General response: We thank all reviewers for their comments. The reviewers agree that the idea of the proposed chiral equivariance is well motivated and interesting (R1) and the paper can be considered as the first to have a chirality transform built into the network structure (R3). The paper focuses on the practical aspects rather than mathematical theory, which makes a very useful contribution (R2) and may inspire a set of work utilizing this property in the field, which potentially has large impact (R3). A very solid/extensive experimental validation is performed (R1&R2&R3). Moreover, the paper is well written, easy to follow (R1&R2) and well-organized (R3). The authors also provide code to ensure reproducibility (R1). In the following we address all comments individually.

Reviewer # 1:

Q1: *Novelty of the technique to achieve equivariance.* Note that chirality equivariance cannot be achieved with prior methods that use parameter sharing to obtain equivariance, *e.g.*, [38]. Beyond parameter sharing, we introduce *odd and even* symmetry: note for example the negative signs in W for fully connected layers discussed in L145.

Q2: *On the loss of model representation power.* Empirically, on chiral data, we do not observe a loss of representation power for chirality nets, even though there are less trainable parameters. Throughout all our experiments, the baselines and their chiral counterparts have the same number of layers. Hence, the chiral net has less trainable parameters for each layer. On three different tasks (2D to 3D pose estimation, 2D pose forecasting and skeleton based action recognition), we observe the chiral net to outperform the baselines despite fewer trainable parameters. Obviously, if a chiral net were applied to tasks that don't exhibit chirality, we expect the representation power to be insufficient.

Q3: *Can parameter sharing be viewed as a kind of model regularization to reduce overfitting?* Yes, this is a compelling interpretation. It can also be viewed as reducing the hypothesis space of the model which decreases sample complexity according to statistical learning theory insights.

Q4: *2d human pose estimation performance of the proposed method.* Chirality equivariance is valid for skeletons like human poses, *i.e*., 2D or 3D key-points. The input for the task of 2D human pose estimation is an image (represented on a 2D-grid). Chirality is not a suitable property for this input data. However, it is certainly an interesting question how to extend the chirality equivariance definition to handle image data.

Reviewer # 2: We thank Reviewer # 2 for appreciating the paper. We'll fix typos.

Reviewer # 3:

Q5: *Jump from sec 3.2 chirality transformation directly into the fully connected layer not smooth. More clarification for better understanding the weight design.* We'll add a description at the beginning of Sec. 3.3, explaining why we need to share the parameters in the form specified in L145. Intuitively, in order to achieve chirality equivariance, we treat the left joints and the right joints equally. Therefore the parameters are shared for joints in the left and right group. Next, to handle the reflection in pose, certain dimensions are negated. We utilized odd symmetry in the parameters to "cancel" the negation. Note, we also provided the proof that verifies the chiral equivariance of the proposed layer in the supplementary material **B.1**. We are happy to extend.

Q6: *Why not verify the proposed network on the 2D pose estimation task?* Please see **Q4.**

Q7: *How does the 3D pose accuracy influence the action recognition accuracy.* We think the accuracy of 3D pose estimation is one of the important factors that influences the performance of skeleton-based action recognition. However, the focus of the paper is to discuss and demonstrate the effectiveness of the proposed chiral nets. Therefore, we used the same input 3D pose and evaluation procedure as prior works. We leave the study of robustness to noise in 3D pose estimation to future work.

Q8: *For limited data setting, the paper lacks an in-depth reasoning why when the training data ratio is extremely small, the performance is inferior to [36], while further increase training data the performance surpasses [36].* We studied this question and report our thoughts in L238-L240. We think that the hyper-parameters are not optimal for this extreme case, in particular, a batch-size of 64, used for the large datasets. As also discussed in [1], a smaller batch-size generally leads to better generalization error. We think that, the "small"-ness of a batch is relative to the size of the dataset. To verify this, we further reduced the batch-size to 16. We observe that Pavllo *et al*. [36] achieves 103.8mm and our approach improved to 98mm. We think that due to parameter sharing, the "effective batch-size" is twice that of the baseline. Note: for a fair and consistent comparison, we used the same hyper-parameters (batch-size of 64) following the setting in [36] for all models in the limited data experiments.

Q9: *Quantitative experiments on the memory saving and computation saving.* Empirically, the running time is highly dependent on the implementation and hardware platform; which deserves a careful study that is beyond the scope of this work. For this reason, we analyze the time complexity and model complexity in Sec. 3.4 and show that chirality nets are more efficient as there are less FLOPs and less model parameters due to symmetry. We analyze the model complexity of the chiral net in the paper L174-L176 and we discuss the FLOPs of a chiral net in the paper L177-L184: we show that the number of parameters is reduced by a factor of $\frac{|(|J_{\mathtt{l}}^{\mathtt{in}}|+|J_{\mathtt{c}}^{\mathtt{in}}|)\cdot(|J_{\mathtt{l}}^{\mathtt{out}}|+|J_{\mathtt{c}}^{\mathtt{out}}|)}{|J^{\mathtt{in}}|\cdot|J^{\mathtt{out}}|}$ in each chiral layer and the number of multiplications reduces by a factor of $\frac{|J_{\mathtt{l}}^{\mathtt{in}}|+|J_{\mathtt{c}}^{\mathtt{in}}|}{|J^{\mathtt{in}}|}$.

References:

[1] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proc. ICLR*, 2017.