$E(2)$-Equivariant Vision Transformer (Supplementary Material)

Renjun Xu[*1] **Kaifan Yang**[*1] **Ke Liu**[*†1] **Fengxiang He**[2, 3]

[1]College of Computer Science and Technology, Zhejiang University
[2]JD Explore Academy, JD.com, Inc.
[3]Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh

A ERRORS IN GSA-NETS

In this part, we first review the proof of equivariance of the GSA-Nets [Romero and Cordonnier, 2020], and then point out the mistakes in the proof process using the positional encoding as:

$$\rho((i,\tilde{h}),(j,\hat{h})) = \rho^P(x(j)-x(i),\tilde{h}^{-1}\hat{h})$$

A.1 DEFINITIONS AND NOTATIONS.

A.1.1 Definition of Group Equivariant Self-Attention.

If the group self-attention formulation $m^r_{\mathcal{G}}[f,\rho](i,h)$ is $\mathcal{G}$-equivariant, if and only if it satisfies:

$$m^r_{\mathcal{G}}[\mathcal{L}_g[f],\rho](i,h) = \mathcal{L}_g[m^r_{\mathcal{G}}[f,\rho]](i,h), \quad g\in\mathcal{G}$$

A.1.2 Input under g-Transformed

A g-transformed input can be expressed as:

$$\begin{aligned}\mathcal{L}_g[f](i,\tilde{h}) = \mathcal{L}_y\mathcal{L}_{\bar{h}}[f](i,\tilde{h}) = f(\rho^{-1}(\bar{h}^{-1}(\rho(i)-y)),\bar{h}^{-1}\tilde{h}),\\ g=(y,\bar{h}),\ y\in\mathbb{R}^d,\ \bar{h}\in\mathcal{H}.\end{aligned}$$

A.2 MISTAKES IN THE PROOF PROCESS OF GSA-NETS

$$\begin{aligned}&m^r_{\mathcal{G}}\left[\mathcal{L}_y\mathcal{L}_{\bar{h}}[f],\rho\right](i,h) \qquad (1)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\tilde{h}\in\mathcal{H}}\sum_{(j,\hat{h})\in\mathcal{N}(i,\tilde{h})}\sigma_{j,\hat{h}}\big(\langle\varphi^{(h)}_{\text{qry}}(\mathcal{L}_y\mathcal{L}_{\bar{h}}[f](i,\tilde{h})),\varphi^{(h)}_{\text{key}}(\mathcal{L}_y\mathcal{L}_{\bar{h}}[f](j,\hat{h})\\
&\qquad + \mathcal{L}_h[\rho]((i,\tilde{h}),(j,\hat{h}))\rangle\big)\varphi^{(h)}_{\text{val}}(\mathcal{L}_y\mathcal{L}_{\bar{h}}[f](j,\hat{h}))\Big)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\tilde{h}\in\mathcal{H}}\sum_{(j,\hat{h})\in\mathcal{N}(i,\tilde{h})}\sigma_{j,\hat{h}}\big(\langle\varphi^{(h)}_{\text{qry}}(f(x^{-1}(\bar{h}^{-1}(x(i)-y)),\bar{h}^{-1}\tilde{h})), \qquad (2)\\
&\qquad \varphi^{(h)}_{\text{key}}(f(x^{-1}(\bar{h}^{-1}(x(j)-y)),\bar{h}^{-1}\hat{h}) + \mathcal{L}_h[\rho]((i,\tilde{h}),(j,\hat{h}))\rangle\big)\\
&\qquad \varphi^{(h)}_{\text{val}}(f(x^{-1}(\bar{h}^{-1}(x(j)-y)),\bar{h}^{-1}\hat{h}))\Big)\end{aligned}$$

[*]Contributed equally.
[†]Corresponding author: Ke Liu <lk2017@zju.edu.cn>

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Using the substitutions:

$$\bar{i} = x^{-1}(\bar{\hbar}^{-1}(x(i) - y)) \Rightarrow i = x^{-1}(\bar{\hbar}x(\bar{i}) + y)), \tilde{\hbar}' = \bar{\hbar}^{-1}\tilde{\hbar}$$

and

$$\bar{j} = x^{-1}(\bar{\hbar}^{-1}(x(j) - y)) \Rightarrow j = x^{-1}(\bar{\hbar}x(\bar{j}) + y)), \hat{\hbar}' = \bar{\hbar}^{-1}\hat{\hbar}$$

the formula can be expressed as:

$$\begin{aligned} = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]} \sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}} \sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')} \sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')), \varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') \\ + \mathcal{L}_{\hbar}[\rho]((x^{-1}(\bar{\hbar}x(\bar{i})+y), \bar{\hbar}\tilde{\hbar}'), (x^{-1}(\bar{\hbar}x(\bar{j})+y), \bar{\hbar}\hat{\hbar}')))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big) \end{aligned} \tag{3}$$

By using the definition:

$$\rho((i,\tilde{\hbar}),(j,\hat{\hbar})) = \rho^P(x(j) - x(i), \tilde{\hbar}^{-1}\hat{\hbar})$$

and

$$\mathcal{L}_{\hbar}[\rho]((i,\tilde{\hbar}),(j,\hat{\hbar})) = \rho^P(\hbar^{-1}(x(j) - x(i)), \hbar^{-1}(\tilde{\hbar}^{-1}\hat{\hbar})).$$

The above formula can be further derived:

$$\begin{aligned} = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]} \sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}} \sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')} \sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')), \varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') \\ + \rho^P(\hbar^{-1}(\bar{\hbar}x(\bar{j}) + y - (\bar{\hbar}x(\bar{i}) + y)), \hbar^{-1}(\bar{\hbar}\tilde{\hbar}')^{-1}(\bar{\hbar}\hat{\hbar}'))))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big) \end{aligned} \tag{4}$$

$$\begin{aligned} = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]} \sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}} \sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')} \sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')), \varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') \\ + \rho^P(\hbar^{-1}(\bar{\hbar}x(\bar{j}) + y - (\bar{\hbar}x(\bar{i}) + y)), \hbar^{-1}\tilde{\hbar}'^{-1}\hat{\hbar}')))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big) \end{aligned} \tag{5}$$

Rather than the formula in the GSA-Nets [Romero and Cordonnier, 2020]:

$$\begin{aligned} = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]} \sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}} \sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')} \sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')), \varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') \\ + \rho^P(\hbar^{-1}(\bar{\hbar}x(\bar{j}) + y - (\bar{\hbar}x(\bar{i}) + y)), \hbar^{-1}\bar{\hbar}\tilde{\hbar}'^{-1}\hat{\hbar}')))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big) \end{aligned} \tag{6}$$

The difference between the formulas is shown in red. Therefore the subsequent derivation in the [Romero and Cordonnier, 2020] is wrong and the group self-attention module of the GSA-Nets is not group equivariant.

B PROOF OF GE-VIT

In this part, we demonstrate that using the positional encoding as follows:

$$\rho((i,\tilde{\hbar}),(j,\hat{\hbar})) = \rho^P(x(j) - x(i), \tilde{\hbar}\hat{\hbar}^{-1}\tilde{\hbar})$$

the group self-attention of GE-ViT is group equivariant.
In the process of proving, We also used the substitutions:

$$\bar{i} = x^{-1}(\bar{\hbar}^{-1}(x(i) - y)) \Rightarrow i = x^{-1}(\bar{\hbar}x(\bar{i}) + y)), \tilde{\hbar}' = \bar{\hbar}^{-1}\tilde{\hbar}$$

and

$$\bar{j} = x^{-1}(\bar{\hbar}^{-1}(x(j) - y)) \Rightarrow j = x^{-1}(\bar{\hbar}x(\bar{j}) + y)), \hat{\hbar}' = \bar{\hbar}^{-1}\hat{\hbar}$$

The complete proof process is as follows:

$$
\begin{aligned}
& m_{\mathcal{G}}^{r}\big[\mathcal{L}_y\mathcal{L}_{\bar{\hbar}}[f],\rho\big](i,\hbar) && (7)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\tilde{\hbar}\in\mathcal{H}}\sum_{(j,\hat{\hbar})\in\mathcal{N}(i,\tilde{\hbar})}\sigma_{j,\hat{\hbar}}\big(\langle\varphi_{\text{qry}}^{(h)}(\mathcal{L}_y\mathcal{L}_{\bar{\hbar}}[f](i,\tilde{\hbar})),\varphi_{\text{key}}^{(h)}(\mathcal{L}_y\mathcal{L}_{\bar{\hbar}}[f](j,\hat{\hbar})\\
&\qquad + \mathcal{L}_{\hbar}[\rho]((i,\tilde{\hbar}),(j,\hat{\hbar}))\rangle\big)\varphi_{\text{val}}^{(h)}(\mathcal{L}_y\mathcal{L}_{\bar{\hbar}}[f](j,\hat{\hbar}))\Big)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\tilde{\hbar}\in\mathcal{H}}\sum_{(j,\hat{\hbar})\in\mathcal{N}(i,\tilde{\hbar})}\sigma_{j,\hat{\hbar}}\big(\langle\varphi_{\text{qry}}^{(h)}(f(x^{-1}(\bar{\hbar}^{-1}(x(i)-y)),\bar{\hbar}^{-1}\tilde{\hbar})), && (8)\\
&\qquad \varphi_{\text{key}}^{(h)}(f(x^{-1}(\bar{\hbar}^{-1}(x(j)-y)),\bar{\hbar}^{-1}\hat{\hbar})+\mathcal{L}_{\hbar}[\rho]((i,\tilde{\hbar}),(j,\hat{\hbar}))\rangle\big)\\
&\qquad \varphi_{\text{val}}^{(h)}(f(x^{-1}(\bar{\hbar}^{-1}(x(j)-y)),\bar{\hbar}^{-1}\hat{\hbar}))\Big)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}}\sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')}\sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')),\varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') && (9)\\
&\qquad + \mathcal{L}_{\hbar}[\rho]((x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}'),(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big)
\end{aligned}
$$

By using the definition:

$$\rho((i,\tilde{\hbar}),(j,\hat{\hbar})) = \rho^{P}(x(j)-x(i),\tilde{\hbar}\hat{\hbar}^{-1}\tilde{\hbar})$$

and

$$\mathcal{L}_{\hbar}[\rho]((i,\tilde{\hbar}),(j,\hat{\hbar})) = \rho^{P}(\hbar^{-1}(x(j)-x(i)),\hbar^{-1}(\tilde{\hbar}\hat{\hbar}^{-1}\tilde{\hbar})).$$

The above formula can be further derived:

$$
\begin{aligned}
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}}\sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')}\sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')),\varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') && (10)\\
&\qquad + \rho^{P}(\hbar^{-1}(\bar{\hbar}x(\bar{j})+y-(\bar{\hbar}x(\bar{i})+y)),\hbar^{-1}(\bar{\hbar}\tilde{\hbar}')(\bar{\hbar}\hat{\hbar}')^{-1}(\bar{\hbar}\tilde{\hbar}')))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}}\sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')}\sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')),\varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') && (11)\\
&\qquad + \rho^{P}(\hbar^{-1}(\bar{\hbar}x(\bar{j})+y-(\bar{\hbar}x(\bar{i})+y)),\hbar^{-1}\bar{\hbar}\tilde{\hbar}'\hat{\hbar}'^{-1}\tilde{\hbar}'))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}}\sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')}\sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')),\varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') && (12)\\
&\qquad + \rho^{P}(\hbar^{-1}\bar{\hbar}(x(\bar{j})-x(\bar{i}),\tilde{\hbar}'\hat{\hbar}'^{-1}\tilde{\hbar}'))))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big)\\
&= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}}\sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')}\sigma_{x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}'}\big(\langle\varphi_{\text{qry}}^{(h)}(f(\bar{i},\tilde{\hbar}')),\varphi_{\text{key}}^{(h)}(f(\bar{j},\hat{\hbar}') && (13)\\
&\qquad + \mathcal{L}_{\bar{\hbar}^{-1}\hbar}[\rho]((\bar{i},\tilde{\hbar}'),(\bar{j},\hat{\hbar}'))))\rangle\big)\varphi_{\text{val}}^{(h)}(f(\bar{j},\hat{\hbar}'))\Big)
\end{aligned}
$$

The subsequent proof is similar to the GSA-Nets [Romero and Cordonnier, 2020]. For unimodular groups, the area of summation remains equal for any transformation $g \in \mathcal{G}$, which means that:

$$
\begin{aligned}
\sum_{(x^{-1}(\bar{\hbar}x(\bar{j})+y),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})+y),\bar{\hbar}\tilde{\hbar}')}[\cdot] &= \sum_{(x^{-1}(\bar{\hbar}x(\bar{j})),\bar{\hbar}\hat{\hbar}')\in\mathcal{N}(x^{-1}(\bar{\hbar}x(\bar{i})),\bar{\hbar}\tilde{\hbar}')}[\cdot]\\
&= \sum_{(x^{-1}(x(\bar{j})),\hat{\hbar}')\in\mathcal{N}(x^{-1}(x(\bar{i})),\tilde{\hbar}')}[\cdot]\\
&= \sum_{(\bar{j},\hat{\hbar}')\in\mathcal{N}(\bar{i},\tilde{\hbar}')}[\cdot].
\end{aligned}
$$

and because of the basic properties of groups, we can get $\sum_{\bar{\hbar}\tilde{\hbar}'\in\mathcal{H}}[\cdot] = \sum_{\tilde{\hbar}'\in\mathcal{H}}[\cdot]$. Consequently, the above formula can be further simplified as:

$$
\begin{aligned}
m^r_{\mathcal{G}}\big[\mathcal{L}_y\mathcal{L}_{\bar{\hbar}}[f],\rho\big](i,\hbar) =&\varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\tilde{\hbar}'\in\mathcal{H}}\sum_{(\bar{j},\hat{\hbar}')\in\mathcal{N}(\bar{i},\tilde{\hbar}')}\sigma_{\bar{j},\hat{\hbar}'}\big(\langle\varphi^{(h)}_{\text{qry}}(f(\bar{i},\tilde{\hbar}')),\\
&\varphi^{(h)}_{\text{key}}(f(\bar{j},\hat{\hbar}')+\mathcal{L}_{\bar{\hbar}^{-1}\hbar}[\rho]((\bar{i},\tilde{\hbar}'),(\bar{j},\hat{\hbar}')))\rangle\big)\varphi^{(h)}_{\text{val}}(f(\bar{j},\hat{\hbar}'))\Big)\\
=&m^r_{\mathcal{G}}[f,\rho](\bar{i},\bar{h}^{-1}\hbar)\\
=&m^r_{\mathcal{G}}[f,\rho](x^{-1}(\bar{\hbar}^{-1}(x(i)-y)),\bar{h}^{-1}\hbar)\\
=&\mathcal{L}_y\mathcal{L}_{\bar{h}}\big[m^r_{\mathcal{G}}[f,\rho]\big](i,\hbar).
\end{aligned}
\tag{14}
$$

From the above formula, it can be seen that:

$$
m^r_{\mathcal{G}}[\mathcal{L}_y\mathcal{L}_{\bar{h}}[f],\rho](i,\hbar)=\mathcal{L}_y\mathcal{L}_{\bar{h}}[m^r_{\mathcal{G}}[f,\rho]](i,\hbar),
$$

which is the same as:

$$
m^r_{\mathcal{G}}[\mathcal{L}_g[f],\rho](i,\hbar)=\mathcal{L}_g[m^r_{\mathcal{G}}[f,\rho]](i,\hbar),\quad g\in\mathcal{G}.
$$

Therefore, with the positional encoding we proposed:

$$
\rho((i,\tilde{\hbar}),(j,\hat{\hbar}))=\rho^P(x(j)-x(i),\tilde{\hbar}\hat{\hbar}^{-1}\tilde{\hbar}),
$$

the group self-attention is group equivariant.

References

David W Romero and Jean-Baptiste Cordonnier. Group equivariant stand-alone self-attention for vision. In *International Conference on Learning Representations*, 2020.