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

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

[1]College of Computer Science and Technology, Zhejiang University
[2]JD Explore Academy, JD.com, Inc.
[3]AIAI, 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:

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

B PROOF OF GE-VIT

In this section, we prove that GE-ViT is group equivariant. For brevity, we also use the substitutions:

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

and

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

[*]Contributed equally.
[†]Corresponding author: Ke Liu

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

The complete proof process is as follows:

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

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

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

By using the definition:

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

and

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

The above formula can be further derived:

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

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

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

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

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 $g\in\mathscr{G}$, which means that:

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

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

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

From the above formula, it can be seen that:

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

which is the same as:

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

Therefore, with the positional encoding we proposed:

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

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.