GROUP EQUIVARIANT
STAND-ALONE SELF-ATTENTION FOR VISION

David W. Romero
Vrije Universiteit Amsterdam
d.w.romeroguzman@vu.nl

Jean-Baptiste Cordonnier
École Polytechnique Fédérale de Lausanne (EPFL)
jean-baptiste.cordonnier@epfl.ch

ABSTRACT

We provide a general self-attention formulation to impose group equivariance to arbitrary symmetry groups. This is achieved by defining positional encodings that are invariant to the action of the group considered. Since the group acts on the positional encoding directly, group equivariant self-attention networks (GSA-Nets) are steerable by nature. Our experiments on vision benchmarks demonstrate consistent improvements of GSA-Nets over non-equivariant self-attention networks.

1 INTRODUCTION

Recent advances in Natural Language Processing have been largely attributed to the rise of the Transformer (Vaswani et al., 2017). Its key difference with previous methods, e.g., recurrent neural networks, convolutional neural networks (CNNs), is its ability to query information from all the input words simultaneously. This is achieved via the self-attention operation (Bahdanau et al., 2015; Cheng et al., 2016), which computes the similarity between representations of words in the sequence in the form of attention scores. Next, the representation of each word is updated based on the words with the highest attention scores. Inspired by the capacity of transformers to learn meaningful inter-word dependencies, researchers have started applying self-attention in vision tasks. It was first adopted into CNNs by channel-wise attention (Hu et al., 2018) and non-local spatial modeling (Wang et al., 2018). More recently, it has been proposed to replace CNNs with self-attention networks either partially (Bello et al., 2019) or entirely (Ramachandran et al., 2019). Contrary to discrete convolutional kernels, weights in self-attention are not tied to particular positions (Fig. A.1), yet self-attention layers are able to express any convolutional layer (Cordonnier et al., 2020). This flexibility allows leveraging long-range dependencies under a fixed parameter budget.

An arguable orthogonal advancement to deep learning architectures is the incorporation of symmetries into the model itself. The seminal work by Cohen & Welling (2016) provides a recipe to extend the translation equivariance of CNNs to other symmetry groups to improve generalization and sample-efficiency further (see §2). Translation equivariance is key to the success of CNNs. It describes the property that if a pattern is translated, its numerical descriptors are also translated, but not modified.

In this work, we introduce group self-attention, a self-attention formulation that grants equivariance to arbitrary symmetry groups. This is achieved by defining positional encodings invariant to the action of the group considered. In addition to generalization and sample-efficiency improvements provided by group equivariance, group equivariant self-attention networks (GSA–Nets) bring important benefits over group convolutional architectures: (i) Parameter efficiency: contrary to conventional discrete group convolutional kernels, where weights are tied to particular positions of neighborhoods on the group, group equivariant self-attention leverages long-range dependencies on group functions under a fixed parameter budget, yet it is able to express any group convolutional kernel. This allows for very expressive networks with low parameter count. (ii) Steerability: since the group acts directly on the positional encoding, GSA–Nets are steerable (Weiler et al., 2018b) by nature. This allows us to go beyond group discretizations that live in the grid without introducing interpolation artifacts.

Contributions:

• We provide an extensive analysis on the equivariance properties of self-attention (§4).
• We provide a general formulation to impose group equivariance to self-attention (§5).
• We provide instances of self-attentive architectures equivariant to several symmetry groups (§6).
• Our results demonstrate consistent improvements of GSA–Nets over non-equivariant ones (§6).
Several approaches exist which provide equivariance to various symmetry groups. The translation equivariance of CNNs has been extended to additional symmetries ranging from planar rotations (Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Weiler et al., 2018b; Li et al., 2018; Cheng et al., 2018; Hoogeboom et al., 2018; Bekkers et al., 2018; Veeling et al., 2018; Lenssen et al., 2018; Graham et al., 2020) to spherical rotations (Cohen et al., 2018; 2019b; Worrall & Brostow, 2018; Weiler et al., 2018a; Esteves et al., 2019a,b; 2020), scaling (Marcos et al., 2018; Worrall & Wellings, 2019; Josovik et al., 2020; Romero et al., 2020b) and more general symmetry groups (Cohen & Wellings, 2016; Konior & Trivedi, 2018; Tai et al., 2019; Weiler & Cesa, 2019; Cohen et al., 2019a; Bekkers, 2020; Venkataraman et al., 2020). Importantly, all these approaches utilize discrete convolutional kernels, and thus, tie weights to particular positions in the neighborhood on which the kernels are defined. As group neighborhoods are (much) larger than conventional ones, the number of weights discrete group convolutional kernels require proportionally increases. This phenomenon is further exacerbated by *attentive group equivariant networks* (Romero & Hoogendoorn, 2019; Diaconu & Worrall, 2019; Romero et al., 2020a). Since attention is used to leverage non-local information to aid local operations, non-local neighborhoods are required. However, as attention branches often rely on discrete convolutions, they effectively tie specific weights to particular positions on a large non-local neighborhood on the group. As a result, attention is bound to growth of the model size, and thus, to negative statistical efficiency. Differently, group self-attention is able to attend over arbitrarily large group neighborhoods under a fixed parameter budget. In addition, group self-attention is steerable by nature (§5.1) a property primarily exhibited by works carefully designed to that end.

Other way to detach weights from particular positions comes by parameterizing convolutional kernels as (constrained) neural networks (Thomas et al., 2018; Finzi et al., 2020). Introduced to handle irregularly-sampled data, e.g., point-clouds, networks parameterizing convolutional kernels receive relative positions as input and output their values at those positions. In contrast, our mappings change as a function of the input content. Most relevant to our work are the SE(3) and Lie Transformers (Fuchs et al., 2020; Hutchinson et al., 2020). However, we obtain group equivariance via a generalization of positional encodings, Hutchinson et al. (2020) does so via operations on the Lie algebra, and Fuchs et al. (2020) via irreducible representations. In addition, our work prioritizes applications on visual data and extensively analyses theoretical aspects and properties of group equivariant self-attention.

## 3 Stand-Alone Self-Attention

In this section, we recall the mathematical formulation of self-attention and emphasize the role of the positional encoding. Next, we introduce a functional formulation to self-attention which will allow us to analyze and generalize its equivariance properties.

**Definition.** Let $X \in \mathbb{R}^{N \times C_{in}}$ be an input matrix consisting of $N$ tokens of $C_{in}$ dimensions each. A self-attention layer maps an input matrix $X \in \mathbb{R}^{N \times C_{in}}$ to an output matrix $Y \in \mathbb{R}^{N \times C_{out}}$ as:

$$Y = SA(X) := \text{softmax}_{(i\cdot)}(A)XW_{val},$$

with $W_{val} \in \mathbb{R}^{C_{in} \times C_{out}}$ the value matrix, $A \in \mathbb{R}^{N \times N}$ the attention scores matrix, and $\text{softmax}_{(i\cdot)}(A)$ the attention probabilities. The matrix $A$ is computed as:

$$A := XW_{qry}(XW_{key})^T,$$

parameterized by query and key matrices $W_{qry}, W_{key} \in \mathbb{R}^{C_{in} \times C_{in}}$. In practice, it has been found beneficial to apply multiple self-attention operations, also called heads, in parallel, such that different

---

1 We consequently consider an image as a set of $N$ discrete objects $i \in \{1, 2, \ldots, N\}$. 

---

Figure 1: Behavior of feature representations in group self-attention networks. An input rotation induces a rotation plus a cyclic permutation to the intermediary feature representations of the network. Additional examples for all the groups used in this work as well as their usage are provided in repo/demo/.
heads are able to attend to different parts of the input. In this multi-head self-attention formulation, the output of $H$ heads of output dimension $C_h$ are concatenated and projected to $C_{\text{out}}$ as:

$$\text{MHSA}(X) := \text{concat}_{h \in [H]} \left[ \text{SA}^{(h)}(X) \right] W_{\text{out}} + b_{\text{out}},$$

with a projection matrix $W_{\text{out}} \in \mathbb{R}^{HC_h \times C_{\text{out}}}$ and a bias term $b_{\text{out}} \in \mathbb{R}^{C_{\text{out}}}$.

### 3.1 The Role of the Positional Encoding

Note that the self-attention operation defined in Eq. 3 is equivariant to permutations of the input rows of $X$. That is, a permutation of the rows of $X$ will produce the same output $Y$ up to this permutation. Hence, self-attention is blind to the order of its inputs, i.e., it is a set operation. Illustratively, an input image is processed as a bag of pixels and the structural content is not considered. To alleviate this limitation, the input representations in self-attention are often enriched with a positional encoding that provides positional information of the set elements.

**Absolute positional encoding.** Vaswani et al. (2017) introduced a (learnable) positional encoding $P \in \mathbb{R}^{N \times C_P}$ for each input position which is added to the inputs when computing the attention scores:

$$A := \left( X + P \right) W_{\text{qy}} \left( \left( X + P \right) W_{\text{key}} \right)^T.$$

More generally, $P$ can be substituted by any function that returns a vector representation of the position and can be incorporated by means of addition or concatenation, e.g., Zhao et al. (2020). This positional encoding injects additional structural information about the tokens into the model, which makes it susceptible to changes in the token’s positions. Unfortunately, the model must learn to recognize similar patterns at every position independently as absolute positional encodings are unique to each position. This undesired data inefficiency is addressed by relative positional encodings.

**Relative positional encoding.** Introduced by Shaw et al. (2018), relative encodings consider the relative distance between the query token $i$ – the token we compute the representation of –, and the key token $j$ – the token we attend to –. The calculation of the attention scores (Eq. 2) then becomes:

$$A_{i,j}^{\text{rel}} := X_i W_{\text{qy}} \left( \left( X_j + P_{x(j-x(i))} \right) W_{\text{key}} \right)^T,$$

where $P_{x(j-x(i))} \in \mathbb{R}^{1 \times C_P}$ is a vector representation of the relative shift and $x(i)$ is the position of the token $i$ as defined in §3.2. Consequently, similar patterns can be recognized at arbitrary positions, as relative query-key distances always remain equal.

### 3.2 A Functional Formulation to Self-Attention

**Notation.** We denote by $[n]$ the set $\{1, 2, \ldots, n\}$. Given a set $S$ and a vector space $V$, $L_V(S)$ will denote the space of functions $\{ f : S \to V \}$. Square brackets are used when functions are arguments.

Let $S = \{ i \}_{i=1}^N$ be a set of $N$ elements. A matrix $X \in \mathbb{R}^{N \times C_P}$ can be interpreted as a vector-valued function $f : S \to \mathbb{R}^{C_n}$ that maps element sets $i \in S$ to $C_n$-dimensional vectors: $f : i \mapsto f(i)$. Consequently, a matrix multiplication, $X W_y$, of matrices $X \in \mathbb{R}^{N \times C_P}$ and $W_y \in \mathbb{R}^{C_P \times C_n}$ can be represented as a function $\varphi_y : L_{\mathbb{C}_n}(S) \to L_{\mathbb{C}_n}(S)$, $\varphi_y : f(i) \mapsto \varphi_y(f(i))$, parameterized by $W_y$, between functional spaces $L_{\mathbb{C}_n}(S) = \{ f : S \to \mathbb{C}_n \}$ and $L_{\mathbb{C}_n}(S) = \{ f : S \to \mathbb{C}_n \}$. Following this notation, we can represent the position-less attention scores calculation (Eq. 2) as:

$$A_{i,j} = \alpha[f](i, j) = \langle \varphi_{\text{qy}}(f(i)), \varphi_{\text{key}}(f(j)) \rangle.$$

The function $\alpha[f] : S \times S \to \mathbb{R}$ maps pairs of set elements $i, j \in S$ to the attention score of $j$ relative to $i$. Therefore, the self-attention (Eq. 1) can be written as:

$$Y_{i,:} = \zeta[f](i) = \sum_{j \in S} \sigma_j(\alpha[f](i, j)) \varphi_{\text{val}}(f(j)) = \sum_{j \in S} \sigma_j(\langle \varphi_{\text{qy}}(f(i)), \varphi_{\text{key}}(f(j)) \rangle) \varphi_{\text{val}}(f(j)),$$

where $\sigma_j = \text{softmax}_j$ and $\zeta[f] : S \to \mathbb{R}^{C_h}$. Finally, multi-head self-attention (Eq. 3) can be written as:

$$\text{MHSA}(X)_{i,:} = m[f](i) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \zeta^{(h)}(f)(i) \right) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{j \in S} \sigma_j(\langle \varphi_{\text{qy}}^{(h)}(f(i)), \varphi_{\text{key}}^{(h)}(f(j)) \rangle) \varphi_{\text{val}}^{(h)}(f(j)) \right),$$

where $\cup$ is the functional equivalent of the concatenation operator $\text{concat}$, and $m[f] : S \to \mathbb{R}^{C_{\text{out}}}$.

**Local self-attention.** Recall that $\alpha[f]$ assigns an attention scores to every other set element $j \in S$ relative to the query element $i$. The computational cost of self-attention is often reduced by restricting
its calculation to a local neighborhood \( \mathcal{N}(i) \) around the query token \( i \) analogous in nature to the local receptive field of CNNs (Fig. A.1a). Consequently, local self-attention can be written as:

\[
m[f](i) = \varphi_{\text{out}} \left( \bigcup_{h \in \mathcal{H}} \sum_{j \in \mathcal{N}(i)} \sigma_r \left( \langle \varphi_{\text{q}_\text{ky}}(f(i)), \varphi_{\text{k}_\text{ey}}(f(j)) \rangle \right) \varphi_{\text{val}}(f(j)) \right).
\]

(9)

Note that Eq. 9 is equivalent to Eq. 8 for \( \mathcal{N}(i) = S \), i.e. when considering global neighborhoods.

**Absolute positional encoding.** The absolute positional encoding is a function \( \rho : S \rightarrow \mathbb{R}^{C_n} \) that maps set elements \( i \in S \) to a vector representation of its position: \( \rho : i \rightarrow \rho(i) \). Note that this encoding is not dependent on functions defined on the set but only on the set itself. Consequently, absolute position-aware self-attention (Eq. 4) can be written as:

\[
m[f, \rho](i) = \varphi_{\text{out}} \left( \bigcup_{h \in \mathcal{H}} \sum_{j \in \mathcal{N}(i)} \sigma_r \left( \langle \varphi_{\text{q}_\text{ky}}(f(i) + \rho(i)), \varphi_{\text{k}_\text{ey}}(f(j) + \rho(j)) \rangle \right) \varphi_{\text{val}}(f(j)) \right).
\]

(10)

The function \( \rho \) can be decomposed as two functions \( \rho^o \circ x \): (i) the position function \( x : S \rightarrow X \), which provides the position of set elements in the underlying space \( X \) (e.g., pixel positions), and, (ii) the positional encoding \( \rho^P : X \rightarrow \mathbb{R}^{C_n} \), which provides vector representations of elements in \( X \). This distinction will be of utmost importance when we pinpoint where exactly (group) equivariance must be imposed to the self-attention operation (§4.3, §5).

**Relative positional encoding.** Here, positional information is provided in a relative manner. That is, we now provide vector representations of relative positions \( \rho(i,j) \equiv \rho^P(x(j) - x(i)) \) among pairs \( (i,j), i \in S, j \in \mathcal{N}(i) \). Consequently, relative position-aware self-attention (Eq. 5) can be written as:

\[
m^r[f, \rho](i) = \varphi_{\text{out}} \left( \bigcup_{h \in \mathcal{H}} \sum_{j \in \mathcal{N}(i)} \sigma_r \left( \langle \varphi_{\text{q}_\text{ky}}(f(i) + \rho(i,j)), \varphi_{\text{k}_\text{ey}}(f(j) + \rho(i,j)) \rangle \right) \varphi_{\text{val}}(f(j)) \right).
\]

(11)

### 4 Equivariance Analysis of Self-Attention

In this section we analyze the equivariance properties of self-attention. Since the analysis largely relies on group theory, we provide all concepts required for proper understanding in Appx. C.

#### 4.1 Group Equivariance and Equivariance for Functions Defined on Sets

First we provide the general definition of group equivariance and refine it to relevant groups next. Additionally, we define the property of unique equivariance to restrict equivariance to a given group.

**Definition 4.1 (Group equivariance).** Let \( G \) be a group (Def. C.1), \( S, S' \) be sets, \( V, V' \) be vector spaces, and \( L^G_\sigma[\cdot], L^G_\sigma'[\cdot] \) be the induced (left regular) representation (Def. C.4) of \( G \) on \( L_\sigma(V) \) and \( L_\sigma(V') \), respectively. We say that a map \( \varphi : L_\sigma(V) \rightarrow L_\sigma(V') \) is equivariant to the action of \( G \) or \( G\)-equivariant if it commutes with the action of \( G \). That is, if:

\[
\varphi[L^G_\sigma(f)] = L^G_\sigma' \left[ \varphi[f] \right], \quad \forall f \in L_\sigma(V), \forall g \in G.
\]

**Example 4.1.1 (Permutation equivariance).** Let \( S = S' = \{ i \}_{i=1}^N \) be a set of \( N \) elements, and \( G = \mathbb{S}_N \) be the group of permutations on sets of \( N \) elements. A map \( \varphi : L_\sigma(V) \rightarrow L_\sigma(V') \) is said to be equivariant to the action of \( \mathbb{S}_N \) or permutation equivariant if:

\[
\varphi[L^\mathbb{S}_N(f)](i) = L^\mathbb{S}_N' \left[ \varphi[f] \right](i), \quad \forall f \in L_\sigma(V), \forall \pi \in \mathbb{S}_N, \forall i \in S,
\]

where \( L^\mathbb{S}_N(f)(i) \equiv f(\pi^{-1}(i)) \), and \( \pi : S \rightarrow S \) is a bijection from the set to itself. The element \( \pi(i) \) indicates the index to which the \( i \)-th element of the set is moved to as an effect of the permutation \( \pi \). In other words, \( \varphi \) is said to be permutation equivariant if it commutes with permutations \( \pi \in \mathbb{S}_N \). That is, if permutations in its argument produce equivalent permutations on its response.

Several of the transformations of interest, e.g., rotations, translations, are not defined on sets. Luckily, as we consider sets gathered from homogeneous spaces \( X \) where these transformations are well-defined, e.g., \( \mathbb{R}^2 \) for pixels, there exists an injective map \( x : S \rightarrow X \) that associates a position in \( X \) to each set element, the *position function*. In Appx. D we show that the action of \( G \) on such a set is well-defined and induces a group representation to functions on it. With this in place, we are now able to define equivariance of set functions to groups whose actions are defined on homogeneous spaces.

**Definition 4.2 (Equivariance of set functions to groups acting on homogeneous spaces).** Let \( G \) be a group acting on two homogeneous spaces \( \mathcal{X} \) and \( \mathcal{X}' \), let \( S, S' \) be sets and \( V, V' \) be vector
spaces. Let \( x : S \to \mathcal{X} \) and \( x' : S' \to \mathcal{X}' \) be injective maps. We say that a map \( \varphi : L_\mathcal{V}(S) \to L_\mathcal{V}(S') \) is equivariant to the action of \( G \) or \( G'-\)equivariant – if it commutes with the action of \( G' \). That is, if:

\[
\varphi\left[ L_\mathcal{V}[f]\right] = L_\mathcal{V'}\left[\varphi[f]\right], \quad \forall f \in L_\mathcal{V}(S), \quad \forall g \in G,
\]

where \( L_\mathcal{V'}[f](i) := f(x^{-1}(g^{-1}x(i))) \) are the induced (left regular) representation of \( G' \) on \( L_\mathcal{V}(S) \) and \( L_\mathcal{V'}(S') \), respectively. I.o.w., \( \varphi \) is said to be \( G'-\)equivariant if a transformation \( g \in G \) on its argument produces a corresponding transformation on its response.

**Example 4.2.1 (Translation equivariance).** Let \( S, S' \) be sets and let \( x : S \to \mathcal{X} \) and \( x' : S' \to \mathcal{X}' \) be injective maps from the sets \( S, S' \) to the corresponding homogeneous spaces \( \mathcal{X}, \mathcal{X}' \) on which they are defined, e.g., \( \mathbb{R}^d \) and \( \vec{G} \). With \( (\mathcal{X}, +) \) the translation group acting on \( \mathcal{X} \), we say that a map \( \varphi : L_\mathcal{V}(S) \to L_\mathcal{V'}(S) \) is equivariant to the action of \( (\mathcal{X}, +) \) – or translation equivariant – if:

\[
\varphi\left[ L_\mathcal{V}[f]\right](i) = L_\mathcal{V'}\left[\varphi[f]\right](i), \quad \forall f \in L_\mathcal{V}(S), \quad \forall y \in \mathcal{X},
\]

with \( L_\mathcal{V'}[f](i) := f(x^{-1}(x(i)−y)) \). I.o.w., \( \varphi \) is said to be translation equivariant if a translation on its argument produces a corresponding translation on its response.

### 4.2 Equivariance properties of self-attention

In this section we analyze the equivariance properties of the self-attention. The proofs to all the propositions stated in the main text are provided in Appx. G.

**Proposition 4.1.** The global self-attention formulation without positional encoding (Eqs. 3, 8) is permutation equivariant. That is, it holds that:

\[
m[L_\mathcal{V}[f]](i) = L_\mathcal{V}m[f](i), \quad \forall f \in L_\mathcal{V}(S), \quad \forall i \in S.
\]

Note that permutation equivariance only holds for global self-attention. The local variant proposed in Eq. 9 reduces permutation equivariance to a smaller set of permutations where neighborhoods are conserved under permutation, i.e., \( \mathcal{S}_N = \{ \pi \in \mathcal{S}_N \mid j \in \mathcal{N}(i) \rightarrow \pi(j) \in \mathcal{N}(i) \}, \forall i \in S \).

**Permutation equivariance induces equivariance to important (sub)groups.** Consider the cyclic group of order 4, \( \mathcal{Z}_4 = \{ e, r, r^2, r^3 \} \) which induces planar rotations by \( 90^\circ \). As every rotation in \( \mathcal{Z}_4 \) effectively induces a permutation of the tokens positions, it can be shown that \( \mathcal{Z}_4 \) is a subgroup of \( \mathcal{S}_N \), i.e., \( \mathcal{S}_N \geq \mathcal{Z}_4 \). Consequently, maps equivariant to permutations are automatically equivariant to \( \mathcal{Z}_4 \). However, as the permutation equivariance constraint is harder than that of \( \mathcal{Z}_4 \)-equivariance, imposing \( \mathcal{Z}_4 \)-equivariance as a result of permutation equivariance is undesirable in terms of expressivity. Consequently, Ravanbakhsh et al. (2017) introduced the concept of unique \( G \)-equivariance to express the family of functions equivariant to \( G \) but not equivariant to other groups \( G' \geq G \):

**Definition 4.3 (Unique \( G \)-equivariance).** Let \( G \) a subgroup of \( G' \), \( G \leq G' \) (Def. C.2). We say that a map \( \varphi \) is uniquely \( G \)-equivariant iff it is \( G'- \)equivariant but not \( G'' \)-equivariant for any \( G'' \geq G \).

In the following sections, we show that we can enforce unique equivariance not only to subgroups of \( \mathcal{S}_N \), e.g., \( \mathcal{Z}_4 \), but also to other interesting groups not contained in \( \mathcal{S}_N \), e.g., groups of rotations finer than 90 degrees. This is achieved by enriching set functions with a proper positional encoding.

**Proposition 4.2.** Absolute position-aware self-attention (Eqs. 4, 10) is neither permutation nor translation equivariant. I.e., \( m[L_\mathcal{V}[f], \rho](i) = L_\mathcal{V}m[f, \rho](i) \) and \( m[L_\mathcal{V'}[f], \rho](i) = L_\mathcal{V'}m[f, \rho](i) \).

Though absolute positional encodings do disrupt permutations equivariance, they are unable to provide translation equivariance. We show next that translation equivariance is obtained via relative encodings.

**Proposition 4.3.** Relative position-aware self-attention (Eq. 11) is translation equivariant. That is, it holds that:

\[
m'\left[ L_\mathcal{V}[f], \rho\right](i) = L_\mathcal{V}'\left[m'[f, \rho]\right](i).
\]

### 4.3 Where exactly is equivariance imposed in self-attention?

In the previous section we have seen two examples of successfully imposing group equivariance to self-attention. Specifically, we see that no positional encoding allows for permutation equivariance and that a relative positional encoding allows for translation equivariance. For the latter, as shown in the proof of Prop. 4.3 (Appx. G), this comes from the fact that for all shifts \( y \in \mathcal{X} \):

\[
\rho(x^{-1}(x(i) + y), x^{-1}(x(j) + y)) + \rho^P(x(j) + y − x(i) + y)) = \rho^P(x(j) − x(i)) = \rho(i, j).
\]

That is, from the fact that the relative positional encoding is invariant to the action of the translation group, i.e., \( L_\mathcal{V}[\rho](i, j) = \rho(i, j), \forall y \in \mathcal{X} \). Similarly, the absence of positional encoding – more precisely, the use of a constant positional encoding –, is what allows for permutation equivariance

\[ e \] represents a \( 0^\circ \) rotation, i.e., the identity. The remaining elements \( r^j \) represent rotations by \( (90\cdot j)^\circ \).
We conjecture this is a result of the networks receiving discrete signals as input. As the action of which the action of the group is well-defined, e.g., which becomes an integral for continuous groups. Interestingly, using discrete groups does not seem (Prop. 4.1, Appx. G). Specifically, constant positional encodings $\rho_c(i) = c, \forall i \in S$ are invariant to the action of the permutation group, i.e., $\mathcal{L}_\pi[\rho_c](i) = \rho_c(i), \forall \pi \in S_N$.

From these observations, we conclude that $G$-equivariance is obtained by providing positional encodings which are **invariant to the action of the group** $G$, i.e., $\mathcal{L}_\pi[\rho] = \rho, \forall \pi \in G$. Furthermore, unique $G$-equivariance is obtained by providing positional encodings which are invariant to the action of $G$ but **not invariant to the action of any other group** $G' \geq G$. This is a key insight that allows us to provide (unique) equivariance to arbitrary symmetry groups, which we provide next.

5 Group Equivariant Stand-Alone Self-Attention

In §4.3 we concluded that unique $G$-equivariance is induced in self-attention by introducing positional encodings which are invariant to the action of $G$ but not invariant to the action of other groups $G' \geq G$. However, this constraint does not provide any information about the expressivity of the mapping we have just made $G$-equivariant. Let us first illustrate why this is important:

Consider the case of imposing rotation and translation equivariance to an encoding defined in $\mathbb{R}^2$. Since translation equivariance is desired, a relative positional encoding is required. For rotation equivariance, we must further impose the positional encoding to be equal for all rotations. That is $\mathcal{L}_\theta[\rho](i, j) = \rho(i, j), \forall \theta \in [0, 2\pi]$, where $\mathcal{L}_\theta[\rho](i, j) := \rho^P(\theta^{-1}x(j) - \theta^{-1}x(i))$, and $\theta^{-1}$ depicts a rotation by $-\theta$ degrees. This constraint leads to an isotropic positional encoding unable to discriminate among orientations, which in turn enforces rotation invariance instead of rotation equivariance.

This is alleviated by **lifting** the underlying function on $\mathbb{R}^2$ to a space where rotations are explicitly encoded (Fig. B.1). To this end, one performs self-attention operations for positional encodings $\mathcal{L}_\theta[\rho]$ of varying values $\theta$ and indexes their responses by the corresponding $\theta$ value. Next, as rotations are now explicitly encoded, a positional encoding can be defined in this space which is able to discriminate among rotations (Fig. B.2). This in turn allows for rotation equivariance instead of rotation invariance.

It has been shown both theoretically (Ravanbakhsh, 2020) and empirically (Weiler & Cesa, 2019) that the most expressive class of $G$-equivariant functions is given by functions that follow the regular representation of $G$. In order to obtain feature representations that behave that way, we introduce a lifting self-attention layer (Fig. B.1, Eq. 14) that receives an input function on $\mathbb{R}^d$ and produces a feature representation on $G$. Subsequently, arbitrarily many group self-attention layers (Fig. B.2, Eq. 16) interleaved with optional point-wise non-linearities can be applied. At the end of the network a feature representation on $\mathbb{R}^d$ can be provided by pooling over $G$. In short, we provide a pure self-attention analogous to Cohen & Welling (2016). However, as the group acts directly on the positional encoding, our networks are **steerable** as well (Weiler et al., 2018b). This allows us to go beyond group discretizations that live in the grid without introducing interpolation artifacts (§5.1).

Though theoretically sound, neural architectures using regular representations are unable to handle continuous groups directly in practice. This is a result of the summation over elements $h \in H$ in Eq. 15, which becomes an integral for continuous groups. Interestingly, using discrete groups does not seem to be detrimental in practice. Our experiments indicate that performance saturates for fine discrete approximations of the underlying continuous group (Tab. 2). In fact, (Weiler & Cesa, 2019, Tab. 3) show via extensive experiments that networks using regular representations and fine enough discrete approximations consistently outperform networks handling continuous groups via irreducible representations. We conjecture this is a result of the networks receiving discrete signals as input. As the action of several group elements fall within the same pixel, no further improvement can be obtained.

5.1 Group Self-Attention is an Steerable Operation

Convolutional filters are commonly parameterized by weights on a discrete grid, which approximate the function implicitly described by the filter at the grid locations. Unfortunately, for groups whose action does not live in this grid, e.g., 45° rotations, the filter must be interpolated. This is problematic as these filters are typically small and the resulting interpolation artifacts can be severe (Fig. 2a). **Steerable CNNs** tackle this problem by parameterizing convolutional filters on a continuous basis on which the action of the group is well-defined, e.g., *circular harmonics* (Weiler et al., 2018b), *B-splines*.

---

4This phenomenon arises from the fact that $\mathbb{R}^2$ is a quotient of the roto-translation group. Consequently, imposing group equivariance in the quotient space is equivalent to imposing an additional homomorphism of constant value over its cosets. Conclusively, the resulting map is of constant value over the rotation elements and, thus, is not able to discriminate among them. See Ch. 3.1 of Dummit & Foote (2004) for an intuitive description.
We have demonstrated that it is sufficient to define self-attention as a function on the group where (group) convolutions are not well-defined, e.g., sets, and it is equivalent to Kondor & Trivedi. This statement is more general than that of Kondor & Trivedi (2018), as it holds for data structures varying positional encodings $\rho$.

Group self-attention (Fig. B.2). Group self-attention is a generalization of Kondor & Trivedi (2018)'s statement as:

$$L^r[\rho](i, \tilde{h}) = \mathcal{L}_h[\rho](i, \tilde{h})$$

Lifting self-attention (Fig. B.1). Let $\mathcal{G} = \mathbb{R}^d \rtimes \mathcal{H}$ be an affine group (Def. C.3) acting on $\mathbb{R}^d$. The lifting self-attention $m^r_{\mathcal{G}}[f, \rho] : L_v(\mathbb{R}^d) \to L_v(\mathcal{G})$ is a map from functions on $\mathbb{R}^d$ to functions on $\mathcal{G}$ obtained by modifying the relative positional encoding $\rho(i, j)$ by the action of group elements $h \in \mathcal{H}$: $L_h[\rho](i, j) = \rho^r(h^{-1}x(j) - h^{-1}x(i))$. It corresponds to the concatenation of multiple self-attention operations (Eq. 11) indexed by $h$ with varying positional encodings $\mathcal{L}_h[\rho] : m^r_{\mathcal{G}}[f, \rho](i, h) = m^r[f, L_h[\rho]](i)$

$$= \varphi_{\text{out}} \left( \sum_{h \in \mathcal{H}} \sigma_i \left( \langle \varphi^{(h)}_{\text{key}}(f(i)), \varphi^{(h)}(f(i) + L_h[\rho](i, j)) \rangle \varphi^{(h)}(f(j)) \right) \right)$$

Proposition 5.1. Lifting self-attention is $\mathcal{G}$-equivariant. That is, it holds that: $m^r_{\mathcal{G}}[L_{\rho}f, \rho](i, \tilde{h}) = \mathcal{L}_g[m^r_{\mathcal{G}}[f, \rho]](i, \tilde{h})$.

Group self-attention (Fig. B.2). Let $\mathcal{G} = \mathbb{R}^d \rtimes \mathcal{H}$ be an affine group acting on itself and $f(i, \tilde{h}) \in L_v(\mathcal{G})$, $i \in S$, $\tilde{h} \in \mathcal{H}$, be a function defined on a set immersed with the structure of the group $\mathcal{G}$. That is, enriched with a positional encoding $\rho(i, \tilde{h}((j, \tilde{h}))) = \rho^r((x(j) - x(i), \tilde{h}^{-1} \tilde{h}))$, $i, j \in S$, $\tilde{h}, \tilde{h} \in \mathcal{H}$.

The group self-attention $m^r_{\mathcal{G}}[f, \rho] : L_v(\mathcal{G}) \to L_v(\mathcal{G})$ is a map from functions on $\mathcal{G}$ to functions on $\mathcal{G}$ obtained by modifying the group positional encoding by the action of group elements $h \in \mathcal{H}$: $L_h[\rho](i, \tilde{h}((j, \tilde{h}))) = \rho^r(h^{-1}(x(j) - x(i)), \tilde{h}^{-1} \tilde{h})$. It corresponds to the concatenation of multiple self-attention operations (Eq. 11) indexed by $h$ with varying positional encodings $L_h[\rho]$ and followed by a summation over the output domain along $\tilde{h}$:

$$m^r_{\mathcal{G}}[f, \rho](i, \tilde{h}) = \sum_{h \in \mathcal{H}} \sigma_i \left( \langle \varphi^{(h)}_{\text{key}}(f(i, \tilde{h})), \varphi^{(h)}(f(j, \tilde{h})) \rangle \right)$$

$$= \varphi_{\text{out}} \left( \sum_{h \in \mathcal{H}} \sigma_i \left( \langle \varphi^{(h)}_{\text{key}}(f(i, \tilde{h})), \varphi^{(h)}(f(j, \tilde{h})) \rangle \right) \right)$$

In contrast to vanilla and lifting self-attention, the group self-attention neighborhood $H(i, \tilde{h})$ is now defined on the group. This allows lifting across group transformations, e.g., rotations.

Proposition 5.2. Group self-attention is $\mathcal{G}$-equivariant. That is, it holds that: $m^r_{\mathcal{G}}[L_{\rho}f, \rho](i, \tilde{h}) = \mathcal{L}_g[m^r_{\mathcal{G}}[f, \rho]](i, \tilde{h})$.

Non-unimodular groups, i.e., groups that modify the volume of the objects they act upon, such as the dilation group, require a special treatment. This treatment is provided in Appx. E.

5.3 GROUP SELF-ATTENTION IS A GENERALIZATION OF THE GROUP CONVOLUTION

We have demonstrated that it is sufficient to define self-attention as a function on the group $\mathcal{G}$ and ensure that $L_h[\rho] = \rho \forall g \in \mathcal{G}$ in order to enforce $\mathcal{G}$-equivariance. Interestingly, this observation is inline with the main statement of Kondor & Trivedi (2018) for (group) convolutions: “the group convolution on $\mathcal{G}$ is the only (unique) $\mathcal{G}$-equivariant linear map”. In fact, our finding can be formulated as a generalization of Kondor & Trivedi (2018)’s statement as:

“Linear mappings on $\mathcal{G}$ whose positional encoding is $\mathcal{G}$-invariant are $\mathcal{G}$-equivariant.”

This statement is more general than that of Kondor & Trivedi (2018), as it holds for data structures where (group) convolutions are not well-defined, e.g., sets, and it is equivalent to Kondor & Trivedi (2018).
We conjecture that 300 epochs are insufficient for all models to converge equally well. Unfortunately, we perform experiments on three image benchmark datasets for which particular forms of equivariance are desirable. Statement (i) follows from the fact that any convolutional layer can be described as a multi-head self-attention layer provided enough heads (Cordonnier et al., 2020), yet self-attention often uses larger receptive fields. As a result, self-attention is able to describe a larger set of functions than convolutions, e.g., Fig. A.1. Cordonnier et al. (2020)’s statement can be seamlessly extended to group self-attention by incorporating an additional dimension corresponding to \( \mathcal{H} \) in their derivations, and defining neighborhoods in this new space with a proportionally larger number of heads. Statement (ii) stems from the finding of Ravanbakhsh (2020) that functions induced by regular group representations are equivariant universal approximators provided full kernels, i.e., global receptive fields. Global receptive fields are required to guarantee that the equivariant map is able to model any dependency among input components. Global receptive fields are readily utilized by our proposed regular global group self-attention and, provided enough heads, one can ensure that any such dependencies is properly modelled.

6 EXPERIMENTS

We perform experiments on three image benchmark datasets for which particular forms of equivariance are desirable.\(^5\) We evaluate our approach by contrasting GSA-Nets equivariant to multiple symmetry groups. Additionally, we conduct an experiment on rotMNIST to evaluate the performance of GSA-Nets as a function of the neighborhood size. All our networks follow the structure shown in Fig. F.1 and vary only in the number of blocks and channels. We emphasize that both the architecture and the number of parameters in GSA-Nets is left unchanged regardless of the group used. Our results illustrate that GSA-Nets consistently outperform equivalent non-equivariant attention networks. We further compare GSA-Nets with convolutional architectures. Though our approach does not build upon these networks, this comparison provides a fair view to the yet present gap between self-attention and convolutional architectures in vision tasks, also present in their group equivariant counterparts.

**Efficient implementation of lifting and group self-attention.** Our self-attention implementation takes advantage of the fact that the group action only affects the positional encoding \( \mathbf{P} \) to reduce the total computational cost of the operation. Specifically, we calculate self-attention scores w.r.t. the content \( \mathbf{X} \) once, and reuse them for all transformed versions of the positional encoding \( \{ \mathcal{L}_k[\rho] \}_{k \in \mathcal{H}} \).

**Model designation.** We refer to translation equivariant self-attention models as Z2_SA. Reflection equivariant models receive the keyword \( \mathbb{H} \), e.g., Z2M_SA, and rotation equivariant models the keyword \( \mathbb{R}_n \), where \( n \) depicts the angle discretization. For example, R8_SA depicts a model equivariant to rotations by 45 degrees. Specific model architectures are provided in Appx. F.

**RotMNIST.** The rotated MNIST dataset (Larochelle et al., 2007) is a classification dataset often used as a standard benchmark for rotation equivariance. It consists of 62k gray-scale 28x28 uniformly rotated handwritten digits, divided into training, validation and test sets of 10k, 2k and 50k images. First, we study the effect of the neighborhood size on classification accuracy and convergence time. We train R4_SA networks for 300 epochs with vicinities \( \times \times N \) of varying size (Tab. 1, Fig. 3). Since GSA-Nets optimize where to attend, the complexity of the optimization problem grows as a function of \( N \). Consequently, models with big vicinities are expected to converge slower. However, as the family of functions describable by big vicinities contains those describable by small ones, models with big vicinities are expected to be at least as good upon convergence. Our results show that models with small vicinities do converge much faster (Fig. 3). However, though some models with large vicinities do outperform models with small ones, e.g., \( 7 \times 7 \) vs. \( 3 \times 3 \), a trend of this behavior is not apparent. We conjecture that 300 epochs are insufficient for all models to converge equally well. Unfortunately, due to computational constraints, we were not able to perform this experiment for a larger number of epochs. We consider an in-depth study of this behavior an important direction for future work.

Next, we compare GSA-Nets equivariant to translation and rotation at different angle discretizations (Tab. 2). Based on the results of the previous study, we select a 5x5 neighborhood, as it provides the

\(^5\)Our code is publicly available at [https://github.com/dwromero/g_selfatt](https://github.com/dwromero/g_selfatt).
Table 1: Accuracy vs. neighborhood size.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>NBHD. SIZE</th>
<th>ACC. (%)</th>
<th>TRAIN. TIME / EPOCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>3x3</td>
<td>96.56</td>
<td></td>
<td>04:53 - 1GPU</td>
</tr>
<tr>
<td>5x5</td>
<td>97.49</td>
<td></td>
<td>05:34 - 1GPU</td>
</tr>
<tr>
<td>7x7</td>
<td>97.33</td>
<td>09:03 - 1GPU</td>
<td></td>
</tr>
<tr>
<td>9x9</td>
<td>97.42</td>
<td>09:16 - 1GPU</td>
<td></td>
</tr>
<tr>
<td>R4_SA</td>
<td>11x11</td>
<td>97.17</td>
<td>12:09 - 1GPU</td>
</tr>
<tr>
<td>15x15</td>
<td>96.89</td>
<td>10:27 - 2GPU</td>
<td></td>
</tr>
<tr>
<td>19x19</td>
<td>96.86</td>
<td>14:47 - 2GPU</td>
<td></td>
</tr>
<tr>
<td>23x23</td>
<td>97.05</td>
<td>06:13 - 3GPU</td>
<td></td>
</tr>
<tr>
<td>28x28</td>
<td>96.81</td>
<td>12:12 - 4GPU</td>
<td></td>
</tr>
</tbody>
</table>

Cohen & Welling (2016).

Table 2: Classification results. All convolutional architectures use 3x3 filters.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ACC. (%)</th>
<th>PARAMS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>22_SA</td>
<td>96.37</td>
<td></td>
</tr>
<tr>
<td>R4_SA</td>
<td>97.46</td>
<td>44.67k</td>
</tr>
<tr>
<td>R12_SA</td>
<td>97.97</td>
<td></td>
</tr>
<tr>
<td>R16_SA</td>
<td>97.66</td>
<td></td>
</tr>
<tr>
<td>22_CNN</td>
<td>94.97</td>
<td>21.75k</td>
</tr>
<tr>
<td>R4_CNN</td>
<td>98.21</td>
<td>77.54k</td>
</tr>
<tr>
<td>αR4_CNN</td>
<td>98.31</td>
<td>73.13k</td>
</tr>
</tbody>
</table>

Cohen & Welling (2016); Romero et al. (2020a).

best trade-off between accuracy and convergence time. Our results show that finer discretizations lead to better accuracy but saturates around R12. We conjecture that this is due to the discrete resolution of the images in the dataset, which leads finer angle discretizations to fall within the same pixel.

**CIFAR-10.** The CIFAR-10 dataset (Krizhevsky et al., 2009) consists of 60k real-world 32x32 RGB images uniformly drawn from 10 classes, divided into training, validation and test sets of 40k, 10k and 10k images. Since reflection is a symmetry that appears ubiquitously in natural images, we compare GSA-Nets equivalent to translation and reflection in this dataset (Tab. 2). Our results show that reflection equivariance indeed improves the classification performance of the model.

**PCam.** The PatchCamelyon dataset (Veeling et al., 2018) consists of 327k 96x96 RGB image patches of tumorous/non-tumorous breast tissues extracted from Camelyon16 (Bejnordi et al., 2017). Each patch is labeled as tumorous if the central region (32x32) contains at least one tumour pixel. As cells appear at arbitrary positions and poses, we compare GSA-Nets equivariant to translation, rotation and reflection (Tab. 2). Our results show that incorporating equivariance to reflection in addition to rotation, as well as providing finer group discretization, improve classification performance.

## 7 DISCUSSION AND FUTURE WORK

Though GSA-Nets perform competitively to G-CNNs for some tasks, G-CNNs still outperforms our approach in general. We conjecture that this is due to the harder nature of the optimization problem in GSA-Nets and the carefully crafted architecture design, initialization and optimization procedures developed for CNNs over the years. Though our theoretical results indicate that GSA-Nets can be more expressive than G-CNNs (§ 5.3), further research in terms of design, optimization, stability and generalization is required. These are in fact open questions for self-attention in general (Xiong et al., 2020; Liu et al., 2020; Zhao et al., 2020) and developments in this direction are of utmost importance.

The main drawback of our approach is the quadratic memory and time complexity typical of self-attention. This is an active area of research, e.g., Kitaev et al. (2020); Wang et al. (2020); Zaheer et al. (2020); Choromanski et al. (2020) and we believe that efficiency advances to vanilla self-attention can be seamlessly integrated in GSA-Nets. Our theoretical results indicate that GSA-Nets have the potential to become the standard solution for applications exhibiting symmetries, e.g., medical imagery. In addition, as self-attention is a set operation, GSA-Nets provide straightforward solutions to set-like data types, e.g., point-clouds, graphs, symmetric sets, which may benefit from additional geometrical information, e.g., Fuchs et al. (2020); Maron et al. (2020). Finally, we hope our theoretical insights serve as a support point to further explore and understand the construction of equivariant maps for graphs and sets, which often come equipped with spatial coordinates: a type of positional encoding.
ACKNOWLEDGMENTS

We gratefully acknowledge Michael Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, and Hyunjik Kim for useful discussions, Robert-Jan Bruintjes, Fabian Fuchs, Erik Bekkers, Andreas Loukas, Mark Hoogendoorn and our anonymous reviewers for their valuable comments on early versions of this work which largely helped us to improve the quality of our work.

David W. Romero is financed as part of the Efficient Deep Learning (EDL) programme (grant number P16-25), partly funded by the Dutch Research Council (NWO) and Semiotic Labs. Jean-Baptiste Cordonnier is financed by the Swiss Data Science Center (SDSC). Both authors thankfully acknowledge everyone involved in funding this work. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

REFERENCES


APPENDIX

A Convolution and the Self-Attention: A Graphical Comparison

Figure A.1: Parameter usage in convolutional kernels (Fig. A.1a) and self-attention (Fig. A.1b). Given a budget of 9 parameters, a convolutional filter ties these parameters to specific positions. Subsequently, these parameters remain static regardless of (i) the query input position and (ii) the input signal itself. Self-attention, on the other hand, does not tie parameters to any specific positions at all. Contrarily, it compares the representations of all tokens falling in its receptive field. As a result, provided enough heads, self-attention is more general than convolutions, as it can represent any convolutional kernel, e.g., Fig. A.1a, as well as several other functions defined on its receptive field.

B Lifting and Group Self-Attention: A Graphical Description

Figure B.1: Lifting self-attention on the roto-translation group for discrete rotations by 90 degrees (also called the $\mathcal{Z}_4$ group). The $\mathcal{Z}_4$ group is defined as $\mathcal{H} = \{ e, h, h^2, h^3 \}$, where $h$ depicts a 90° rotation. The lifting self-attention corresponds to the concatenation of $|\mathcal{H}| = 4$ self-attention operations between the input $f$ and $h$-transformed versions of the positional encoding $\mathcal{L}[\rho]$, $\forall h \in \mathcal{H}$. As a result, the model “sees” the input $f$ at each of the rotations in the group at once. Since $\mathcal{Z}_4$ is a cyclic group, i.e., $h^4 = e$, functions on this group are often represented as responses on a ring (right side of the image). This is a self-attention analogous to the regular lifting group convolution broadly utilized in group equivariant learning literature, e.g., Cohen & Welling (2016); Romero et al. (2020a).
C  CONCEPTS FROM GROUP THEORY

Definition C.1 (Group). A group is an ordered pair \((G, \cdot)\) where \(G\) is a set and \(\cdot : G \times G \to G\) is a binary operation on \(G\), such that (i) the set is closed under this operation, (ii) the operation is associative, i.e., \((g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)\), \(g_1, g_2, g_3 \in G\), (iii) there exists an identity element \(e \in G\) s.t. \(\forall g \in G\) we have \(e \cdot g = g \cdot e = g\), and (iv) for each \(g \in G\), there exists an inverse \(g^{-1}\) s.t. \(g \cdot g^{-1} = e\).

Definition C.2 (Subgroup). Let \((G, \cdot)\) be a group. A subset \(H\) of \(G\) is a subgroup of \(G\) if \(H\) is nonempty and closed under the group operation and inverses (i.e., \(h_1, h_2 \in H\) implies that \(h_1^{-1} \in H\) and \(h_1, h_2 \in H\)). If \(H\) is a subgroup of \(G\) we write \(H \leq G\).

Definition C.3 (Semi-direct product and affine groups). In practice, one is mainly interested in the analysis of data defined on \(\mathbb{R}^d\), and, consequently, in groups of the form \(G = \mathbb{R}^d \rtimes H\), resulting from the semi-direct product \((\rtimes)\) between the translation group \((\mathbb{R}^d, +)\) and an arbitrary (Lie) group \(H\) that acts on \(\mathbb{R}^d\), e.g., rotation, scaling, mirroring, etc. This family of groups is referred to as affine groups and their group product is defined as:

\[
g_1 \cdot g_2 = (x_1, h_1) \cdot (x_2, h_2) = (x_1 + h_1 \circ x_2, h_1 \cdot h_2),
\]

with \(g = (x_1, h_1), g_2 = (x_2, h_2) \in G, x_1, x_2 \in \mathbb{R}^d\) and \(h_1, h_2 \in H\). The operator \(\circ\) denotes the action of \(h \in H\) on \(x \in \mathbb{R}^d\), and it describes how a vector \(x \in \mathbb{R}^d\) is modified by elements \(h \in H\).

Definition C.4 (Group representation). Let \(G\) be a group and \(L^2(X)\) be a space of functions defined on some vector space \(X\). The (left) regular group representation of \(G\) is a linear transformation \(\mathcal{L} : G \times L^2(X) \to L^2(X), (g, f) \mapsto \mathcal{L}_g[f] = f(g^{-1} \circ x)\), that shares the group structure via:

\[
\mathcal{L}_{g_1} \mathcal{L}_{g_2} [f] = \mathcal{L}_{g_1 \cdot g_2} [f]
\]

for any \(g_1, g_2 \in G, f \in L^2(X)\). That is, concatenating two such transformations, parameterized by
A solution to this problem is to scale our neighborhood with a corresponding group representation on $L$ which the set element has been gathered. The lifting and group self-attention formulation provided in §5.2 are only valid for unimodular problematic, as the response produced by the same circle, would still be different for different scales. Specifically, this result would always be bigger for larger versions of the neighborhood. This is $\sum$ problematic: self-attention utilizes summations over its neighborhood. Since the circle would fall within the neighborhood for any scale factor $h \in \mathbb{R}_0^+$ of these transformations. If we scale the circle by a factor of $R$ for which the original circle fits perfectly. Since the size of the circle is not modified for any $h \in \mathbb{R}_0^+$ kept constant. If we increase its radius by a factor $R$. Imagine we have a circle on $H = \mathbb{R}_0^+$. To illustrate why this distinction is important, consider the following example: $HE$ $ET$ $E$ $ON$ $PACES$ $FOR$ $GROUPE$ $S$ $ACTING$ $ON$ $HOMOGENEOUS$ $SPACES$ $FOR$ $FUNCTIONS$ $DEFINED$ $ON$ $SETS$ In this section we show that the action of a group $G$ acting on a homogeneous space $X$ is well defined on sets $S$ gathered from $X$, and that it induces a group representation of functions defined on $S$. Let $S = \{i\}$ be a set and $X$ be a homogeneous space on which the action of $G$ is well-defined, i.e., $gx \in X$, $\forall g \in G$, $\forall x \in X$. Since $S$ has been gathered from $X$, there exists an injective map $x : S \to X$, that maps set elements $i \in S$ to unique elements $x_i \in X$. That is, there exists a map $x : i \mapsto x_i$ that assigns an unique value $x_i \in X$ to each $i \in S$ corresponding to the coordinates from which the set element has been gathered.

Since the action of $G$ is well defined on $X$, it follows that the left regular representation (Def. C.4) $L_g[f^X](x_i) = f^X(g^{-1}x_i) \in L_{ij}(X)$ of functions $f^X \in L_{ij}(X)$ exists and is well-defined. Since $x$ is injective, the left regular representation $L_g[f^X](x_i) = f^X(g^{-1}x_i)$ can be expressed uniquely in terms of set indices as $L_g[f^X](x_i) = f^X(g^{-1}x_i) = f^X(g^{-1}x(i))$. Furthermore, its inverse $x^{-1} : X \to S$, $x^{-1} : x_i \mapsto i$ also exist and is well-defined. As a consequence, points $x_i \in X$ can be expressed uniquely in terms of set indices as $i = x^{-1}(x_i)$, $i \in S$. Consequently, functions $f^X \in L_{ij}(X)$ can be expressed in terms of functions $f \in L_{ij}(S)$ by means of the equality $f^X(i) = f(x^{-1}(x_i))$. Resultantly, we see that the group representation $L_g[f^X](x_i) = f^X(g^{-1}x(i))$ can be described in terms of functions $f \in L_{ij}(S)$ as:

$$L_g[f^X](x_i) = f^X(g^{-1}x(i)) = f(x^{-1}(g^{-1}x(i))) = L_g[f](i),$$

with a corresponding group representation on $L_{ij}(S)$ given by $L_g[f](i) = f(x^{-1}(g^{-1}x(i)))$, and an action of group elements $g \in G$ on set elements $i$ given by $gi = x^{-1}(gx(i))$.

E  THE CASE OF NON-UNIMODULAR GROUPS: SELF-ATTENTION ON THE DILATION-TRANSLATION GROUP

The lifting and group self-attention formulation provided in §5.2 are only valid for unimodular groups. That is, for groups whose action does not change the volume of the objects they act upon, e.g., rotation, mirroring, etc. Non-unimodular groups, however, do modify the volume of the acted objects (Bekkers, 2020). The most relevant non-unimodular group for this work is the dilation group $H = (\mathbb{R}_0^+, \times)$. To illustrate why this distinction is important, consider the following example:

Imagine we have a circle on $\mathbb{R}^2$ of area $\pi r^2$. If we rotate, mirror or translate the circle, its size is kept constant. If we increase its radius by a factor $h \in \mathbb{R}_0^+$, however, its size would increase by $h^2$. Imagine that we have an application for which we would like to recognize this circle regardless of any of these transformations by means of self-attention. For this purpose, we define a neighborhood $N$ for which the original circle fits perfectly. Since the size of the circle is not modified for any translated, rotated or translated versions of it, we would still be able to detect the circle regardless of these transformations. If we scale the circle by a factor of $h > 1$, however, the circle would fall outside of our neighborhood $N$ and hence, we would not be able to recognize it.

A solution to this problem is to scale our neighborhood $N$ in a proportional way. That is, if the circle is scaled by a factor $h \in \mathbb{R}_0^+$, we scale our neighborhood by the same factor $h$: $N \to hN$. Resultantly, the circle would fall within the neighborhood for any scale factor $h \in \mathbb{R}_0^+$. Unfortunately, there is a problem: self-attention utilizes summations over its neighborhood. Since $\sum_{i \in hN} i > \sum_{i \in hN} i$, for $h > 1$, and $\sum_{i \in hN} i < \sum_{i \in hN} i$, for $h < 1$, the result of the summations would still differ for different scales. Specifically, this result would always be bigger for larger versions of the neighborhood. This is problematic, as the response produced by the same circle, would still be different for different scales.
In order to handle this problem, one utilizes a normalization factor proportional to the change of size of the neighborhood considered. This ensures that the responses are equivalent for any scale $h \in \mathbb{R}_{>0}$. That is, one normalizes all summations proportionally to the size of the neighborhood. As a result, we obtain that $\sum_{i \in \mathcal{N}(h^{-1})} i = \sum_{i \in \mathcal{N}(h^{-2})} i, \forall h, h_1, h_2 \in \mathbb{R}_{>0}$.

In the example above we have provided an intuitive description of the (left invariant) Haar measure $d\mu(h)$. As its name indicates, it is a measure defined on the group, which is invariant over all group elements $h \in \mathcal{H}$. For several unimodular groups, the Haar measure corresponds to the Lebesgue measure as the volume of the objects the group acts upon is kept equal, i.e., $d\mu(h) = dh$.

For non-unimodular groups, however, the Haar measure requires a normalization factor proportional to the change of volume of these objects. Specifically, the Haar measure corresponds to the Lebesgue measure times a normalization factor $h^d$, where $d$ corresponds to the dimensionality of the space $\mathbb{R}^d$ the group acts upon (Bekkers, 2020; Romero et al., 2020b), i.e., $d\mu(h) = \frac{1}{|\mathcal{H}|}dh$.

In conclusion, in order to obtain group equivariance to non-unimodular groups, the lifting and group self-attention formulation provided in Eqs. 14, 16 must be modified via normalization factors proportional to the group elements $h \in \mathcal{H}$. Specifically, they are redefined as:

$$m_{G_r}^h[f, \rho](i, h) = \varphi_{\text{out}}\left( \sum_{h \in \mathcal{H}} h_i \sum_{h \in \mathcal{H}} \frac{1}{|\mathcal{H}|} \sigma_j \left( \langle \varphi_{\text{out}}^{(h)}(f(i)), \varphi_{\text{key}}^{(h)}(f(i)) + \mathcal{L}_h(\rho)(i, j) \rangle \right) \varphi_{\text{val}}^{(h)}(f(j)) \right)$$

$$m_{G_r}^h[f, \rho](i, h) = \varphi_{\text{out}}\left( \sum_{h \in \mathcal{H}} h_i \sum_{h \in \mathcal{H}} \frac{1}{|\mathcal{H}|} \sigma_j \left( \langle \varphi_{\text{out}}^{(h)}(f(i)), \varphi_{\text{key}}^{(h)}(f(j)) \rangle \right) \varphi_{\text{val}}^{(h)}(f(j)) \right)$$

The factor $d + 1$ in Eq. 21 results from the fact that the summation is now performed on the group $G_r = \mathbb{R}^{d+1} \times \mathcal{H}$, an space of dimensionality $d + 1$. An interesting case emerges when global neighborhoods are considered, i.e., s.t. $\mathcal{N}(i) = \mathcal{S}, \forall i \in \mathcal{S}$. Since $h\mathcal{N}(i) = \mathcal{N}(i) = \mathcal{S}$ for any $h > 1$, approximation artifacts are introduced. It is not clear if it is better to introduce normalization factors in these situations or not. An in-depth investigation of this phenomenon is left for future research.

### E.1 Current Empirical Aspects of Scale Equivariant Self-Attention

Self-attention suffers from quadratic memory and time complexity proportional to the size of the neighborhood considered. This constraint is particularly important for the dilation group, for which these neighborhoods grow as a result of the group action. We envisage two possible solutions to this limitation left out for future research:

The most promising solution is given by incorporating recent advances in efficient self-attention in group self-attention, e.g., Kitaev et al. (2020); Wang et al. (2020); Zaheer et al. (2020); Katharopoulos et al. (2020); Choromanski et al. (2020). By reducing the quadratic complexity of self-attention, the current computational constraints of scale equivariant self-attention can be (strongly) reduced. Importantly, resulting architectures would be comparable to Bekkers (2020); Sosnowik et al. (2020); Romero et al. (2020b) in terms of their functionality and the group discretizations they can manage.

The second option is to draw a self-attention analogous to Worrall & Welling (2019), where scale equivariance is implemented via dilated convolutions. One might consider an analogous to dilated convolutions via "sparse" dilations of the self-attention neighborhood. As a result, scale equivariance can be implemented while retaining an equal computational cost for all group elements. Importantly however, this strategy is viable for a dyadic set of scales only, i.e., a set of scales given by a set $\{2^n\}_{n=0}^{\text{max}}$, and thus, less general that the scale-equivariant architectures listed before.

### F Experimental Details

In this section we provide extended details over our implementation as well as the exact architectures and optimization schemes used in our experiments. All our models follow the structure shown in Fig. F.1 and vary only in the number of blocks and channels. All self-attention operations utilize 9 heads. We utilize PyTorch for our implementation. Any missing specification can be safely considered to be the PyTorch default value. Our code is publicly available at https://github.com/dwromero/g_selfatt.

---

6The squared factor in $h_1^2$ and $h_2^2$ appears as a result that the neighborhood growth is quadratic in $\mathbb{R}^2$.

7This is why this subtlety is often left out in group equivariance literature.
Published as a conference paper at ICLR 2021

Figure F.1: Graphical description of group self-attention networks. Dot-lined blocks depict optional blocks. Linear layers are applied point-wise across the feature map. Swish non-linearities (Ramachandran et al., 2017) and layer normalization (Ba et al., 2016) are used all across the network. The GlobalPooling block consists of max-pool over group elements followed by spatial mean-pool.

F.1 ROTATED MNIST

For rotational MNIST we use a group self-attention network composed of 5 attention blocks with 20 channels. We utilize automatic mixed precision during training to reduce memory requirements. attention_dropout_rate and value_dropout_rate are both set to 0.1. We train for 300 epochs and utilize the Adam optimizer, batch size of 8, weight decay of 0.0001 and learning rate of 0.001.

F.2 CIFAR-10

For CIFAR-10 we use a group self-attention network composed of 6 attention blocks with 96 channels for the first two blocks and 192 channels for the rest. attention_dropout_rate and value_dropout_rate are both set to 0.1. We use dropout on the input with a rate of 0.3 and additional dropout blocks of rate 0.2 followed by spatial max-pooling after the second and fourth block. We did not use automatic mixed precision training for this dataset as it made all models diverge. We perform training for 350 epochs and utilize stochastic gradient descent with a momentum of 0.9 and cosine learning rate scheduler with base learning rate 0.01 (Loshchilov & Hutter, 2016). We utilize a batch size of 24, weight decay of 0.0001 and He’s initialization.

F.3 PATCHCAMELYON

For PatchCamelyon we use a group self-attention network composed of 4 attention blocks with 12 channels for the first block, 24 channels for the second block, 48 channels for the third and fourth blocks and 96 channels for the last block. attention_dropout_rate and value_dropout_rate are both set to 0.1. We use an additional max-pooling block after the lifting block to reduce memory requirements. We did not use automatic mixed precision training for this dataset as it made all models diverge. We perform training for 100 epochs, utilize stochastic gradient descent with a momentum of 0.9 and cosine learning rate scheduler with base learning rate 0.01 (Loshchilov & Hutter, 2016). We utilize a batch size of 8, weight decay of 0.0001 and He’s initialization.

G PROOFS

Proof of Proposition 4.1. If the self-attention formulation provided in Eqs. 3, 8, is permutation equivariant, then it must hold that \( m[L_\pi[f]](i) = L_\pi[m[f]](i) \). Consider a permuted input signal \( L_\pi[f](i) = f(\pi^{-1}(i)) \). The self-attention operation on \( L_\pi[f] \) is given by:

\[
m[L_\pi[f]](i) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{j \in S} \sigma_j \langle \varphi_{\text{qry}}(L_\pi[f](i)), \varphi_{\text{key}}(L_\pi[f](j)) \rangle \varphi_{\text{val}}(L_\pi[f](j)) \right)
\]

\[
= \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{j \in S} \sigma_j \langle \varphi_{\text{qry}}(f(\pi^{-1}(i))), \varphi_{\text{key}}(f(\pi^{-1}(j))) \rangle \varphi_{\text{val}}(f(\pi^{-1}(j))) \right)
\]

18
Translation equivariance.

If the self-attention formulation provided in Eq. 10 is permutation equivariant, then it must hold that $m[\mathcal{L}_x[f], \rho](i) = L_x[m[f, \rho]](i)$. Consider a permuted input signal $\mathcal{L}_x[f](i) = f(\pi^{-1}(i))$. The self-attention operation on $\mathcal{L}_x[f]$ is given by:

$$
m[\mathcal{L}_x[f], \rho](i) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{j \in S} \sigma_j \left( \varphi_{\text{qy}}^h \left( f(\pi^{-1}(i)) + \rho(i) \right) \right) \varphi_{\text{key}}^h \left( f(\pi^{-1}(j)) + \rho(j) \right) \right) \varphi_{\text{val}}^h \left( f(\pi^{-1}(j)) \right)
$$

As discussed in §3.2, since there exists permutations in $S_N$ able to send elements $j$ in $\mathcal{N}(i)$ to elements $\tilde{j}$ outside of $\mathcal{N}(i)$, it is trivial to show that Eq. 10 is not equivariant to $S_N$. Consequently, in order to provide a more interesting analysis, we consider global attention here, i.e., cases where $\mathcal{N}(i) = S$. As shown for Proposition 4.1, this self-attention instantiation is permutation equivariant. Consequently, by considering this particular case, we are able to explicitly analyze the effect of introducing absolute positional encodings into the self-attention formulation. We have then that:

$$
m[\mathcal{L}_x[f], \rho](i) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{j \in S} \sigma_j \left( \varphi_{\text{qy}}^h \left( f(\pi^{-1}(i)) + \rho(i) \right) \right) \varphi_{\text{key}}^h \left( f(\pi^{-1}(j)) + \rho(j) \right) \right) \varphi_{\text{val}}^h \left( f(\pi^{-1}(j)) \right)
$$

Here we have used the substitution $i = \pi(i)$ and $\tilde{j} = \pi(j)$. Since the summation is defined over the entire set we have that $\sum_{\pi(j) \in S}[] = \sum_{j \in S}[]$. Consequently, we conclude that absolute position-aware self-attention is not permutation equivariant.

Translation equivariance. If the self-attention formulation provided in Eq. 10, is translation equivariant, then it must hold that $m(\mathcal{L}_y[f], \rho)(i) = \mathcal{L}_y[m(f, \rho)](i)$. Consider a translated input signal $\mathcal{L}_y[f](i) = f(x^{-1}(x(i) - y))$. The self-attention operation on $\mathcal{L}_y[f]$, is given by:

$$
m[\mathcal{L}_y[f], \rho](i) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{j \in \mathcal{N}(i)} \sigma_j \left( \varphi_{\text{qy}}^h \left( \mathcal{L}_y[f](i) + \rho(i) \right) \right) \varphi_{\text{key}}^h \left( \mathcal{L}_y[f](j) + \rho(j) \right) \right) \varphi_{\text{val}}^h \left( \mathcal{L}_y[f](j) \right)
$$

Published as a conference paper at ICLR 2021
Here, we have used the substitution $\tilde{i} = x^{-1}(i) - y \Rightarrow i = x^{-1}(\tilde{i} + y)$ and $\tilde{j} = x^{-1}(j) - y \Rightarrow j = x^{-1}(\tilde{j} + y)$. Since the area of summation remains equal to any translation $y \in \mathbb{R}^d$, we have:

$$
\sum [\cdot]_{x^{-1}(\tilde{j})+y \in \mathbb{H}(x^{-1}(\tilde{j})+y)} = \sum [\cdot]_{x^{-1}(\tilde{i})+y \in \mathbb{H}(x^{-1}(\tilde{i}))} = \sum [\cdot].
$$

Hence, we can further reduce the expression above as:

$$
m_r[\mathcal{L}_y[f], \rho](i) = \varphi_{\text{out}} \left( \sum_{h \in \mathbb{H}} \sum_{j \in \mathbb{H}(i)} \sigma_j(\varphi_{\text{query}}(f(\tilde{i})), \varphi_{\text{key}}(\mathcal{L}_y[f](i) + \rho(i, j))) \varphi_{\text{val}}(f(\tilde{j})) \right)
$$

Since, $\rho(\tilde{i}) + y \neq \rho(\tilde{j})$ and $\rho(\tilde{j}) + y \neq \rho(\tilde{j})$, we are unable to reduce the expression further towards the form of $m(f, \rho)(i)$. Consequently, we conclude that the absolute positional encoding does not allow for translation equivariance either.

\[\square\]

**Proof of Claim 4.3.** If the self-attention formulation provided in Eq. 11 is translation equivariant, then it must hold that $m_r[\mathcal{L}_y[f], \rho](i) = \mathcal{L}_y[m_r[f, \rho]](i)$. Consider a translated input signal $\mathcal{L}_y[f](i) = f(x^{-1}(x(i) - y))$. The self-attention operation on $\mathcal{L}_y[f]$ is given by:

$$
m_r[\mathcal{L}_y[f], \rho](i) = \varphi_{\text{out}} \left( \sum_{h \in \mathbb{H}} \sum_{j \in \mathbb{H}(i)} \sigma_j(\varphi_{\text{query}}(f(\tilde{i})), \varphi_{\text{key}}(\mathcal{L}_y[f](i) + \rho(i, j))) \varphi_{\text{val}}(f(\tilde{j})) \right)
$$

Here, we have used the substitution $\tilde{i} = x^{-1}(x(i) - y) \Rightarrow i = x^{-1}(\tilde{i} + y)$ and $\tilde{j} = x^{-1}(x(j) - y) \Rightarrow j = x^{-1}(\tilde{j} + y)$. By using the definition of $\rho(i, j)$ we can further reduce the expression above as:

$$
m_r[\mathcal{L}_y[f], \rho](i) = \varphi_{\text{out}} \left( \sum_{h \in \mathbb{H}} \sum_{j \in \mathbb{H}(i)} \sigma_j(\varphi_{\text{query}}(f(\tilde{i})), \varphi_{\text{key}}(f(\tilde{j})) + \rho(\tilde{i}, j))) \varphi_{\text{val}}(f(\tilde{j})) \right)
$$

Since the area of summation remains equal to any translation $y \in \mathbb{R}^d$, we have that:

$$
\sum [\cdot]_{x^{-1}(\tilde{j})+y \in \mathbb{H}(x^{-1}(\tilde{j})+y)} = \sum [\cdot]_{x^{-1}(\tilde{i})+y \in \mathbb{H}(x^{-1}(\tilde{i}))} = \sum [\cdot].
$$

Resultantly, we can further reduce the expression above as:

$$
m_r[\mathcal{L}_y[f], \rho](i) = \varphi_{\text{out}} \left( \sum_{h \in \mathbb{H}} \sum_{j \in \mathbb{H}(i)} \sigma_j(\varphi_{\text{query}}(f(\tilde{i})), \varphi_{\text{key}}(f(\tilde{j}) + \rho(\tilde{i}, j))) \varphi_{\text{val}}(f(\tilde{j})) \right)
$$

$$
= m_r[f, \rho](\tilde{i}) = m_r[f, \rho](x^{-1}(x(i) - y))
$$

$$
= \mathcal{L}_y[m_r[f, \rho]](i)
$$

We see that indeed $m_r[\mathcal{L}_y[f], \rho](i) = \mathcal{L}_y[m_r[f, \rho]](i)$. Consequently, we conclude that the relative positional encoding allows for translation equivariance. We emphasize that this is a consequence of the fact that $\rho(x^{-1}(x(i) - y), x^{-1}(x(j) + y)) = \rho(i, j), \forall y \in \mathbb{R}^d$. In other words, it comes from the fact that relative positional encoding is invariant to the action of the translation group. \(\square\)
Proof of Claim 5.1. If the lifting self-attention formulation provided in Eq. 11 is $G$-equivariant, then it must hold that $m_{g}^{r}(L_{y}^{L}[f], \rho)(i, \tilde{h}) = L_{y}[m_{g}^{r}(f, \rho)](i, \tilde{h})$. Consider a $g$-transformed input signal $L_{g}[f](i) = L_{y}L_{\tilde{h}}[f](i) = f(x^{-1}(\tilde{h}^{-1}(x(i) - y)))$, $g = (y, \tilde{h})$, $y \in \mathbb{R}^d$, $\tilde{h} \in \mathcal{H}$. The lifting group self-attention operation on $L_{g}[f]$ is given by:

\[
m^{r}_{g}(L_{y}L_{\tilde{h}}[f], \rho)(i, \tilde{h}) = \varphi_{\text{out}}\left( \bigcup_{\tilde{h} \in \mathcal{H}} \sum_{j \in \Omega(i)} \sigma_{f}(\varphi_{\text{qy}}^{(h)}(L_{y}L_{\tilde{h}}[f](i)), \varphi_{\text{key}}^{(h)}(L_{y}L_{\tilde{h}}[f](i)) + L_{\tilde{h}}[\rho](i, j))) \varphi_{\text{val}}^{(h)}(L_{y}L_{\tilde{h}}[f](j)) \right)
\]

Here we have used the substitution $\tilde{i} = x^{-1}(\tilde{h}^{-1}(x(i) - y)) \Rightarrow i = x^{-1}(\tilde{h}x(\tilde{i}) + y)$ and $\tilde{j} = x^{-1}(\tilde{h}^{-1}(x(j) - y)) \Rightarrow j = x^{-1}(\tilde{h}x(\tilde{j}) + y)$. By using the definition of $\rho(i, j)$ we can further reduce the expression above as:

\[
= \varphi_{\text{out}}\left( \bigcup_{\tilde{h} \in \mathcal{H}} \sum_{j \in \Omega(i)} \sigma_{f}(\varphi_{\text{qy}}^{(h)}(f(\tilde{i})), \varphi_{\text{key}}^{(h)}(f(\tilde{j})) + L_{\tilde{h}}[\rho](x^{-1}(\tilde{h}x(\tilde{i}) + y), x^{-1}(\tilde{h}x(\tilde{j}) + y))) \varphi_{\text{val}}^{(h)}(f(\tilde{j})) \right)
\]

Since, for unimodular groups, the area of summation remains equal for any $g \in G$, we have that:

\[
\sum_{j \in \Omega(i)} [\cdot] = \sum_{j \in \Omega(i)} [\cdot] = \sum_{j \in \Omega(i)} [\cdot] = \sum_{j \in \Omega(i)} [\cdot].
\]

Resultantly, we can further reduce the expression above as:

\[
m_{G}^{r}[L_{y}L_{\tilde{h}}[f], \rho](i, \tilde{h}) = \varphi_{\text{out}}\left( \bigcup_{\tilde{h} \in \mathcal{H}} \sum_{j \in \Omega(i)} \sigma_{f}(\varphi_{\text{qy}}^{(h)}(f(\tilde{i})), \varphi_{\text{key}}^{(h)}(f(\tilde{j})) + L_{\tilde{h}}[\rho](\tilde{i}, \tilde{j}))) \varphi_{\text{val}}^{(h)}(f(\tilde{j})) \right)
= m_{G}^{r}[f, \rho](i, \tilde{h}) = m_{G}^{r}[f, \rho](x^{-1}(\tilde{h}^{-1}(x(i) - y)), \tilde{h}^{-1} \tilde{h}) = L_{y}L_{\tilde{h}}[m_{G}^{r}[f, \rho]](i, \tilde{h}).
\]

We see indeed that $m_{G}^{r}[L_{y}L_{\tilde{h}}[f], \rho](i, \tilde{h}) = L_{y}L_{\tilde{h}}[m_{G}^{r}[f, \rho]](i, \tilde{h})$. Consequently, we conclude that the lifting group self-attention operation is group equivariant. We emphasize once more that this is a consequence of the fact that $L_{\rho}(\rho(i, j)) \Rightarrow i = \rho(i, j)$, $\forall g \in G$. In other words, it comes from the fact that the positional encoding used is invariant to the action of elements $g \in G$.

\[
\]

Proof of Claim 5.2. If the group self-attention formulation provided in Eq. 11 is $G$-equivariant, then it must hold that $m_{G}^{r}[L_{g}[f], \rho](i, \tilde{h}) = L_{y}[m_{G}^{r}[f, \rho]](i, \tilde{h})$. Consider a $g$-transformed input signal $L_{g}[f](i, \tilde{h}) = L_{y}L_{\tilde{h}}[f](i, \tilde{h}) = f(x^{-1}(\tilde{h}^{-1}(\rho(i) - y))$, $g = (y, \tilde{h})$, $y \in \mathbb{R}^d$, $\tilde{h} \in \mathcal{H}$. The group self-attention operation on $L_{g}[f]$ is given by:

\[
m_{g}^{r}(L_{y}L_{\tilde{h}}[f], \rho)(i, \tilde{h}) = \varphi_{\text{out}}\left( \bigcup_{\tilde{h} \in \mathcal{H}} \sum_{j \in \Omega(i)} \sigma_{f}(\varphi_{\text{qy}}^{(h)}(L_{y}L_{\tilde{h}}[f](i, \tilde{h})), \varphi_{\text{key}}^{(h)}(L_{y}L_{\tilde{h}}[f](j, \tilde{h}))) \varphi_{\text{val}}^{(h)}(L_{y}L_{\tilde{h}}[f](j, \tilde{h})) \right)
= \varphi_{\text{out}}\left( \bigcup_{\tilde{h} \in \mathcal{H}} \sum_{j \in \Omega(i)} \sigma_{f}(\varphi_{\text{qy}}^{(h)}(f(x^{-1}(\tilde{h}^{-1}(x(i) - y)), \tilde{h}^{-1} \tilde{h})), \varphi_{\text{key}}^{(h)}(f(x^{-1}(\tilde{h}^{-1}(x(j) - y)), \tilde{h}^{-1} \tilde{h})) \varphi_{\text{val}}^{(h)}(f(x^{-1}(\tilde{h}^{-1}(x(j) - y)), \tilde{h}^{-1} \tilde{h})) \right)
\]
We see that indeed
\[
\sum \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{\tilde{h} \in \mathbb{H}} \sigma_{x^{-1}(\tilde{h},x(i)+y),\tilde{h}} \left( \left( \varphi_{\text{query}}^{(h)}(f(\tilde{i},\tilde{h}')), \varphi_{\text{key}}^{(h)}(f(\tilde{j},\tilde{h}')) + \mathcal{L}_{\tilde{h}}[\rho]\left( (x^{-1}(\tilde{h},x(i)+y),\tilde{h}) \right) \right) \left( x^{-1}(\tilde{h},x(j)+y),\tilde{h} \right) \right) \right) \right) \right).
\]

Here we have used the substitutions \( \tilde{i} = x^{-1}(\tilde{h}^{-1}(x(i) - y)) \Rightarrow i = x^{-1}(\tilde{h}(x(i) + y)), \tilde{h}' = \tilde{h}^{-1}\tilde{h} \). By using the definition of \( \rho((i,\tilde{h}),(j,\tilde{h})) \) we can further reduce the expression above as:
\[
= \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{\tilde{h} \in \mathbb{H}} \left( \sum_{x^{-1}(\tilde{h},x(i)+y),\tilde{h}} \left( \left( \varphi_{\text{query}}^{(h)}(f(\tilde{i},\tilde{h}')), \varphi_{\text{key}}^{(h)}(f(\tilde{j},\tilde{h}')) + \mathcal{L}_{\tilde{h}}[\rho]\left( (x^{-1}(\tilde{h},x(i)+y),\tilde{h}) \right) \right) \left( x^{-1}(\tilde{h},x(j)+y),\tilde{h} \right) \right) \right) \right).
\]

Furthermore, since for unimodular groups the area of summation remains equal for any transformation \( g \in G \), we have that:
\[
\sum \left[ \cdot \right]_{(x^{-1}(\tilde{h},x(i)+y),\tilde{h})} = \sum \left[ \cdot \right]_{(x^{-1}(\tilde{h},x(i)+y),\tilde{h})} = \sum \left[ \cdot \right]_{(x^{-1}(\tilde{h},x(i)+y),\tilde{h})} = \sum \left[ \cdot \right]_{(x^{-1}(\tilde{h},x(i)+y),\tilde{h})}.
\]

Additionally, we have that \( \sum_{\tilde{h} \in \mathbb{H}} [\cdot] = \sum_{\tilde{h} \in \mathbb{H}} [\cdot] \). Resultantly, we can further reduce the expression above as:
\[
\mathcal{L}_{g}[\mathcal{L}_{\tilde{h}}[L_y \mathcal{L}_{\tilde{h}}[f]],\rho](i,\tilde{h}) = \varphi_{\text{out}} \left( \bigcup_{h \in [H]} \sum_{\tilde{h} \in \mathbb{H}} \left( \sum_{x^{-1}(\tilde{h},x(i)+y),\tilde{h}} \left( \left( \varphi_{\text{query}}^{(h)}(f(\tilde{i},\tilde{h}')), \varphi_{\text{key}}^{(h)}(f(\tilde{j},\tilde{h}')) + \mathcal{L}_{\tilde{h}}[\rho]\left( (x^{-1}(\tilde{h},x(i)+y),\tilde{h}) \right) \right) \left( x^{-1}(\tilde{h},x(j)+y),\tilde{h} \right) \right) \right) \right).
\]

We see that indeed \( \mathcal{L}_{g}[\mathcal{L}_{y} \mathcal{L}_{\tilde{h}}[f],\rho](i,\tilde{h}) = \mathcal{L}_{y} \mathcal{L}_{\tilde{h}}[\mathcal{L}_{g}[f,\rho]](i,\tilde{h}) \). Consequently, we conclude that the group self-attention operation is group equivariant. We emphasize once more that this is a consequence of the fact that \( \mathcal{L}_{g}[\rho](i,\tilde{h}),(j,\tilde{h}) = \rho((i,\tilde{h}),(j,\tilde{h})), \forall g \in G \). In other words, it comes from the fact that the positional encoding used is invariant to the action of elements \( g \in G \).