estimation", details of following topics are given:

- Results for BIWI Dataset [3] using protocol P_1
- Results for BIWI Dataset using protocol P_2
- Sine Embedding
- Transformer Encoder Layer
- Orthogonality Loss

1. Qualitative Results on Protocol P_1 with BIWI dataset

Figure 1 illustrate the qualitative head poses on the test set of BIWI dataset using protocol P_1. In this protocol P_1, *LwPosr* is both trained with artificial images from 300W-LP dataset and tested using BIWI dataset. Visualization of comparison of ground truth head poses and predicted head poses in the Figure 1 shows that *LwPosr* can efficiently recognize the head poses on real head pose images even when trained with artificially generated images. The results with FSA-Net [7] are also compared which shows that the *LwPosr* performs to the same level as FSA-Net but *LwPosr* takes less parameters than FSA-Net.

2. Qualitative Results on Protocol P_2 with BIWI dataset

Figure 2 illustrate the qualitative head poses on the test set of BIWI dataset using protocol P_2. In this protocol P_2, *LwPosr* is both trained and tested using BIWI dataset. Visualization of comparison of ground truth head poses and predicted head poses in the Figure 2 shows that *LwPosr* performs efficiently as all three line vectors almost match with the ground truth line vectors.

3. Sine Embedding

If position embedding is not present in the Transformer Encoder, then it behaves like a permutation equivariant structure as shown in Eq. 1. The ρ is the sequence order or it is any permutation for the pixel locations.

$$Encoder(\rho(input)) = \rho(Encoder(input)) \tag{1}$$

The sine positional embeddings P are used to include the order of sequence and spatial structure of the pixels in the image. It is hypothesized that the horizontal and vertical position information in an image is independent [2, 4, 6].

$$\begin{aligned} P_{(2i,p_y,:)} &= \sin\left(2\pi * p_y/(H * 10000^{2s/\frac{C}{2}})\right), \\ P_{(2i+1,p_y,:)} &= \cos\left(2\pi * p_y/(H * 10000^{2s/\frac{C}{2}})\right), \\ P_{(2i,:,p_x)} &= \sin\left(2\pi * p_x/(W * 10000^{2s/\frac{C}{2}})\right), \\ P_{(2i+1,:,p_x)} &= \cos\left(2\pi * p_x/(W * 10000^{2s/\frac{C}{2}})\right), \end{aligned} \tag{2}$$

The original 2D-structure is retained for x and y-direction with $C/2$ channels. In the Eq. 2, $s = 0, 1, ..., C/2-1$. p_x and p_y are the position indexes for the x and y direction, respectively. W, H are the width and height of the input. They are further stacked and flattened to have a shape $\mathbb{R}^{A\times C}$. The input sequence is injected with the position embedding before the self-attention is computed.

4. Transformer Encoder Layer

The Transformer Encoder layer [5, 6] used in the paper can be defined as:

$$\begin{aligned} \mathbf{X}' &= \text{LayerNorm}\left(\text{MultiheadSelfAttention}\left(\mathbf{X}\right) + \mathbf{X}\right), \\ \mathbf{X}^* &= \text{LayerNorm}\left(\text{FFN}\left(\mathbf{X}'\right) + \mathbf{X}'\right), \end{aligned} \tag{3}$$

where $\mathbf{X}$ is the input sequence without the position embedding. The querys and keys are computed when position

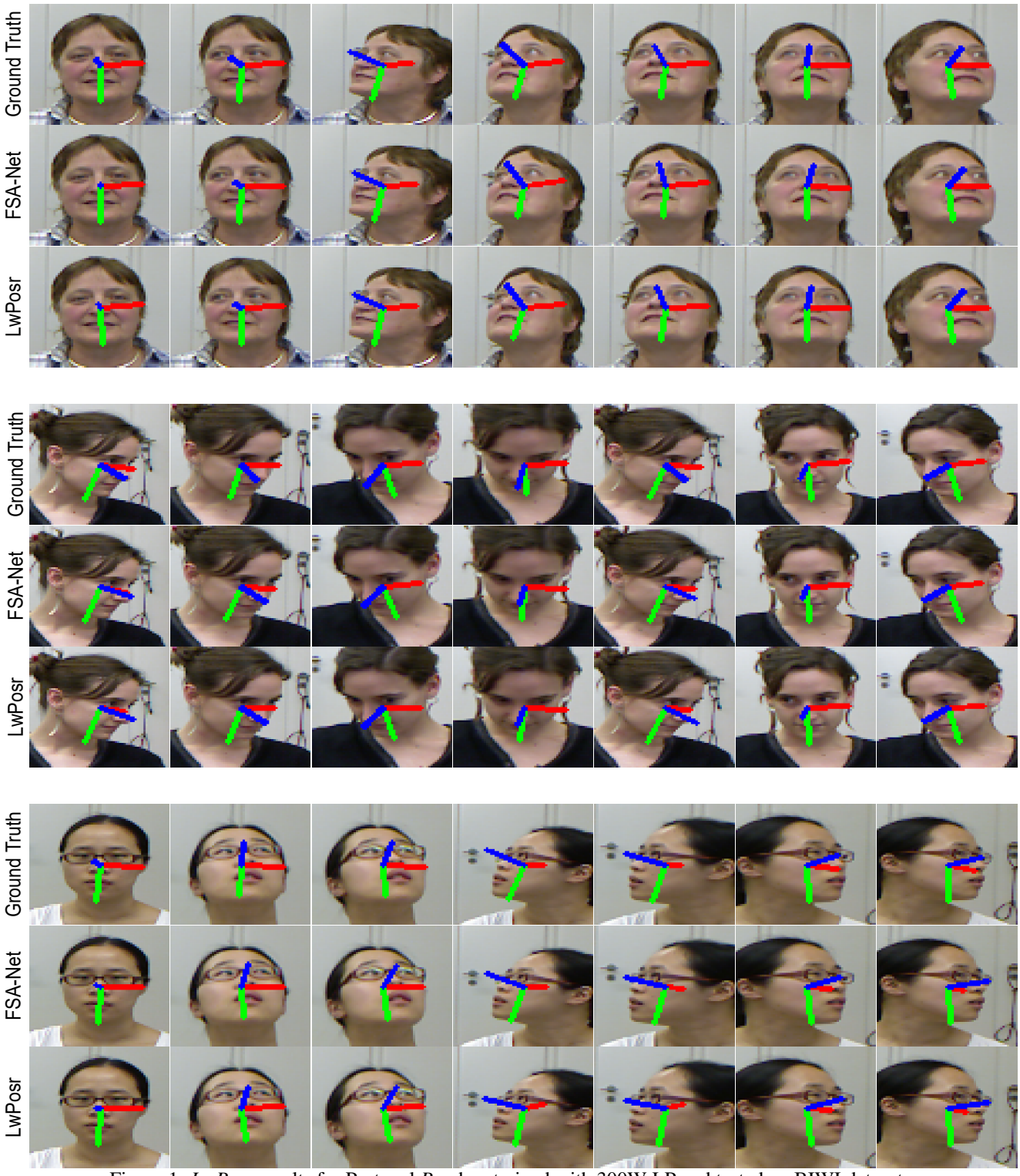

Figure 1: *LwPosr* results for Protocol P_1 when trained with 300W-LP and tested on BIWI dataset.

embedding is added to $\mathbf{X}$. The output of the Transformer Encoder layer is given by $\mathbf{X}^*$. This $\mathbf{X}^*$ is used as an input sequence for the next encoder layer. The definitions of Multihead Self-Attention and FFN are given in [5].

5. Orthogonality Loss

The orthogonality loss for head pose estimation is adapted from [1]. They represent the head poses as described below.

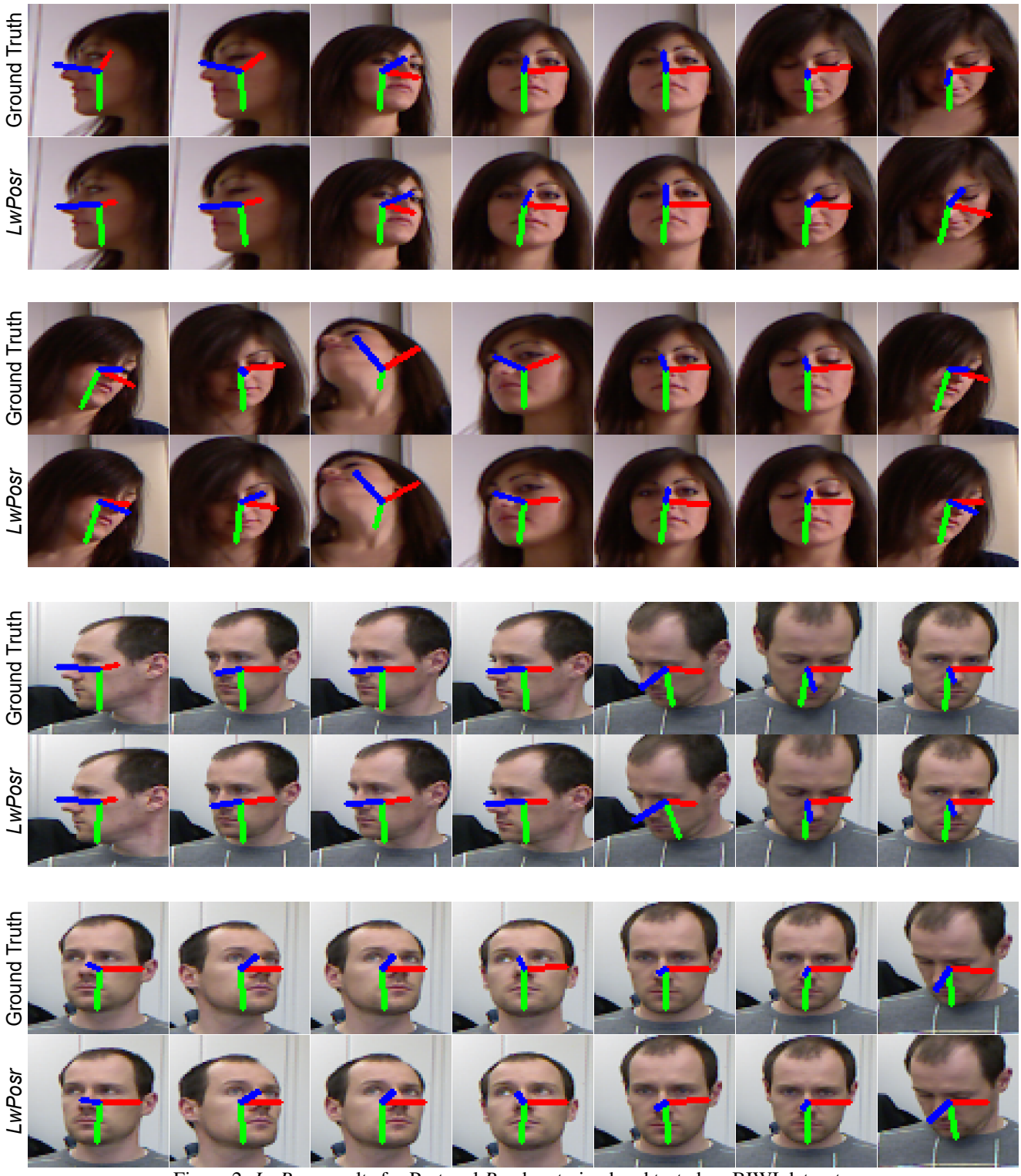

Figure 2: *LwPosr* results for Protocol P_2 when trained and tested on BIWI dataset.

5.1. Representation of Rotation

As described in [1], 3D Rotation can be represented in several ways such as Euler angle, axis-angle, quaternion, and Lie algebra. The rotation is described using 4 dimensions. In [8], it is shown that atleast 5 dimensions are needed for continuous description of 3D rotation which means that the above representations will have ambiguity problem. So, ro-

tation matrix becomes good option. It has nine elements and can have orthogonality (determinant $+1$). Continuous special orthogonal group $SO(3)$ is formed by the set of rotation matrices. It does not have problems related to ambiguity and discontinuity. The next step is to choose metric which can quantify the closeness between the two rotation matrices. The Frobenius norm of rotation matrices can be used as measure. It is a square root of sum of squares of differences of the elements of the rotation matrices. As used and shown in [1], the three vectors of head pose can be considered equivalent and corresponding to the three columns of rotation matrices. For instance, at the starting reference point, the left, down and front vectors are given as $v_l = \begin{bmatrix}1,0,0\end{bmatrix}^T$, $v_d = \begin{bmatrix}0,1,0\end{bmatrix}^T$ and $v_f = \begin{bmatrix}0,0,1\end{bmatrix}^T$ respectively. So, the rotation matrix is applied to these vectors, the resulting vectors are $v'_l = Rv_l = r_1$, $v'_d = Rv_d = r_2$ and $v'_f = Rv_f = r_3$. There are two options that can be used as a metric: (1) Frobenius norm- has a drawback that it is not intiuitve and percievable to see the difference of rotation angles by using the distance measure between the endpoints of corresponding vectors; (2) the mean absolute error of vectors (MAEV) [1], it is absolute error between the ground truth vector and predicted vector and then mean of the three error gives the final error.

Unlike [1], we used the combination of mean absolute error of angles (MAE) instead of using MAEV. But, we adapt their orthogonality loss between vectors. Loss of orthogonality between vectors is defined as:

$$\mathscr{L}_{ortho} = \sum_{i \neq j} \text{mse}(\hat{v}_i \hat{v}_j, 0) \text{ where } i, j = 1,2,3 \qquad (4)$$

where $\hat{v}_i$ is the predicted vector from the proposed network. Total loss is given as by the Eq. 5

$$\mathscr{L} = MAE + \alpha \mathscr{L}_{ortho}(\hat{v}_i, \hat{v}_j) \qquad (5)$$

As in [1], we use $\alpha = 0.5$. We also found out the $\alpha > 0.1$ did not converge as smoothly as when α is small. From ablations, it is clear that orthogonality loss did not improve the performance of the network. Hence, it increased the complexity of the optimization technique for the proposed approach. So, to keep the algorithm and optimization technique simple, we use only MAE loss for our other experiments.

References

[1] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1188–1197, 2021.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[3] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International journal of computer vision*, 101(3):437–458, 2013.

[4] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[6] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv preprint arXiv:2012.14214*, 2020.

[7] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019.

[8] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.