Group Equivariant Vision Transformer

Renjun Xu[*1,2] **Kaifan Yang**[1] **Ke Liu**[1] **Fengxiang He**[3,4]

[1] Zhejiang University
[2] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[3] JD Explore Academy, JD.com, Inc.
[4] Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh

Abstract

Vision Transformer (ViT) has achieved remarkable performance in computer vision. However, positional encoding in ViT makes it substantially difficult to realize the equivariance, compared to models based on convolutional operations which are translation-equivariant. Initial attempts have been made on designing equivariant ViT but proved not effective in some cases in this paper. To address this issue, we propose a Group Equivariant Vision Transformer (GE-ViT) via a novel, effective positional encoding operation. We prove that GE-ViT meets all the theoretical requirements of an equivariant neural network. Comprehensive experiments are conducted on standard benchmark datasets. The empirical results demonstrate that GE-ViT has made significant improvement over non-equivariant self-attention networks. The code will be released publicly.

1 INTRODUCTION

Transformer Vaswani et al. [2017] and its series of variants Devlin et al. [2018] have achieved remarkable success in natural language processing (NLP) Vaswani et al. [2017] and computer vision (CV) Carion et al. [2020], Dosovitskiy et al. [2020], Liu et al. [2021]. Different from previous methods, e.g., recurrent neural networks (RNNs) Elman [1990] and convolutional neural networks (CNNs) LeCun et al. [1989], transformer handles the input tokens simultaneously, which has shown competitive performance and superior ability in capturing long-range dependencies between these tokens. The core of transformer is the self-attention operation Vaswani et al. [2017], which excels at modeling the relationship of tokens in a sequence. Self-attention takes the similarity of token representations as attention scores and update the representations with the score weighted sum of them in an iterative manner.

Equivariance is an intrinsic property of many majors types of data in image processing Krizhevsky et al. [2012], 3D point cloud processing Li et al. [2018], chemistry Faber et al. [2016], astronomy Ntampaka et al. [2016], Ravanbakhsh et al. [2016] , etc. Zaheer et al. [2017], Zaheer et al. [2020], Cohen and Welling [2016a], Cohen and Welling [2016b], Cohen et al. [2018], Cohen et al. [2019] have adopted machine learning to realize the equivariance via modifying classic neural networks. In visual tasks, the equivariance has been highlighted in the aspects of permutation Romero and Cordonnier [2020], symmetry Krizhevsky et al. [2012], and translation Worrall et al. [2017]. CNNs guarantee the translation equivariance, i.e., the property that if a pattern in an image is translated, the numerical representation of the image is also translated instead of modified. The translation-equivariance of CNN contributes significantly to its success. The equivariance of CNNs is then extended to other symmetry groups by Cohen and Welling.

Positional encoding in vision transformer (ViT) Dosovitskiy et al. [2020] makes it substantially difficult to realize the equivariance. Initial attempts have been made to modify the self-attention to be equivariant Romero et al. [2020]. However, the efforts met significant challenges. We prove that the existing positional encoding operations are not effective in maintaining the group equivariance.

To address this issue, we propose a Group Equivariant Vision Transformer (GE-ViT) via a novel, effective equivariant positional encoding operation. We prove that the GEViT has met the theoretical requirements of a group equivariant neural network. Benefited from the group equivariance, GE-ViT significantly improves the generalization. Parameter efficiency and steerability Cohen and Welling Cohen and Welling [2016b], Weiler et al. [2018] are also guaranteed. The weights of group equivariant CNN kernels are tied to particular positions of neighborhoods on the group, which requires a mass of parameters. While GE-ViT leverages

[*]Corresponding author: rux@zju.edu.cn

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

long-range dependencies on group functions under a fixed parameter budget, which can express any group convolutional kernel Romero and Cordonnier [2020]. GE-ViT is steerable since group operations are performed directly on the positional encoding Weiler et al. [2018]. This performance of GE-ViT is evaluated by experiments which fully support our algorithm.

The main contributions of our work can be summarized as follows:

- We propose a novel Group Equivariant Vision Transformer (GE-ViT). Rigorous mathematical analysis demonstrates that the theoretical requirerments of an equivariance neural network are met in GE-ViT.
- We prove that Parameter efficiency and steerability are also guaranteed in GE-ViT.
- Experiments are conducted on standard benchmark datasets. The empirical results demonstrate consistent improvements of GE-ViT over previous works. The code will be released publicly.

The rest of this paper is orgainized as follows. Section 2 reviews the related works. Section 3 introduces the self-attention in detail and gives the notations in our paper. Preliminary concepts on groups and equivariance are introduced in Section 4. The theory of our GE-ViT, especially the positional encoding, is explained in Section 5 We report our experiments in Section 6. The discussion and future work are given in7

2 RELATED WORK

The group equivariant neural network was first proposed by Cohen and Welling [2016a], which extended the equivariance of CNNs from translation to discrete groups. The main idea of the approach is that it uses standard convolutional kernels and transforms them or the feature maps for each of the elements in the group Cohen et al. [2019]. This approach is easy to implement and has been used widely Marcos et al. [2017], Zhou et al. [2017]. However, this kind of approach can only be used in particular circumstances where locations are discrete and the group cardinality is small such as image data. Nowadays, many methods have been proposed for designing group equivariant networks. The equivariance of networks has been extended to general symmetry groups Bekkers [2019], Venkataraman et al. [2019], Weiler and Cesa [2019]. Macroscopically, equivariant neural networks can be broadly categorised by whether the input spatial data is lifted onto the space of functions on group G or not Hutchinson et al. [2021]. Without lifting, the equivariant map is defined on the homogeneous input space X. For convolutional networks, the kernel is always expressed using a basis of equivariant functions, such as circular harmonics Weiler et al. [2018], Worrall et al. [2017], spherical harmonics Thomas et al. [2018]. With lifting, the equivariant map is defined on G Cohen et al. [2018], Esteves et al. [2018], Finzi et al. [2020], Hutchinson et al. [2021], Romero and Hoogendoorn [2019]. Both GE-ViT and GSANets use lifting to define equivariant self-attention Romero and Cordonnier [2020].

ViT based on self-attention has been widely used in CV. According to the theoretical analyze $(\$4.3)$, it is the positional encoding that destroys the equivariance of self-attention. To extend the equivariance of ViT to arbitrary affine groups, a new positional encoding should be designed to replace the traditional one. Research on how to make the self-attention satisfy the general group equivariance has already existed Romero et al. [2020]. The $\mathrm{SE}(3)$-Transformers Fuchs et al. [2020] achieves this goal via the irreducible representations of $\mathrm{SO}(3)$ and LieTransformer Hutchinson et al. [2021] achieves this by means of Lie algebra. However, GE-ViT, the model proposed by this paper, achieved this by designing a new positional encoding. Besides, the above two models are specifically designed for processing 3-D point cloud data while GE-ViT is good at processing regular image data.

Based on GSA-Nets, we propose a novel positional encoding to fix the mathematic errors in GSA-Nets and get the self-attention operation theoretically equivariant. The main difference is the positional encoding which will be explained in $\$5.2$ in detail. Since the positional encoding is key to the equivariance of self-attention, our work is of great significance.

3 SELF-ATTENTION

Attention mechanism has been widely used in computer vision tasks since Mnih et al. [2014]. It was then used in the field of natural language processing(NLP) in Bahdanau et al. [2014] to improve translation accuracy. Self-attention, its variants, was proposed in Vaswani et al. [2017] and has achieved the state-of-the-art performance in various tasks of NLP. Nowadays, it has become the core module of the model in the field of NLP and the current research hotspot Jumper et al. [2021], Townshend et al. [2021], Liu et al. [2022].

In this section, We mathematically formulate the self-attention mechanism to better analyze the equivariant properties.

3.0.1 Definition of Self-Attention.

The overview of self-attention is shown in Fig.1 A self-attention module takes in N inputs and returns N outputs. Let $\mathbf{X} \in \mathbb{R}^{N \times C_{\text{in}}}$ be an input matrix consisting of N tokens of C_{in} dimensions. Let $\mathbf{Y} \in \mathbb{R}^{N \times C_{\text{out}}}$ be an output matrix consisting of N tokens of C_{out} dimensions obtained from $\mathbf{X}$ through self-attention. The whole calculation process can

be divided into the following two steps:

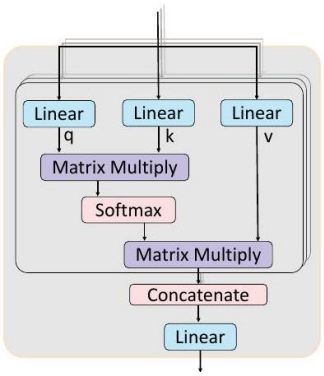

Figure 1: Illustration of self-attention. q, k, and v denote the query, key, and value respectively. Linear denotes the fully connected neural network layers. For multi-head self-attention, each black box denotes one head and gives a representation. Finally, all the representations are concatenated through the Concatenate layer and input into the Linear layer.

1. Calculate the attention scores matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$.

$$\mathbf{A} := \mathbf{X}\mathbf{W}_{\text{qry}} \left(\mathbf{X}\mathbf{W}_{\text{key}}\right)^{\top}, \tag{1}$$

where $\mathbf{W}_{\text{qry}}, \mathbf{W}_{\text{key}} \in \mathbb{R}^{C_{\text{in}}\times C_{\text{h}}}$ represent query and key matrices respectively. $\mathbf{A}_{i,j}$ represents the correlation between the i-th item and the j-th item of the input.

2. Get the output through softmax and summation.

$$\mathbf{Y} = \text{SA}(\mathbf{X}) := \text{softmax}_{[,:]}(\mathbf{A})\mathbf{X}\mathbf{W}_{\text{val}}, \tag{2}$$

where $\mathbf{W}_{\text{val}} \in \mathbb{R}^{C_{\text{in}}\times C_{\text{h}}}$ represents value matrix.

In practical application, Multi-Headed Self-Attention (MHSA) that focuses on different aspects of the input is applied. The outputs of different heads of dimension C_{h} are concatenated firstly and then projected to output via a projection matrix $\mathbf{W}_{\text{out}} \in \mathbb{R}^{HC_{\text{h}}\times C_{\text{out}}}$. The H denotes the number of heads.

$$\text{MHSA}(\mathbf{X}) := \underset{h\in[H]}{\text{concat}} \left[\text{SA}^{(h)}(\mathbf{X})\right] \mathbf{W}_{\text{out}}. \tag{3}$$

3.1 POSITIONAL ENCODING

The self-attention operation defined in Eq. 2 and Eq. 3 do not take into account structural information. Specifically, a permutation of the input $\mathbf{X}$ will only result in the same permutation transformation of the output $\mathbf{Y}$, and it does not change the concrete value of the single token of $\mathbf{Y}$. This indicates that the self-attention without positional encoding does not capture structural information. To solve this shortcoming, positional encoding $\mathbf{P}$ that contains structural information of the input is added to $\mathbf{X}$ to enrich representation. There are two positional encoding methods, absolute positional encoding, and relative positional encoding.

3.1.1 Absolute Positional Encoding.

This encoding scheme was firstly proposed in Vaswani et al. [2017]. For each position, there is a unique positional encoding. And the positional encoding can be represented by a matrix $\mathbf{P} \in \mathbb{R}^{N\times C_{\text{in}}}$. Therefore the attention scores matrix A can be expressed as follows:

$$\mathbf{A} := (\mathbf{X}+\mathbf{P})\mathbf{W}_{\text{qry}} \left((\mathbf{X}+\mathbf{P})\mathbf{W}_{\text{key}}\right)^{\top}. \tag{4}$$

$\mathbf{P}$ can be replaced by functions that return vector representation of the position index.

3.1.2 Relative Positional Encoding.

Proposed by Shaw et al. [2018], relative positional encoding considers the relative distance between the query token i and the key token j. The corresponding attention score $\mathbf{A}_{i,j}$ can be calculated by the following formula:

$$\mathbf{A}^{\text{rel}}_{i,j} := \mathbf{X}_i\mathbf{W}_{\text{qry}} \left(\left(\mathbf{X}_j + \mathbf{P}_{x(j)-x(i)}\right)\mathbf{W}_{\text{key}}\right)^{\top}. \tag{5}$$

In the above formula, $x(i)$ is the position of token i used to calculated the relative distances to other tokens. The detailed definition of $x(i)$ can be seen in $3.2 $\mathbf{P}_{x(j)-x(i)} \in \mathbb{R}^{1\times C_{\text{in}}}$ is the positional encoding of the relative distance of token i and token j.

3.2 SYMBOL DEFINITION

In this section, We recall the notations defined in GSANets Romero and Cordonnier [2020] to do a subsequent analysis of the equivariant properties of self-attention theoretically.

3.2.1 Notation

The set $\{1, 2, 3, \ldots, n\}$ are denoted by $[n]$ and let $\mathcal{S} = [N]$. $L_{\mathcal{V}}(\mathcal{S})$ denotes the space of functions $\{f : \delta \to \mathcal{V}\}$, where $\mathcal{V}$ represents a vector space. With the above definition, a matrix $\mathbf{X} \in \mathbb{R}^{N\times C_{\text{in}}}$ can be interpreted as a vector valued function $f_X : \delta \to \mathbb{R}^{C_{\text{in}}}$ that maps element $i \in \delta$ to C_{in} -dimension vector $\mathbf{X}_i \in \mathbb{R}^{C_{\text{in}}}$. A matrix multiplication, $\mathbf{X}\mathbf{W}_y^{\top}$, of matrices $\mathbf{X} \in \mathbb{R}^{N\times C_{\text{in}}}$ and $\mathbf{W}_y \in \mathbb{R}^{C_{\text{out}}\times C_{\text{in}}}$ can be represented as a function $\varphi_y : L_{\mathbb{R}^{C_{\text{in}}}}(\mathcal{S}) \to L_{\mathbb{R}^{C_{\text{out}}}}(\mathcal{S})$,

$$\varphi_y(f_X) = f_{XW_y^T}$$

With the above definitions, the attention score matrix (Eq.1) without positional encoding can be represented as:

$$\mathbf{A}_{i,j} = \alpha[f](i,j) = \langle \varphi_{\text{qry}}(f(i)), \varphi_{\text{key}}(f(j)) \rangle \tag{6}$$

The function $\alpha[f] : \mathcal{S}\times\mathcal{S} \to \mathbb{R}$ maps pairs of set elements $i, j \in \mathcal{S}$ to the attention score $\mathbf{A}_{i,j}$. As a result, the self-

attention (Eq.2) can be represented as:

$$\begin{aligned}\mathbf{Y}_{i,:} = \zeta[f](i) &= \sum_{j\in\delta} \sigma_j(\alpha[f](i,j))\varphi_{\text{val}}\left(f(j)\right) \\ &= \sum_{j\in\delta} \sigma_j\left(\langle\varphi_{\text{qry}}(f(i)), \varphi_{\text{key}}(f(j))\rangle\right)\varphi_{\text{val}}(f(j)).\end{aligned} \tag{7}$$

In the above formula, $\zeta[f] : \mathcal{S} \to \mathbb{R}^{C_{\text{h}}}, \sigma = \text{softmax}$, and $\sigma_j = \frac{e^{z_j}}{\sum_{i=1}^{N} e^{z_i}}$. Similarly, the MHSA (Eq. 3) can be expressed as:

$$\begin{aligned}&\text{MHSA}\left(\mathbf{X}_i\right) = m[f](i) = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]} \zeta^{(h)}[f](i)\Big) \\ &= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{j\in\mathcal{S}} \sigma_j\left(\left\langle\varphi_{\text{qry}}^{(h)}(f(i)), \varphi_{\text{key}}^{(h)}(f(j))\right\rangle\right)\varphi_{\text{val}}^{(h)}(f(j))\Big)\end{aligned} \tag{8}$$

In the above formula, $\cup$ is the concatenation operator and $m[f] : \delta \to \mathbb{R}^{C_{\text{out}}}$. Unlike the field of natural language processing, computer vision tasks often deal with images composed of many pixels. Because of the quadratic time complexity of the self-attention, only part of the image which always are nearest to the i_{th} item is selected when calculating the output of the i_{th} item. Let $\eta_{(i)}$ be the selected part related to the i_{th} item. $\eta_{(i)}$ is also called the local neighborhood of the token i in the later section. Therefore, replacing $\mathcal{S}$ with $\eta_{(i)}$ in Eq. 8 can be written as:

$$\begin{aligned}&\text{MHSA}\left(\mathbf{X}_i\right) = m[f](i) = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]} \zeta^{(h)}[f](i)\Big) \\ &= \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{j\in\eta(i)} \sigma_j\left(\left\langle\varphi_{\text{qry}}^{(h)}(f(i)), \varphi_{\text{key}}^{(h)}(f(j))\right\rangle\right)\varphi_{\text{val}}^{(h)}(f(j))\Big)\end{aligned} \tag{9}$$

3.2.2 Absolute Positional Encoding.

The positional encoding is a function $\rho : \delta \to \mathbb{R}^{C_{\text{in}}}$ that maps set elements $i \in \mathcal{S}$ to a vector representation. Using this definition, Eq. 4 can be written as:

$$\begin{aligned}m[f,p](i) = \varphi_{\text{out}}\Bigg(\cup_{h\in[H]}\sum_{j\in\eta(i)} \sigma_j\Big(\Big\langle\varphi_{\text{qry}}^{(h)}(f(i)+\rho(i)), \\ \varphi_{\text{key}}^{(h)}(f(j)+\rho(j))\Big\rangle\Big)\varphi_{\text{val}}^{(h)}(f(j))\Bigg)\end{aligned} \tag{10}$$

As defined in Romero and Cordonnier [2020], the function ρ can be decomposed as two functions $\rho^P \circ x : (i)$ the position function $x : \delta \to X$, which provides the position of set elements in the underlying homogeneous space, and, (ii) the positional encoding $\rho^P : X \to \mathbb{R}^{C_{\text{in}}}$, which provides vector representations of elements in X.

3.2.3 Relative Positional Encoding.

Similar to absolute positional encoding, relative positional encoding can be defined as $\rho(i,j) := \rho^P(x(j) - x(i))$ among pairs $(i,j), i \in \delta, j \in \eta(i)$. Therefore, the Eq. 5 can be written as:

$$\begin{aligned}m[f,p](i) = \varphi_{\text{out}}\Bigg(\bigcup_{h\in[H]}\sum_{j\in\eta(i)} \sigma_j\Big(\Big\langle\varphi_{\text{qry}}^{(h)}(f(i)), \\ \varphi_{\text{key}}^{(h)}(f(j)+\rho(i,j))\Big\rangle\Big)\varphi_{\text{val}}^{(h)}(f(j))\Bigg)\end{aligned} \tag{11}$$

4 GROUP EQUIVARIANCE

This section lays down some of the necessary definitions and notations in group theory and representation theory. Then recall the equivariance properties of the self-attention proposed by Romero and Cordonnier [2020].

4.1 BASIC CONCEPTS AND NOTATIONS

4.1.1 Definition of Group

A group is an abstract mathematical concept. Formally a group $(G;\circ)$ consists of a set G and a binary composition operator $\circ : G \times G \to G$. All groups must adhere to the following 4 axioms:

1. Closure: $g \circ h \in G$ for all $g, h, \in G$
2. Associativity: $f \circ (g \circ h) = (f \circ g) \circ h = f \circ g \circ h$ for all $f, g, h \in G$
3. Identity: There exists an element such that $e \circ g = g \circ e = g$ for all $g \in G$
4. Inverses: For each $g \in G$ there exists a $g^{-1} \in G$ such that $g^{-1} \circ g = g \circ g^{-1} = e$.

Each group element $g \in G$ corresponds to a symmetry transformation. In practice, the binary composition operator $\circ$ can be omitted, so would write gh instead of $g \circ h$. Groups can be finite or infinite, countable or uncountable, compact or non-compact. Note that they are not necessarily commutative; that is, $gh \neq hg$ in general. If a group is commutative, that is $gh = hg$ for all $g, h \in G$, it is called the Abelian Group. One example of the infinite group is $SE(2)$, the set of all 2D rotations about the origin and the 2D translation. Because the image is being transformed in 2D, $SE(2)$ is the focus of this paper.

4.1.2 Group Action

A group action $\rho(g)$ is an bijective map from a space into itself: $\rho(g) : \mathcal{X} \to \mathcal{X}$. It is parameterized by an element g of a group G. For the expression $\rho(g)x$, we say that $\rho(g)$ acts on x. A symmetry transformation of group element $g \in G$ on object $x \in X$ is referred to as the group action of G on X.$\rho(g)x$ is often written as gx to reduce clutter. In the context of group equivariant neural networks, grouping

action object X is commonly defined to be the space of scalar-valued functions or vector-valued functions on some set $\mathcal{E}$, so that $X = \{f \mid f : \mathcal{E} \to \mathbb{R}^d\}$. This set $\mathcal{E}$ could be a Euclidean input space, e.g. a grey-scale image can be expressed as a feature map $f : \mathbb{R}^2 \to \mathbb{R}$ from pixel coordinate x_i to pixel intensity f_i, supported on the grid of pixel coordinates.

4.1.3 Group Representation.

A group representation $\rho : G \to GL(N)$ is a map from a group G to the set of $N \times N$ invertible matrices $GL(N)$. Critically ρ is a group homomorphism, that is, it satisfies the following property:

$$\rho(g1 \circ g2) = \rho(g1)\rho(g2), \quad \forall g1, g2 \in G. \tag{12}$$

For SO(2), the standard rotation matrix is an example of a representation that acts on $\mathbb{R}^2$:

$$\rho(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{13}$$

According to the above definition, the rotation of the image can be expressed as a representation of $SO(2)$ by extending the action ρ on the pixel coordinates x to a representation π that acts on the space of feature maps $\{f \mid f : \mathcal{E} \to \mathbb{R}^d\}$:

$$[\pi(g)(f)](x) \triangleq f\left(\rho\left(g^{-1}\right)x\right), \tag{14}$$

where $\mathcal{E} = \{x_i\}$, or write gx instead of $\rho(g)x$ to reduce clutter:

$$[\pi(g)(f)](x) \triangleq f\left(g^{-1}x\right). \tag{15}$$

And it is equivalent to the mapping:

$$(x_i, f_i)_{i=1}^n \to (\rho(g)x_i, f_i)_{i=1}^n, \tag{16}$$

where n is the total number of pixels in the image.

4.1.4 Affine Group

Affine groups have the following form: $\mathcal{G} = \mathbb{R}^d \rtimes \mathcal{H}$. It is resulting from the semi-direct product $(\rtimes)$ between the translation group $\left(\mathbb{R}^d, +\right)$ and an group $\mathscr{H}$ that acts on $\mathbb{R}^d$. $\mathcal{H}$ can be rotation, mirroring and so on.

4.1.5 Group Equivariance

A map $\Phi : V_1 \to V_2$ is G-equivariant with respect to actions $\rho 1, \rho 2$ of G acting on V_1, V_2 respectively if:

$$\Phi[\rho 1(g)f] = \rho 2(g)[\Phi[f]], \quad \forall g \in G, f \in V_1. \tag{17}$$

As is well-known, convolution is an equivariant map for the translation group.

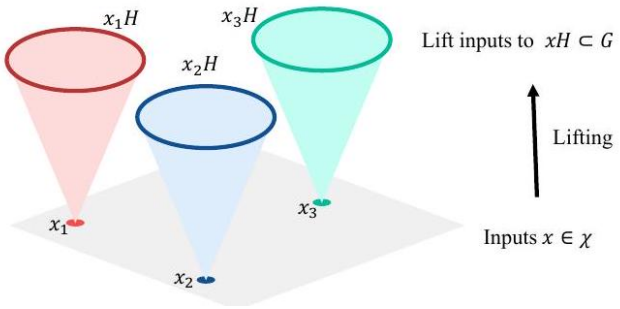

Figure 2: The illustration of lifting. For any $x \in X$, $f(x)$ equals to $\mathcal{L}(f)(g)$ on G, where $g \in xH$, $\mathcal{L}$ is the lifting operation, and f is a function defined on X.

4.2 LIFTING

We can view X as a quotient group G/H for some subgroup H of a group G, that means X is isomorphic to G/H. Then naturally the function f defined on X can be viewed as defined on G/H. Thus we define the lifting operation $\mathcal{L}$ on the function f as

$$\mathcal{L}(f)(g) = f([g]), \tag{18}$$

where $[g] \in G/H$ is the equivalent class of g.

For example, $\mathbb{R}^2$ is isomorphic to $SE(2)/SO(2)$, and every element $g \in SE(2)$ can be written as tr uniquely, where $t \in \mathbb{R}^2$ and $r \in SO(2)$. Furthermore, for any function f on $\mathbb{R}^2$, the lifting function $\mathcal{L}(f)$ is defined as $\mathcal{L}(f)(g) = f(t)$.

4.3 EQUIVARIANCE OF SELF-ATTENTION

There are several important conclusions about the equivariance of self-attention which has been proved correctly by Romero and Cordonnier [2020]:

1. The global self-attention formulation without positional encoding (Eq.3) is permutation equivariant.
2. Absolute position-aware self-attention (Eq. 4) is neither permutation nor translation equivariant.
3. Relative position-aware self-attention (Eq. 5) is translation equivariant.

5 GROUP EQUIVARIANT SELF-ATTENTION

In §4.3, it has shown that translation equivariance can be achieved via relative positional encoding. For 2D images, there are usually translation and rotation transformations. Therefore, for the model, not only the translation equivariance but also the rotation equivariance need to be satisfied.

To achieve the above goals, an improved version of positional encoding based on relative encoding needs to be designed. When designing equivariant networks, there are usually two choices of group representation: irreducible representation and regular representation. The experimental results Fuchs et al. [2020], Hutchinson et al. [2021], Weiler and Cesa [2019] shows that regular representation is more expressive and Ravanbakhsh [2020] has proved it theoretically. A lifting self-attention layer is an essential module to obtain feature representation based on regular representation. The main function of the lifting layer is mapping fx (a function defined on $\mathbb{R}^d$) to $\mathcal{L}\left[f_x\right]$ (a function defined on G). After the lifing layer, the feature has been defined on the group G, which brings practical implementation problems when the group G is infinite. Because the summation over group elements $g \in G$ (Eq. 22) is an essential step. Fortunately, extensive experiments Weiler and Cesa [2019] have shown that via proper discrete approximations, networks using regular representations can achieve satisfactory results.

In this section, we recall the lifting and the group self-attention of the GSA-Nets Romero and Cordonnier [2020] and point out that the positional encoding of the GSA-Nets makes the whole network does not satisfy the rotation equivariance. At the end, a modified version of the positional encoding has been proposed to impose the rotation equivariance to the network.

$$\rho((i,\tilde{h}),(j,\hat{h})) = \rho^P\left(x(j)-x(i),\tilde{h}^{-1}\hat{h}\right) \tag{19}$$

5.1 LIFTING SELF-ATTENTION

As previously mentioned, the lifting self-attention is a map from functions on $\mathbb{R}^d$ to functions on $\mathcal{G}$ and can be expressed as: $m^r_{\mathcal{G}\uparrow}[f,\rho] : L_{\mathcal{V}}\left(\mathbb{R}^d\right) \rightarrow L_{V'}(\mathcal{L})$, where $\mathcal{G}$ is an affine group and $\mathcal{G} = \mathbb{R}^d \rtimes \mathcal{H}$ as notated in 4.1.4. The action of group element $h \in \mathscr{H}$ on relative positional encoding $\rho(i,j)$ is defined as: $\{\mathcal{L}_h[\rho](i,j)\}_{h\in\mathscr{H}}$, $\mathcal{L}_h[\rho](i,j) = \rho^P\left(h^{-1}x(j) - h^{-1}x(i)\right)$. Consequently, the formula of lifting self-attention can be expressed as:

$$\begin{aligned} m^r_{\mathcal{G}\uparrow}[f,\rho](i,h) &= m^r\left[f,\mathcal{L}_h[\rho]\right](i) \\ = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{j\in\eta(i)} &\sigma_j(\langle\varphi^{(h)}_{\text{qry}}(f(i)),\varphi^{(h)}_{\text{key}}(f(j)+ \\ &\mathcal{L}_h[\rho](i,j))\rangle)\varphi^{(h)}_{\text{val}}(f(j))\Big). \end{aligned} \tag{20}$$

It has been proven that the lifting self-attention defined above is equivariant to the affine group $\mathcal{G}$ Romero and Cordonnier [2020].

5.2 GROUP SELF-ATTENTION

After the lifting self-attention layer, the feature map can be viewed as a function defined on $\mathcal{G}_{\mathcal{L}}$. So the action of group elements $h \in \mathcal{H}$ on relative positional encoding $\rho(i,j)$ is defined as: $\left\{\mathcal{L}_h[\rho]((i,\tilde{h}),(j,\hat{h}))\right\}_{h\in\mathscr{H}}$. The positional encoding used in Romero and Cordonnier [2020] is:

$$\rho((i,\tilde{h}),(j,\hat{h})) = \rho^P\left(x(j)-x(i),\tilde{h}^{-1}\hat{h}\right). \tag{21}$$

Therefore, the group action on relative positional encoding can be expressed as:

$$\left\{\mathcal{L}_h[\rho]((i,\tilde{h}),(j,\hat{h}))\right\}_{h\in\mathcal{H}} = \rho^P\left(h^{-1}(x(j)-x(i)),h^{-1}\left(\tilde{h}^{-1}\hat{h}\right)\right).$$

Similar to the lifting self-attention layer, the formula of group self-attention can be expressed as:

$$\begin{aligned} m^r_{\mathcal{G}}[f,\rho](i,h) &= \sum_{\tilde{h}\in\mathcal{H}} m^r\left[f,\mathcal{L}_h[\rho]\right](i,\tilde{h}) \\ = \varphi_{\text{out}}\Big(\bigcup_{h\in[H]}\sum_{\tilde{h}\in\mathcal{H}}\sum_{(j,\hat{h})\in\eta(i,\tilde{h})} &\sigma_{j,\hat{h}}(\langle\varphi^{(h)}_{\text{qry}}(f(i,\tilde{h})), \\ \varphi^{(h)}_{\text{key}}\left(f(j,\hat{h})+\mathcal{L}_h[\rho]((i,\tilde{h}),(j,\hat{h}))\right)\rangle)&\varphi^{(h)}_{\text{val}}(f(j,\hat{h}))\Big). \end{aligned} \tag{22}$$

However, we prove the group self-attention using the positional encoding defined at Eq. 21 is not $\mathcal{G}$-equivariant.

$$m^r_{\mathcal{G}}\left[\mathcal{L}_g[f],\rho\right](i,h) \neq \mathcal{L}_g\left[m^r_{\mathcal{G}}[f,\rho]\right](i,h),\ g\in\mathcal{G}_{\mathcal{L}} \tag{23}$$

Appendix A shows the detailed proofs. In order to make the module satisfy the equivariant property, we propose a novel positional encoding to replace the old one (Eq. 21):

$$\rho((i,\tilde{h}),(j,\hat{h})) = \rho^P\left(x(j)-x(i),\tilde{h}\hat{h}^{-1}\tilde{h}\right). \tag{24}$$

Correspondingly, the group action on relative positional encoding can be expressed as:

$$\mathcal{L}_h[\rho]((i,\tilde{h}),(j,\hat{h})) = \rho^P\left(h^{-1}(x(j)-x(i)),h^{-1}\left(\tilde{h}\hat{h}^{-1}\tilde{h}\right)\right). \tag{25}$$

It can be proven (Appendix B) that using the modified version of positional encoding (Eq.24), the group self-attention is $\mathcal{G}$-equivariant. That is,

$$m^r_{\mathcal{G}}\left[\mathcal{L}_g[f],\rho\right](i,h) = \mathcal{L}_g\left[m^r_{\mathcal{G}}[f,\rho]\right](i,h),\quad g\in\mathcal{G}_{\mathcal{L}}. \tag{26}$$

6 EXPERIMENTS

We conduct a study on standard benchmark datasets, including rotMNIST, to evaluate the performance of GE-ViT compared with GSA-Nets using neighborhood size. In order to demonstrate the superiority of our positional encoding fairly, the structure except for positional encoding module and the number of parameters of the models used to compare keep the same. That is, the GE-ViT and GSA-Nets used to compare have the same number of parameters and structure except for the positional encoding module. Experimental results illustrate that GE-ViT consistently outperforms not only equivalent non-equivariant attention networks but also the GSA-Nets. The code will be released publicly.

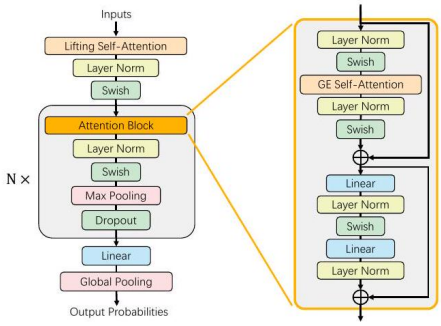

Figure 3: Illustration of GE-ViT and Attention Block. The blocks on the left show the structure of GE-ViT. Functions are transformed from R^2 to Group through Lifting Self-Attention. N denotes the number of blocks in the black box. The Global Pooling block consists of max-pool over group elements followed by spatial mean-pool. Swish is an activation function Ramachandran et al. [2017]. The flow on the right illustrates the structure of the Attention Block. Linear denotes the fully connected neural network layers.

6.1 EXPERIMENT SETUP

This section explains the experiment setup.

6.1.1 Dataset.

RotMNIST dataset is constructed by rotating the MNIST dataset. It is a classification dataset often used as a standard benchmark for rotation equivariance Weiler and Cesa [2019]. RotMNIST contains $62k$ gray-scale 28×28 uniformly rotated handwritten digits. The rotMNIST has been divided into training, validation, and test sets of $10\mathrm{k}, 2\mathrm{k}$, and $50\mathrm{k}$ images.

6.1.2 Compared Approaches.

Following Romero and Cordonnier [2020], we compare our GE-ViT with Z2_SA and GSA-Nets. Z2_SA is a translation equivariant self-attention model. GSA-Nets is also a self-attention-based model, which tried to introduce more kinds of equivariance to Z2_SA.

6.2 IMPLEMENTATION DETAILS

This section gives the experimental implementation details.

6.2.1 Invariant Network.

The invariant network is a special case of the equivariant network. It makes sense that the invariant network is more suitable for classification tasks than the equivariant network.

Table 1: Classification accuracy (%) of R4_SA with different neighborhood size on rotMNIST.

MODEL	GSA-Nets	GE-ViT (ours)
3×3	96.28	**96.63**
5×5	97.47	**97.58**
7×7	97.33	**97.45**
9×9	97.10	**97.15**
11×11	97.06	**97.16**
15×15	96.89	**97.12**
19×19	96.86	**97.37**
23×23	96.90	**97.01**

The function composition of several equivariant functions followed by an invariant function f, is an invariant function Hutchinson et al. [2021]. Therefore, the Gobal Pooling layer, an invariant map, is added to the end of the GE-ViT and GSA-Nets in our experiments.

6.2.2 Model Structure

Fig. 3 shows the structure of our GE-ViT, which is similar with GSA-Nets Romero and Cordonnier [2020]. The core modules of the GE-ViT and GSA-Nets are the lifting self-attention and group self-attention. Linear map, layer normalization, and activation function are interspersed in the model. Following Dosovitskiy et al. [2020], Liu et al. [2021], Romero and Cordonnier [2020], the Global Pooling block, in the end, consists of max-pool over group elements followed by spatial mean-pool. In our experiments, we choose the local self-attention because of the computational constraints. The neighborhood size $\mathrm{n} \times \mathrm{n}$ denotes the chosen size of the local region. Following Romero and Cordonnier [2020], rotation equivariant models are notated as $\mathbf{Rn}$, where $\mathbf{n}$ represents the angle discretization. Specifically speaking, R4_SA depicts a model equivariant to rotations by 90 degrees and R8_SA depicts a model equivariant to rotations by 45 degrees.

6.2.3 Hyperparameters Setting

To ensure fairness, the hyperparameters remain fixed for all experiments. The number of epochs is 300 and the batch size is 8. The learning rate is set to 0.001 and the weight decay is set to 0.0001. Attention dropout rate and value dropout rate are both set to 0.1. Adam optimizer is applied.

6.3 EXPERIMENTS AND RESULTS

Experiments are conducted to compare our GE-ViT with previous methods. Table 1 shows the classification results of **R4_SA** with different neighborhood size. Table 2 shows the classification results of different equivariant models

Table 2: Classification accuracy (%) of different equivariant models on rotMNIST. All architectures based on self-attention use 5×5 neighborhood size.

MODEL	GSA-Nets	GE-ViT (ours)
Z2_SA	96.63	
R4_SA	97.46	**97.58**
R8_SA	97.79	**97.88**
R12_SA	97.97	**98.01**
R16_SA	97.66	**97.83**

with 5×5 neighborhood size. The reported performance of GSANets is recorded from the official released code (GSA-Nets). It can be seen from the experimental results that our GE-ViT outperforms other methods consistently under any setup. With more kinds of equivariance, GSA-Nets beats **Z2_SA** on most settings since some errors exist in GSA-Nets. Our novel positional encoding improves the classification accuracy in GE-ViT and makes a new SOTA. Besides, R4_SA with the neighborhood size of 5×5 achieves the best accuracy. This finding is also available in Romero and Cordonnier [2020]. Since in the whole experiment, only the positional encodings are different and the rest remains the same, the experimental results can demonstrate the superiority of the positional encoding we proposed.

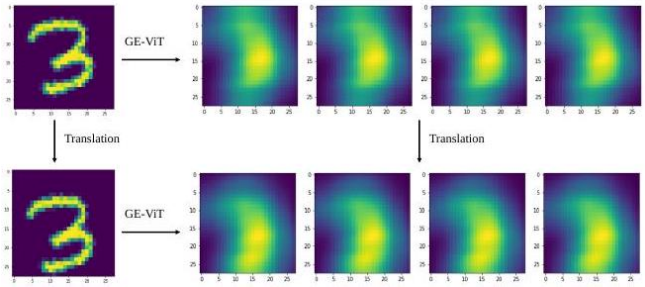

Figure 4: Translation equivariance of GE-ViT. The images on the left are the raw data and the right images are feature representations. Specifically speaking, feature representations of the original data are shown in the top right of the image, and feature representations obtained by translating the original data are in the lower right of the image.

Fig. 4 demonstrates the translational equivariance of our GEViT: a translation of an input image induces a translation of the intermediate feature representations of the GE-ViT. Fig. 5 demonstrates the rotational equivariance of our GEViT: a rotation of an input image induces a rotation plus a circular permutation of the intermediate feature representations of the GE-ViT. Fig. 6 demonstrates the reflectional equivariance of our GE-ViT: a reflection of an input image induces a reflection plus a circular permutation of the intermediate feature representations of the GE-ViT.

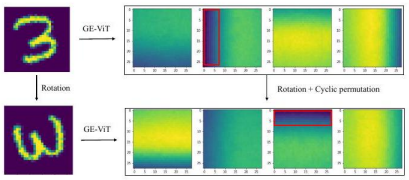

Figure 5: Rotation equivariance of GE-ViT. The images on the left are the raw data and the images on the right are feature representations. Specifically speaking, feature representations of the original data are shown in the top right of the image, and feature representations obtained by rotating the original data are in the lower right of the image.

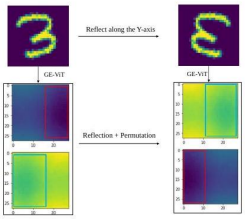

Figure 6: Reflection equivariance of GE-ViT. The images on the top are the raw data and the images on the bottom are feature representations. Specifically speaking, feature representations of the original data are shown in the lower left of the image, and feature representations obtained by flipping the original data are in the lower right of the image.

7 DISCUSSION AND FUTURE WORK

GE-ViT with a novel and effective positional encoding outperforms GSA-Nets and non-equivariant self-attention networks are competitive to G-CNNs. However, G-CNNS still performs better on most data sets Romero et al. [2020], which may be due to the optimization problem of GE-ViT or the limits on computing resources Liu et al. [2020]. From the theoretical perspective, the group equivariant self-attention can be more expressive than G-CNNS Cordonnier et al. [2019], so the GE-ViT has a lot of potential for improvement in the aspect of initialization, optimization, generalization and so on Zhao et al. [2020].

Quadratic space and time complexity in the number of inputs is the inevitable problem and the major flaws of the model based on self-attention. There have been many researches that study efficient variants of self-attention to alleviate this issue Katharopoulos et al. [2020], Kitaev et al. [2019], Wang et al. [2020], Zaheer et al. [2020]. And the above variants can be directly integrated into GE-ViT. Besides, we hope that our proposed positional encoding will provide a novel perspective on designing more robust models.

References

D Bahdanau, K Cho, and Y Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

EJ Bekkers. B-spline cnns on lie groups. In *International Conference on Learning Representations*, 2019.

N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, and S Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

T Cohen and M Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2016a.

TS Cohen and M Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

TS Cohen, M Geiger, J Köhler, and M Welling. Spherical cnns. In *International Conference on Learning Representations*, 2018.

TS Cohen, M Geiger, and M Weiler. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32:9142–9153, 2019.

J-B Cordonnier, A Loukas, and M Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and et al. Gelly, Sylvain. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.

Felix A Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million elpasolite (abC_2D_6) crystals. *Physical review letters*, 117(13):135502, 2016.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3) transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 1970–1981, 2020.

Michael J Hutchinson, Charles Le Lan, Syed Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pages 4533–4543. PMLR, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Ž'ıdek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Angelos Katharopoulos, Apoorva Vyas, Nikolaos Pappas, and Francois Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, 2012.

Yann LeCun, Bernhard Boser, John S Denker, Don Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551, 1989.

Yangyan Li, Rongchang Bu, Mingchao Sun, Wei Wu, Xiaoxiang Di, and Baoquan Chen. Pointenne: Convolution on x-transformed points. In *Advances in neural information processing systems*, volume 31, 2018.

Kezhi Liu, Kun Yang, Jun Zhang, and Rui Xu. S2snet: A pretrained neural network for superconductivity discovery. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22 AI for Good*, pages 5101–5107, 2022.

Lilian Weng Liu, Xiaodan Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Brian Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *Advances in neural information processing systems*, 2: 2204–2212, 2014.

Michelle Ntampaka, Hy Trac, David J Sutherland, Sebastien Fromenteau, Barnabás Póczos, and Jeff Schneider. Dynamical mass measurements of contaminated galaxy clusters using machine learning. *The Astrophysical Journal*, 831(2):135, 2016.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Siamak Ravanbakhsh. Universal equivariant multilayer perceptrons. In *International Conference on Machine Learning*, pages 7996–8006, 2020.

Siamak Ravanbakhsh, Joydeep Oliva, Sebastien Fromenteau, Luke C Price, Shirley Ho, Jeff Schneider, and Barnabás Póczos. Estimating cosmological parameters from the dark matter distribution. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 2407–2416, 2016.

Daniel Romero, Erik Bekkers, Jakub Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 8188–8199, 2020.

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

Daniel W Romero and Mark Hoogendoorn. Co-attentive equivariant neural networks: Focusing equivariance on transformations co-occurring in data. In *International Conference on Learning Representations*, 2019.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. Tensor field networks: Rotation and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

R.J. Townshend, S. Eismann, A.M. Watkins, R. Rangan, M. Karelina, R. Das, and R.O. Dror. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

S.R. Venkataraman, S. Balasubramanian, and R.R. Sarma. Building deep equivariant capsule networks. In *International Conference on Learning Representations*, 2019.

S. Wang, B.Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

M. Weiler and G. Cesa. General e(2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pages 14334–14345, 2019.

M. Weiler, F.A. Hamprecht, and M. Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018.

D.E. Worrall, S.J. Garbin, D. Turmukhambetov, and G.J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R.R. Salakhutdinov, and A.J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Manzil Zaheer, Guru Guruganesh, Kunal A Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Hang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020.

Yang Zhou, Qixiang Ye, Qiangui Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 519528, 2017.