Supplementary Material
Attentive Group Equivariant Convolutional Networks

A. Generalized Visual Self-Attention

Before we derive the constraints for general visual self-attention and prove Thm. 1 of the main article, we first motivate our definition of group equivariant visual self-attention. In the subsequent subsections we explain that our definition of attentive group convolution, as given in Eq. 14 of the main article, and reformulated in Eq. 25, essentially describes a group equivariant linear mapping that is augmented with an additional attention function.

A.1. Self-attention: From Vectors to Feature Maps

Let us first consider the general form of a linear map between respectively vector spaces (used in multi-layer perceptrons) and feature maps (used in (group) convolutional neural nets), defined as follows:

$$\text{vectors:} \qquad \mathbf{x}_c^{out} = \sum_{\tilde{c}}^{N_{\tilde{c}}} \mathbf{W}_{c,\tilde{c}}\, \mathbf{x}_{\tilde{c}}^{in}, \tag{21}$$

$$\text{feat maps:} \quad f_c^{out}(g) = \sum_{\tilde{c}}^{N_{\tilde{c}}} \int_G \Psi_{c,\tilde{c}}(g,\tilde{g}) f_{\tilde{c}}^{in}(\tilde{g}) \mathrm{d}\tilde{g} \tag{22}$$

Here, the first equation describes a linear map between vectors $\mathbf{x}^{in} \in \mathbb{R}^{N_{\tilde{c}}}$ and $\mathbf{x}^{out} \in \mathbb{R}^{N_c}$ via matrix-vector multiplication with matrix $\mathbf{W} \in \mathbb{R}^{N_c \times N_{\tilde{c}}}$. The second equation describes a linear map between feature maps $f^{in} \in (\mathbb{L}_2(G))^{N_{\tilde{c}}}$ and $f^{out} \in (\mathbb{L}_2(G))^{N_c}$, via a two argument kernel $\Psi \in \mathbb{L}_1(G \times G)^{N_{\tilde{c}} \times N_c}$. The two argument kernel Ψ can be seen as the continuous counterpart of the matrix $\mathbf{W}$, and matrix-vector multiplication (sum over input indices) is augmented with an integral over the input coordinates $\tilde{g}$.

Keeping this form of linear mapping, we define the self-attentive map as the regular linear map augmented with attention weights computed from the input. Consequently, we formally define the self-attentive mappings as:

$$\text{vectors:} \qquad \mathbf{x}_c^{out} = \sum_{\tilde{c}}^{N_{\tilde{c}}} \mathbf{A}_{c,\tilde{c}} \mathbf{W}_{c,\tilde{c}}\, \mathbf{x}_{\tilde{c}}^{in}, \tag{23}$$

$$\text{feat maps:} \quad f_c^{out}(g) = \sum_{\tilde{c}}^{N_{\tilde{c}}} \int_G \alpha_{c,\tilde{c}}(g,\tilde{g}) \Psi_{c,\tilde{c}}(g,\tilde{g}) f_{\tilde{c}}^{in}(\tilde{g}) \mathrm{d}\tilde{g} \tag{24}$$

in which the attention weights are computed from the input via some operator $\mathcal{A}$, i.e., $\mathbf{A}_{c,\tilde{c}} = \mathcal{A}[\mathbf{x}^{in}]_{c,\tilde{c}}$ in the vector case and $\alpha_{c,\tilde{c}} = \mathcal{A}[f^{in}]_{c,\tilde{c}}$ in the case of feature maps.

A.2. Equivariant Linear Maps are Group Convolutions

Now, since we want to preserve the spatial correspondences between the input and output feature maps, special attention should be paid to the continuous self-attentive mappings. In other words, these operators should be equivariant. By including an equivariance constraint on the linear mapping of Eq. 22 we obtain a group convolution (see e.g. Kondor & Trivedi (2018); Cohen et al. (2019a); Bekkers (2020)). The derivation is as follows:

Imposing the equivariance constraint $\mathcal{L}_g[f^{in}] \underset{\text{Eq.22}}{\mapsto} \mathcal{L}_g[f^{out}]$ means that for all $\overline{g}, g \in G$ and all $f \in \mathbb{L}_2(G)^{N_c}$ we must guarantee that:

$$\begin{aligned}
&\mathcal{L}_g[f^{in}] = \mathcal{L}_g[f^{out}] \\
&\Leftrightarrow \\
&\int_G \Psi_{c,\tilde{c}}(g,\tilde{g}) \mathcal{L}_{\overline{g}}[f](\tilde{g}) \mathrm{d}\tilde{g} = \int_G \Psi_{c,\tilde{c}}(\overline{g}^{-1}g,\tilde{g}) f(\tilde{g}) \mathrm{d}\tilde{g} \\
&\Leftrightarrow \\
&\int_G \Psi_{c,\tilde{c}}(g,\tilde{g}) f(\overline{g}^{-1}\tilde{g}) \mathrm{d}\tilde{g} = \int_G \Psi_{c,\tilde{c}}(\overline{g}^{-1}g,\tilde{g}) f(\tilde{g}) \mathrm{d}\tilde{g} \\
&\Leftrightarrow \\
&\int_G \Psi_{c,\tilde{c}}(g,\tilde{g}) f(\overline{g}^{-1}\tilde{g}) \mathrm{d}\tilde{g} = \\
&\qquad \int_G \Psi_{c,\tilde{c}}(\overline{g}^{-1}g, \overline{g}^{-1}\tilde{g}) f(\overline{g}^{-1}\tilde{g}) \mathrm{d}\tilde{g},
\end{aligned}$$

where the change of variables $\tilde{g} \rightarrow \overline{g}^{-1}\tilde{g}$ as well as the left-invariance of the Haar measure ($\mathrm{d}(\overline{g}^{-1}\tilde{g}) = \mathrm{d}\tilde{g}$) is used in the last step. Since this equality must hold for all $f \in \mathbb{L}_2(G)^{N_{\tilde{c}}}$ we obtain that Ψ should be left-invariant in both input arguments. In other words, we have that

$$\forall_{\overline{g} . \in G} : \Psi(\overline{g}g, \overline{g}\tilde{g}) = \Psi(g,\tilde{g})$$

Resultantly, we can always multiply both arguments with g^{-1} and obtain $\Psi(e, g^{-1}\tilde{g})$, which is effectively a single argument function $\psi(g^{-1}\tilde{g}) := \Psi(e, g^{-1}\tilde{g})$ that takes as input a relative "displacement" $g^{-1}\tilde{g}$. Consequently, under the equivariance constraint, Eq. 22 becomes a group convolution:

$$f_c^{out}(g) = \sum_{\tilde{c}}^{N_{\tilde{c}}} \int_G \psi_{c,\tilde{c}}(g^{-1}\tilde{g}) f_{\tilde{c}}^{in}(\tilde{g}) \mathrm{d}\tilde{g}.$$

A.3. Proof of Theorem 1

We can apply the same type of derivation to reduce the general form of visual self-attention of Eq. 24 to our main

definition of attentive group convolution:

$$f_c^{out}(g) = \sum_{\tilde{c}}^{N_{\tilde{c}}} \int_G \alpha_{c,\tilde{c}}(g,\tilde{g}) \psi_{c,\tilde{c}}(g^{-1}\tilde{g}) f_{\tilde{c}}^{in}(\tilde{g}) \mathrm{d}\tilde{g}. \quad (25)$$

However, we cannot reduce attention map α to a single argument function like we did for the kernel Ψ since α depends on the input f^{in}. To see this consider the following:

Without loss of generality, let $\mathfrak{A} : \mathbb{L}_2(G) \to \mathbb{L}_2(G)$ denote the attentive group convolution defined by Eq. 25, with $N_c = N_{\tilde{c}} = 1$, and some ψ which in the following we omit in order to simplify our derivation. Equivariance of $\mathfrak{A}$ implies that $\forall_{f \in \mathbb{L}_2(G)}, \forall_{\overline{g}, g \in G}$:

$$\begin{gathered}
\mathfrak{A}\left[\mathcal{L}_{\overline{g}}[f]\right](g) = \mathcal{L}_{\overline{g}}\left[\mathfrak{A}[f]\right](g) \\
\Leftrightarrow \\
\mathfrak{A}\left[\mathcal{L}_{\overline{g}}[f]\right](g) = \mathfrak{A}[f](\overline{g}^{-1}g) \\
\Leftrightarrow \\
\int_G \mathcal{A}\left[\mathcal{L}_{\overline{g}}[f]\right](g,\tilde{g}) \mathcal{L}_{\overline{g}}[f](\tilde{g}) \mathrm{d}\tilde{g} = \\
\int_G \mathcal{A}[f](\overline{g}^{-1}g, \tilde{g}) f(\tilde{g}) \mathrm{d}\tilde{g} \\
\Leftrightarrow \\
\int_G \mathcal{A}\left[\mathcal{L}_{\overline{g}}[f]\right](g,\tilde{g}) f(\overline{g}^{-1}\tilde{g}) \mathrm{d}\tilde{g} = \\
\int_G \mathcal{A}[f](\overline{g}^{-1}g, \overline{g}^{-1}\tilde{g}) f(\overline{g}^{-1}\tilde{g}) \mathrm{d}\tilde{g},
\end{gathered}$$

where we once again perform the variable substitution $\tilde{g} \to \overline{g}^{-1}\tilde{g}$ at the right hand side of the last step. This must hold for all $f \in \mathbb{L}_2(G)$ and hence:

$$\forall_{\overline{g} \in G} : \mathcal{A}\left[\mathcal{L}_{\overline{g}} f\right](g,\tilde{g}) = \mathcal{A}[f](\overline{g}^{-1}g, \overline{g}^{-1}\tilde{g}), \quad (26)$$

which proves the constraint on $\mathcal{A}$ as given in Thm. 1 of the main article. Just as for convolutions in Sec. A.2, we can turn this into a single argument function as:

$$\mathcal{A}[f](g,\tilde{g}) = \mathcal{A}\left[\mathcal{L}_{g^{-1}} f\right](e, g^{-1}\tilde{g}) =: \mathcal{A}'[\mathcal{L}_{g^{-1}} f](g^{-1}\tilde{g}), \quad (27)$$

in which $\mathcal{A}'$ is an attention operator that generates a single argument attention map from an input f. However, this would mean that for each g the input should be transformed via $\mathcal{L}_{g^{-1}}$, which does not make things easier for us. Things do get easier when we choose to attend to either the input *or* the output, which we discuss next.

Corollary 1. *Each attention operator $\mathcal{A}$ that generates an attention map $\alpha : G \times G \to [0,1]$ which is left-invariant to either one of the arguments, and thus exclusively attends either the input or output domain, satisfies the equivariance constraint of Eq. 26, iff the operator is G-equivariant, i.e., a group convolution.*

Proof. Left-invariant to either one of the arguments (let us now consider invariance in the first argument) means that:

$$\forall_{g,\tilde{g}} : \quad \mathcal{A}[f](g,\tilde{g}) = \mathcal{A}[f](e,\tilde{g}),$$

and hence, we are effectively dealing with a single argument attention map, which we define as $\mathcal{A}'[f](\tilde{g}) := \mathcal{A}(e,\tilde{g})$. Consequently, the equivariance constraint of Eq. 26 becomes:

$$\begin{gathered}
\forall_{\overline{g} \in G} : \mathcal{A}\left[\mathcal{L}_{\overline{g}} f\right](g,\tilde{g}) = \mathcal{A}[f](\overline{g}^{-1}g, \overline{g}^{-1}\tilde{g}) \Leftrightarrow \\
\forall_{\overline{g} \in G} : \mathcal{A}'\left[\mathcal{L}_{\overline{g}} f\right](\tilde{g}) = \mathcal{A}'[f](\overline{g}^{-1}\tilde{g}) \Leftrightarrow \\
\forall_{\overline{g} \in G} : \mathcal{A}'\left[\mathcal{L}_{\overline{g}} f\right] = \mathcal{L}_{\overline{g}}\left[\mathcal{A}'\right][f].
\end{gathered}$$

Conclusively, $\mathcal{A}'$ must be an equivariant operator. □

The derivation of the Eq. 26 together with the proof of Corollary 1 completes the proof of Theorem 1 of the main article.

A.4. Equivariance Proof of the Proposed Visual Attention

In this section we revisit the proposed attention mechanisms and prove that they indeed satisfy Thm. 1 of the main article. Recall the general formulation of attentive group convolution given in Eq. 25. Inspired by the work of Woo et al. (2018), we reduce the computation load by factorizing the attention map α into channel and spatial components via:

$$\alpha_{c,\tilde{c}}(g,\tilde{g}) = \alpha^{\mathcal{X}}(x,h,\tilde{h}) \alpha^{C}_{c,\tilde{c}}(h,\tilde{h})$$

where $\alpha^{\mathcal{C}}$ attends to both input and output channels as well as input and output poses $h, \tilde{h} \in H$, and spatial attention attends to the output domain $g = (x,h) \in G$ for all input poses $\tilde{h} \in H$ but does not change for input spatial positions $\tilde{x} \in \mathbb{R}^d$. We denote the operators $\mathcal{A}^C$, $\mathcal{A}^X$ utilized to compute the attention maps as $\alpha^{\mathcal{C}} = \mathcal{A}^{\mathcal{C}}[f]$ and $\alpha^{\mathcal{X}} = \mathcal{A}^{\mathcal{X}}[f]$, respectively.

A.4.1. Channel attention

We compute channel attention via:

$$\begin{aligned}
\mathcal{A}^{\mathcal{C}}[f](h,\tilde{h}) &= \varphi^{\mathcal{C}}\left[s^{\mathcal{C}}\left[\tilde{f}[f]\right]\right](h,\tilde{h}) \quad (28) \\
&= \sigma\Big(\left[\mathbf{W}_2(h^{-1}\tilde{h}) \cdot \left[\mathbf{W}_1(h^{-1}\tilde{h}) \cdot s^{\mathcal{C}}_{\text{avg}}(h,\tilde{h})\right]^+\right] \\
&\quad + \left[\mathbf{W}_2(h^{-1}\tilde{h}) \cdot \left[\mathbf{W}_1(h^{-1}\tilde{h}) \cdot s^{\mathcal{C}}_{\text{max}}(h,\tilde{h})\right]^+\right]\Big)
\end{aligned}$$

with

$$\tilde{f}_{c,\tilde{c}}(x,h,\tilde{h}) := \left[f_{\tilde{c}} \star_{\mathbb{R}^d} \mathcal{L}_h[\psi_{c,\tilde{c}}]\right](x,\tilde{h}) \quad (29)$$

the intermediary result from the convolution between the input f and the h-transformation of the filter ψ, $\mathcal{L}_h[\psi]$ before pooling over $\tilde{c}$ and $\tilde{h}$. $s^{\mathcal{C}}_{\text{avg}}$ and $s^{\mathcal{C}}_{\text{max}}$ denote respectively average and max pooling over the x coordinate.

Here, we apply a slight abuse of notation with $\tilde{f}[f]$ and $s^{\mathcal{C}}[\tilde{f}]$ in order to keep track of the dependencies. In order to proof equivariance of the attention operator $\mathcal{A}^{\mathcal{C}}$ we need to proof that $\forall_{\overline{g}\in G} : \mathcal{A}^{\mathcal{C}}[\mathcal{L}_{\overline{g}}[f]](h,\tilde{h}) = \mathcal{A}^{\mathcal{C}}[f](\overline{h}^{-1}h, \overline{h}^{-1}\tilde{h})$, with $\overline{g} = (\overline{x}, \overline{h})$. To this end, we first identify the equivariance and invariance properties of the functions used in Eq. 28.

From Eq. 29 we see that the intermediate convolution result $\tilde{f}$ is equivariant via $\tilde{f}[\mathcal{L}_{\overline{g}}[f]](\tilde{x},\tilde{h}) = \tilde{f}[f](\overline{g}^{-1}x, \overline{h}^{-1}h, \overline{h}^{-1}\tilde{h})$. For the statistics operators $s^{\mathcal{C}}$ we have invariance w.r.t. translations due to the pooling over x, and equivariance w.r.t. parameter $\overline{h}$ via $s^{\mathcal{C}}[\tilde{f}[\mathcal{L}_{\overline{g}}f]](h,\tilde{h}) = s^{\mathcal{C}}[\tilde{f}[f]](\overline{h}^{-1}h, \overline{h}^{-1}\tilde{h})$. Now, we propagate the transformation on the input and compute the result of $\mathcal{A}^{\mathcal{C}}[\mathcal{L}_{\overline{g}}[f]](g,\tilde{g})$. That is, we compute the left-hand side of the constraint given in Eq. 26, where, for brevity, we omit the $s^{\mathcal{C}}_{\text{max}}$ term:

$$\mathcal{A}^{\mathcal{C}}[\mathcal{L}_{\overline{g}}[f]](g,\tilde{g}) = \mathbf{W}_2(h^{-1}\tilde{h}) \cdot [\mathbf{W}_1(h^{-1}\tilde{h}) \cdot s^{\mathcal{C}}_{\text{avg}}(\overline{h}^{-1}h, \overline{h}^{-1}\tilde{h})]^{+}.$$

The right-hand side of Eq. 26 is given by:

$$\mathcal{A}^{\mathcal{C}}[f](\overline{g}^{-1}g, \overline{g}^{-1}\tilde{g}) = \mathbf{W}_2(h^{-1}\tilde{h}) \cdot [\mathbf{W}_1(h^{-1}\tilde{h}) \cdot s^{\mathcal{C}}_{\text{avg}}(\overline{h}^{-1}h, \overline{h}^{-1}\tilde{h})]^{+}.$$

and hence, Eq. 26 is satisfied for all $\overline{g} \in G$. Resultantly, $\mathcal{A}^{\mathcal{C}}$ is a valid attention operator.

A.4.2. Spatial attention

The spatial attention map $\alpha^{\mathcal{X}}$ is computed via:

$$\begin{aligned}\alpha^{\mathcal{X}}(g,\tilde{h}) &= \mathcal{A}^{\mathcal{X}}[f](g,\tilde{h}) \\ &= \varphi^{\mathcal{X}}\left[s^{\mathcal{X}}\left[\tilde{f}[f]\right]\right](g,\tilde{h}) \\ &= \sigma\left([s^{\mathcal{X}} \star_{\mathbb{R}^d} \mathcal{L}_h[\psi^{\mathcal{X}}]\right)(x,\tilde{h}) \qquad (30)\end{aligned}$$

where σ is a point-wise logistic sigmoid, $\psi^{\mathcal{X}} : G \to \mathbb{R}^2$ is a group convolution filter and $s^{\mathcal{X}}[\tilde{f}] : G \times H \to \mathbb{R}^2$ is a map of averages and maximum values taken over the channel axis at each $g \in G$ in $\tilde{f}$ for each $\tilde{h} \in H$. Note that Eq. 30 corresponds to a group convolution up to the final pooling operation over $\tilde{h}$. Since the statistics operator $s^{\mathcal{X}}$ is invariant w.r.t. translations in the input and Eq. 30 corresponds to a group convolution (up to pooling over $\tilde{h}$), we have that $\mathcal{A}^{X}$ is a valid attention operator as well.

B. Extended Implementation Details

In this section we provide extended details over our implementation. For the sake of completeness and reproducibility, we summarize the exact training procedures utilized during our experiments. Moreover, we delve into some important changes performed to some network architectures during our experiments to ensure *exact* equivariance, and shed light into their importance for our equivariant attention maps.

B.1. General Observations

We utilize `PyTorch` for our implementation. Any missing parameter specification in the following sections can be safely considered to be the default value of the corresponding parameter. For batch normalization layers, we utilize eps=0.00002 similarly to Cohen & Welling (2016a).

B.2. rot-MNIST

For rotational MNIST, we utilized the same backbone network as in Cohen & Welling (2016a). During training we utilize Adam (Kingma & Ba, 2014), batches of size 128, weight decay of 0.0001, learning rate of 0.001, drop-out rate of 0.3 and perform training for 100 epochs. Importantly and contrarily to Cohen & Welling (2016a), we consistently experience improvements when utilizing drop-out and therefore we do not exclude it for any model.

B.3. CIFAR-10

It is not clear from Springenberg et al. (2014); Cohen & Welling (2016a) which batch size is used in their experiments. For our experiments, we always utilize batches of size 128.

B.3.1. All-CNN

We utilize the All-CNN-C structure of Springenberg et al. (2014). Analogously to Springenberg et al. (2014); Cohen & Welling (2016a), we utilize stochastic gradient descent, weight decay of 0.001 and perform training for 350 epochs. We utilize a grid search on the set $\{0.01, 0.05, 0.1, 0.25\}$ for the learning rate and report the best obtained performance. Furthermore, we reduce the learning rate by a factor of 10 at epochs 200, 250 and 300.

B.3.2. ResNet44

Similar to Cohen & Welling (2016a), we utilize stochastic gradient descent, learning rate of 0.05 and perform training for 300 epochs. Furthermore, we reduce the learning rate by a factor of 10 at epochs 50, 100 and 150.

B.4. PCam

During training on the PatchCamelyon dataset, we utilize Adam (Kingma & Ba, 2014), batches of size 64, weight decay of 0.0001, learning rate of 0.001 and perform training for 100 epochs. Furthermore, we reduce the learning rate by a factor of 2 after 20 epochs of no improvement in the validation loss.

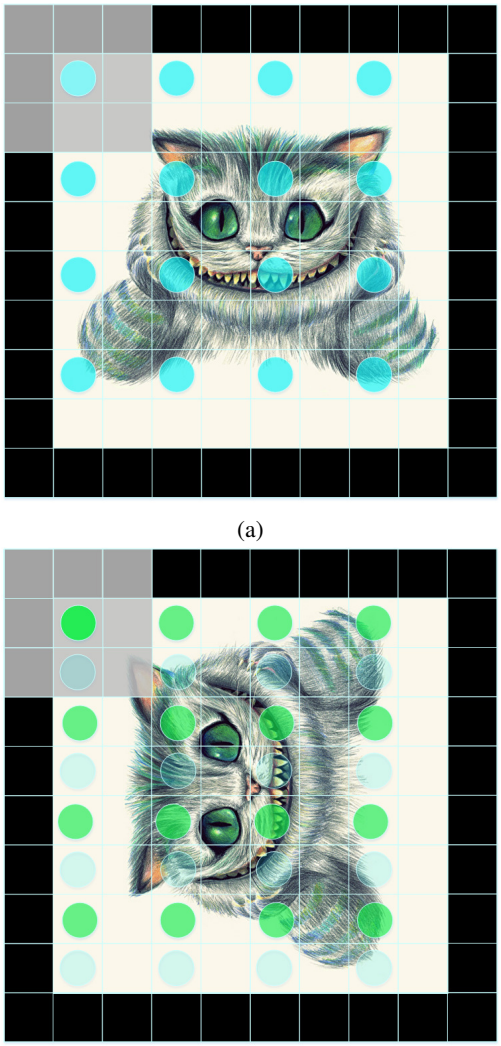
C. Effects of Stride and Input Size on Equivariance

Theoretically seen, the usage of stride during pooling and during convolution is of no relevance for the equivariance properties of the corresponding mapping (Cohen & Welling, 2016a). However, we see that in practice stride can affect equivariance for specific cases as is the case for our experiments on CIFAR-10.

Consider the convolution between an input of even size and a small 3x3 filter as shown in Fig. 9a. Via group convolutions, we can ensure that the output of the original input and a rotated one (Fig. 9b) will be exactly equal (up to the same rotation). Importantly however, note that for Fig. 9, the local support of the filter, i.e., the input section with which the filter is convolved at a particular position, *is not equivalent* for rotated versions of the input (denoted by blue circles for the non-rotated case and by green circles by for t he rotated case). As a result, despite the group convolution itself being equivariant, the responses of both convolutions do not entirely resemble one another and, consequently, the depicted strided group convolution is *not* exactly equivariant.

It is important to highlight that this behaviour is just exhibited for the special case when the residual between the used stride and the input size is even. Unfortunately, this is the case both for the ResNet44 as well as the All-CNN networks utilized in our CIFAR-10 experiments. However, as neighbouring pixels are extremely correlated with one another, the effects of this phenomenon are not of much relevance for the classification task itself. As a matter of fact, it can be interpreted as a form of data augmentation by skipping intermediary pixel values. Consequently, we can say that these networks are *approximately equivariant*.

Importantly, this phenomenon does affect the resulting equivariant attention maps generated via attentive group convolutions as shown in Fig. 10. As these networks are only equivariant in an approximate manner, the generated attention maps are slightly deformed versions of one another for multiple orientations. In order to alleviate this problem, we replace all strided convolutions in the All-CNN and ResNet44 architectures by conventional convolutions (stride=1), followed by spatial max pooling. Resultantly, we are able to produce exactly equivariant attention maps as shown in Fig. 7 in the main text and Fig. 11 here.



(a)

(b)

Figure 9. Effect of stride and input size on exact equivariance. Although group convolutions are ensured to be group equivariant, in practice, if the residual between the stride and the input size is even, as it's the case for the networks utilized in the CIFAR-10 experiments, equivariance is only approximate. This has important effects on equivariant attention maps (Fig. 10).

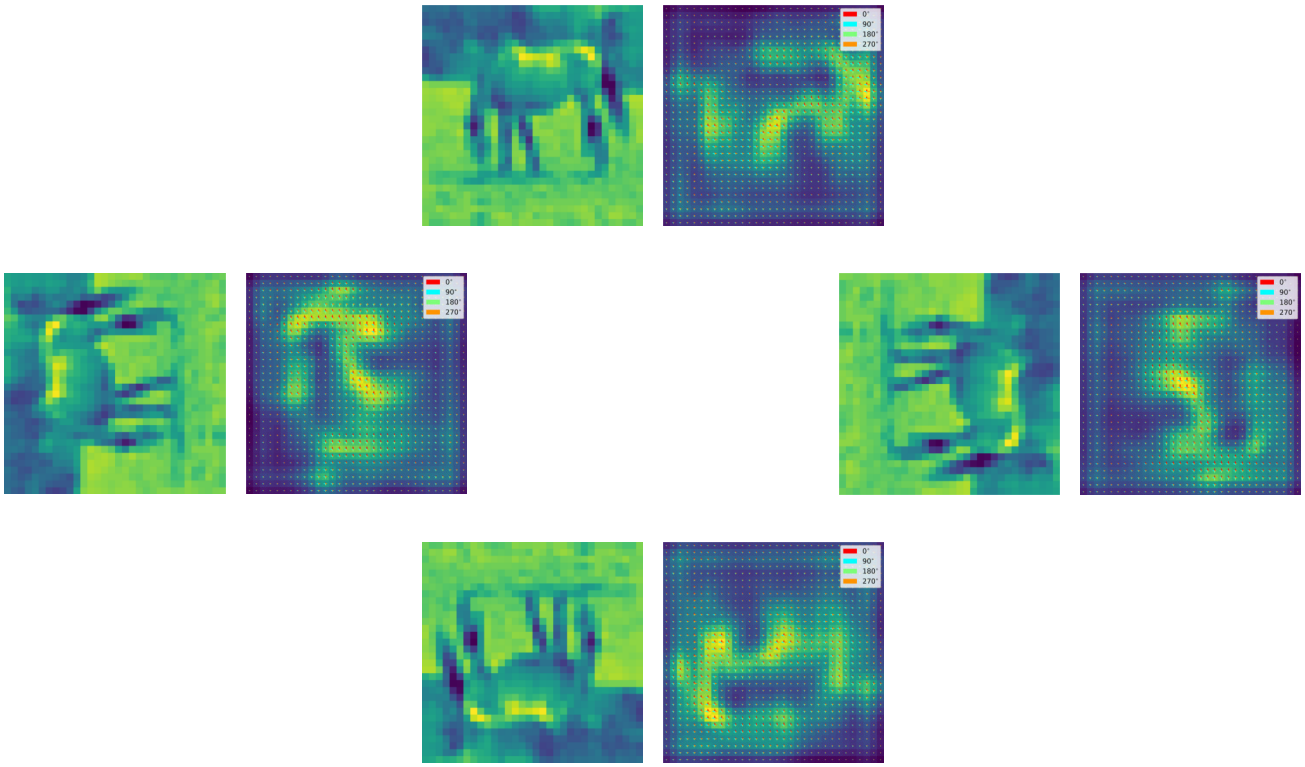


Figure 10. Examples of equivariant attention maps under the approximate equivariance regime. Note, for example, that attention around the horse's back changes for different orientations.

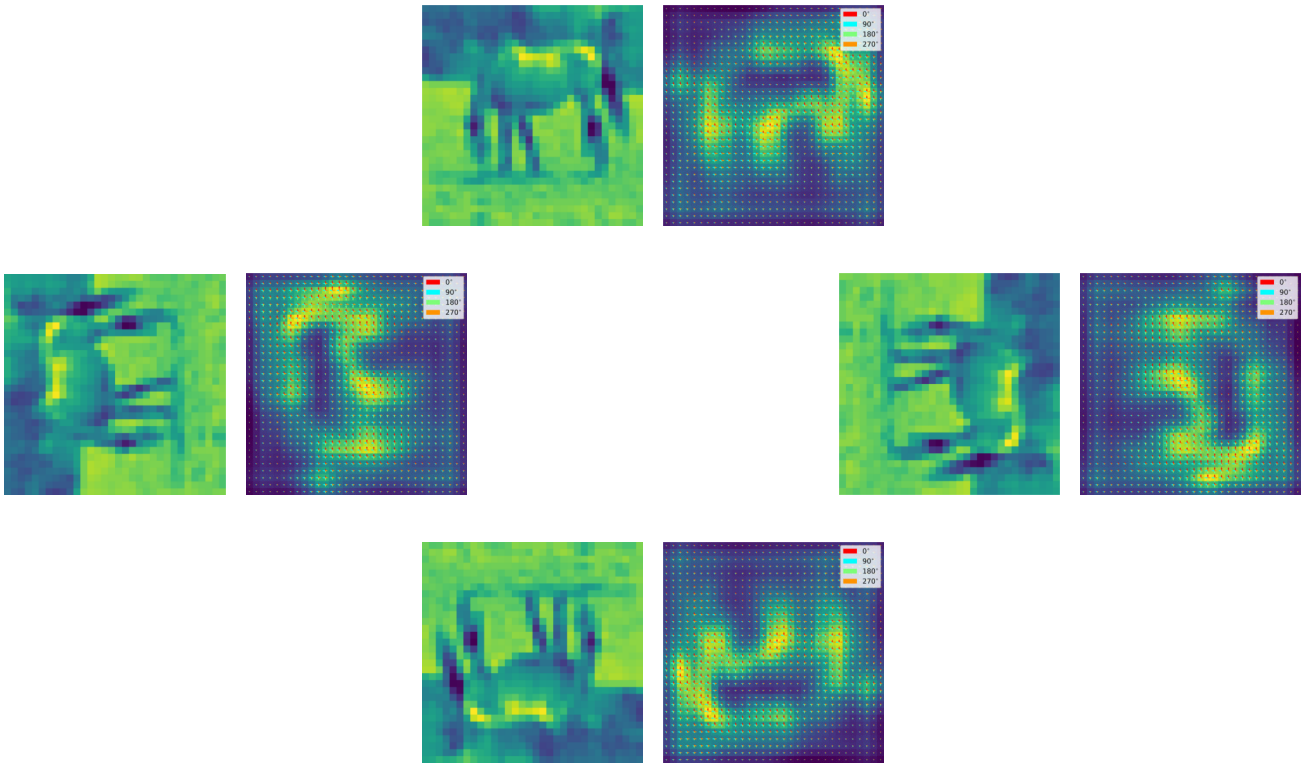


Figure 11. Examples of equivariant attention maps under the exact equivariance regime.