

EQUISELD: EFFICIENT TRAINING OF EQUIVARIANT SOUND EVENT LOCALIZATION AND DETECTION NETWORKS

Goksenin Yuksel¹, Marcel van Gerven¹, Kiki van der Heijden^{1,2}

¹Dept. of Machine Learning and Neural Computing, Donders Institute,
Radboud University, Nijmegen, the Netherlands

²Mortimer B. Zuckerman Institute, Columbia University, New York, USA

ABSTRACT

First-order Ambisonics (FOA) signals exhibit exact $O(3)$ symmetry: The rotation or reflection of the FOA signal modifies the direction of arrival of the sound sources, while preserving the sound sources themselves. Prior attempts to utilize this spatial symmetry of FOA to improve the efficiency and robustness of sound event detection and localization (SELD) systems either learned only an approximation of the symmetry through rotation-based augmentation or relied on computationally expensive methods to integrate equivariance. Furthermore, prior work focused exclusively on $SO(3)$ equivariance, leaving the potential of incorporating $O(3)$ equivariance for SELD tasks unclear. To address these limitations, we developed EQUISELD. This equivariant attention network processes first-order Ambisonics as paired streams of $O(3)$ -invariant scalars and equivariant intensity vectors, producing an invariant activity magnitude and an equivariant DOA with a Multi-ACCDOA readout. To compare the impact of $O(3)$ versus $SO(3)$ -equivariance, we designed a matched $SO(3)$ -only variant. EQUISELD outperforms prior equivariant networks on both simulated scenes with measured RIRs and recordings of real-world sound scenes at a fraction of the training cost. EQUISELD additionally surpasses the performance of non-equivariant SELD networks of a similar size on the simulated real-world sound scenes and achieves competitive performance on the real-world sound scenes.

Index Terms— Sound event localization and detection, Ambisonics, Group equivariant networks

1. INTRODUCTION

Sound Event Localization and Detection (SELD) provides a joint spatio-temporal characterization of acoustic scenes, supporting applications ranging from computational auditory scene analysis to immersive audio and smart environments [1, 2]. First-order Ambisonics (FOA) has become the de facto standard input for SELD systems [3, 4, 5]. The mathematical definition of the FOA signal ensures that the sound scene’s directional components transform equivariantly under any rotation or reflection of the scene, while its omnidirectional component (W) remains invariant. This spatial equivariance is a natural candidate to improve the efficiency and robustness of SELD systems [6].

Yet, most SELD systems process FOA with standard CNNs or Transformers that do not encode spatial equivariance directly, instead relying on rotational data augmentation to inject the spatial symmetry statistically [7, 8, 3]. However, this approach consumes data and computational resources to learn a constraint that could be hard-coded and, even though such rotational transformations are physically exact in the Ambisonic domain, the trained network only approximates the desired symmetry.

Group-equivariant networks address this limitation by incorporating equivariance into the network directly. Sato et al. [6] employed Clebsch–Gordan product networks [9] to integrate $SO(3)$ rotation equivariance into a unified rotation, scale, and time-translation framework. However, their architecture does not take into account reflection equivariance, is computationally expensive, and cannot represent overlapping instances of the same sound class (common in real-world acoustic scenes, [4, 10]). Outside the audio domain, equivariant attention networks have gained traction for processing sets of geometric features, combining the expressivity of transformers with exact symmetry guarantees [11, 12] and building on the permutation-equivariant set attention of [13]. A complementary line of work achieves equivariance through typed scalar–vector streams, in which invariant scalars gate equivariant vector channels via their norms [14, 15].

Inspired by this direction, we propose, to the best of our knowledge, the first SELD network that is exactly equivariant to the full orthogonal group $O(3)$. EQUISELD¹ is an equivariant attention network, processing each time–mel token as a pair of typed streams. Each token comprises $O(3)$ -invariant scalars and equivariant first-order intensity vectors. By computing attention weights and all nonlinearities solely from invariant scalars, and mixing vector channels via bias-free linear maps, we achieve exact equivariance. Our network accommodates overlapping sound sources of the same type, outputting Multi-ACCDOA [16] activity as an invariant magnitude vector and DOA as an equivariant unit vector. We furthermore isolate $O(3)$ ’s contribution with a matched $SO(3)$ -only variant.

The results demonstrate that EQUISELD produces exactly equivariant predictions, surpasses the performance of an existing $SO(3)$ -equivariant network [6] with $2.9\times$ fewer parameters and $19\times$ less training time, and obtains a lower SELD score than SELD-Nets trained with and without exact $O(3)$ data augmentation.

2. METHOD

2.1. Invariant and Equivariant FOA Features

First-order Ambisonics (FOA) comprises an omnidirectional channel W and three directional channels $\mathbf{B} = [X, Y, Z]^\top$. Under rotations or reflections $Q \in O(3)$, W is unchanged and $\mathbf{B} \mapsto Q\mathbf{B}$.

For each time–frequency bin of an STFT (1024-sample window, 20 ms hop), we compute the complex intensity $\mathbf{i} = W^*\mathbf{B}$, its active and reactive components $\mathbf{i}_a = \Re\{\mathbf{i}\}$ and $\mathbf{i}_r = \Im\{\mathbf{i}\}$, and channel powers $E_W = |W|^2$ and $E_B = \|\mathbf{B}\|^2$.

Features are pooled to 64 mel bands, with $\langle \cdot \rangle_m$ denoting the weighted frequency sum under the m -th mel filter. The normalized

¹Code&Models: <https://github.com/labhamlet/equiseLD>

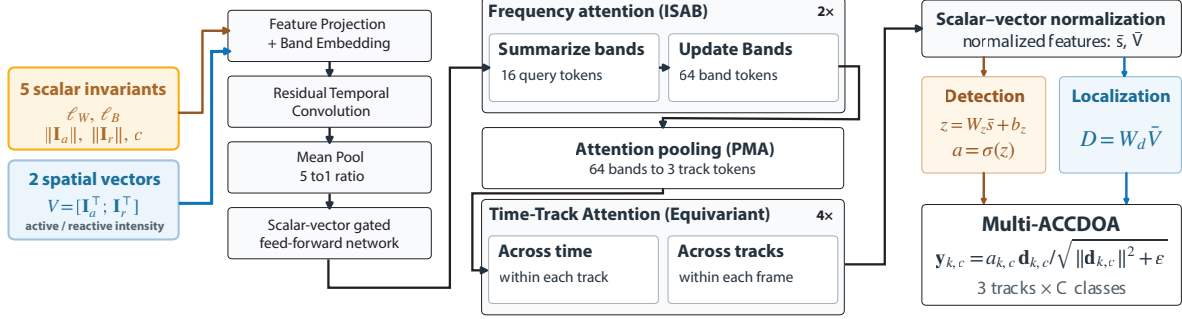


Fig. 1. EquiSELD architecture. Five invariant scalar features (orange) and two equivariant intensity vectors (blue) are processed with frequency attention, attention pooling, and attention across time and tracks. The network produces multi-ACCDOA outputs for C sound classes for K tracks of which direction estimates are equivariant. ISAB = induced set attention block; PMA = pooling by multihead attention [13].

intensities are

$$\mathbf{I}_{p,m} = \frac{\langle \mathbf{I}_p \rangle_m}{\bar{E}_m + \epsilon}, \quad p \in \{a, r\}, \quad (1)$$

where $\bar{E}_m = \langle \frac{1}{2}(E_W + \frac{1}{3}E_B) \rangle_m$ is the mean band energy and $\epsilon = 10^{-8}$ is used throughout for numerical stability.

Each time-mel bin contains two equivariant vectors, stacked as rows, and five invariant scalars:

$$\mathbf{V}_m = \begin{bmatrix} \mathbf{I}_{a,m}^\top \\ \mathbf{I}_{r,m}^\top \end{bmatrix}, \quad (2)$$

$$\mathbf{s}_m = [\ell_W, \ell_B, \|\mathbf{I}_{a,m}\|, \|\mathbf{I}_{r,m}\|, c_m]^\top.$$

Here $\ell_X = \log(\langle E_X \rangle_m + \epsilon)$, $X \in \{W, B\}$, are log band energies, and c_m is the cosine similarity between $\mathbf{I}_{a,m}$ and $\mathbf{I}_{r,m}$.

2.2. Equivariant Token Blocks

A token $(\mathbf{s}, \mathbf{V})$ pairs d_s scalar channels with d_v vector channels, stored as rows $\mathbf{v}_j^\top$ of $\mathbf{V}$. The group action is $Q \cdot (\mathbf{s}, \mathbf{V}) = (\mathbf{s}, \mathbf{V}Q^\top)$. A block F is equivariant if $F(Q \cdot \mathcal{T}) = Q \cdot F(\mathcal{T})$ for every $Q \in O(3)$ and every token collection $\mathcal{T}$.

Input tokens ($d_s = 5$, $d_v = 2$) are mapped to higher dimension by a scalar linear layer and a vector map (3). Throughout, $\mathbf{W}(\cdot)$ and $\mathbf{U}$ are learnable real matrices; $\sigma(\cdot)$, $\text{diag}(\cdot)$, $[\cdot; \cdot]$, and $\|\cdot\|_F$ denote the logistic sigmoid, a diagonal matrix, vertical concatenation, and the Frobenius norm.

Vector linear maps. Vector channels are mixed without bias:

$$\mathcal{L}_v(\mathbf{V}) = \mathbf{W}_v \mathbf{V}. \quad (3)$$

The same weights act on all three spatial coordinates, giving $\mathcal{L}_v(\mathbf{V}Q^\top) = \mathcal{L}_v(\mathbf{V})Q^\top$.

Channel norms. The per-channel norms

$$\nu(\mathbf{V}) = [\sqrt{\|\mathbf{v}_1\|^2 + \epsilon}, \dots, \sqrt{\|\mathbf{v}_{d_v}\|^2 + \epsilon}]^\top. \quad (4)$$

satisfy $\nu(\mathbf{V}Q^\top) = \nu(\mathbf{V})$. Within the token blocks, vector information enters the scalar stream only through these invariant norms.

Normalization. Scalars use LayerNorm [17], denoted LN. Vectors are normalized by their RMS channel norm [15] and rescaled by learnable channel gains γ :

$$\mathcal{N}_v(\mathbf{V}) = \frac{\text{diag}(\gamma)\mathbf{V}}{\sqrt{d_v^{-1}\|\mathbf{V}\|_F^2 + \epsilon}}. \quad (5)$$

Since $\|\mathbf{V}Q^\top\|_F = \|\mathbf{V}\|_F$, the denominator is invariant and $\mathcal{N}_v(\mathbf{V}Q^\top) = \mathcal{N}_v(\mathbf{V})Q^\top$.

Gated feed-forward block. An MLP produces scalar updates and one gate per vector channel:

$$\begin{aligned} \tilde{\mathbf{V}} &= \mathcal{L}_{v,1}(\mathcal{N}_v(\mathbf{V})), \\ \mathbf{h} &= \text{GELU}(\mathbf{W}_1[\text{LN}(\mathbf{s}); \nu(\tilde{\mathbf{V}})] + \mathbf{b}_1), \\ [\Delta \mathbf{s}; \mathbf{g}] &= \mathbf{W}_2 \mathbf{h} + \mathbf{b}_2, \\ \mathbf{s}' &= \mathbf{s} + \Delta \mathbf{s}, \\ \mathbf{V}' &= \mathbf{V} + \mathcal{L}_{v,2}(\text{diag}(\sigma(\mathbf{g}))\tilde{\mathbf{V}}). \end{aligned} \quad (6)$$

where $\mathcal{L}_{v,1}, \mathcal{L}_{v,2}$ follow (3), and $\mathbf{b}_1, \mathbf{b}_2$ are learnable scalar biases. Both $\Delta \mathbf{s}$ and $\mathbf{g}$ are invariant. Applying each gate uniformly to a vector's three coordinates therefore preserves equivariance, as do the vector maps and residual addition.

Equivariant attention. Following equivariant attention networks [12], we compute attention weights from invariant token descriptors:

$$\phi_i = \text{LN}(\mathbf{s}_i) + \mathbf{U} \nu(\mathcal{N}_v(\mathbf{V}_i)). \quad (7)$$

For one head with query/key width d_h , let $\mathbf{q}_i = \mathbf{W}_q \phi_i$ and $\mathbf{k}_j = \mathbf{W}_k \phi_j$. The attention weights are

$$A_{ij} = \text{softmax}_j \left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_h}} \right). \quad (8)$$

Here i indexes query tokens and j indexes key/value tokens. The same weights update both streams:

$$\begin{aligned} \mathbf{s}'_i &= \mathbf{s}_i + \mathbf{W}_o^s \sum_j A_{ij} \mathbf{W}_v^s \phi_j, \\ \mathbf{V}'_i &= \mathbf{V}_i + \mathcal{L}_o^v \left(\sum_j A_{ij} \mathcal{L}_v^v(\mathcal{N}_v(\mathbf{V}_j)) \right). \end{aligned} \quad (9)$$

The matrices $\mathbf{W}_v^s, \mathbf{W}_o^s$ are scalar value and output projections, and $\mathcal{L}_v^v, \mathcal{L}_o^v$ are their vector counterparts (3).

Because ϕ_i is invariant, queries, keys, and attention weights are unchanged under Q . For vector values $\mathbf{Z}_j = \mathcal{L}_v^v(\mathcal{N}_v(\mathbf{V}_j))$, aggregation satisfies the equivariance constraint. Channel concatenation, residual addition, and composition preserve this property, so all blocks above are equivariant to rotations and reflections.

2.3. SELD Architecture

EQUISELD adapts Set Transformer attention [13] to paired scalar-vector features, using $d_s = 128$, $d_v = 32$, and 2.25M parameters. Figure 1 specifies layer order, token counts, and block repetitions. Temporal vector convolutions are bias-free and share kernels across Cartesian components. We add learnable mel-band positional embeddings to the scalar stream, shared across time. Frequency attention and pooling follow the induced set attention block (ISAB) and pooling by multihead attention (PMA), respectively. ISAB/PMA seeds have learnable scalar components only, with seed vectors fixed at zero. Time-track attention operates across time within each track and across tracks within each frame. Output heads combine invariant activities with normalized equivariant directions into multi-ACCDOA outputs [16] for C classes and K tracks per class.

3. EXPERIMENTS

3.1. Setup

Datasets. We evaluate our method on two established SELD tasks: TAU2021 [10] and STARSS23 [4]. TAU2021 is a synthetically generated dataset based on real recorded room impulse responses (RIRs) featuring moving sound sources and interfering noise (sampling rate = 24 kHz). It comprises six folds, each containing approximately 100 minutes of audio. We use folds 1–4 for training, fold 5 for validation, and fold 6 for testing. STARSS23 is a recorded dataset of real-world listening scenes that presents a more challenging acoustic environment with up to six overlapping sound events (sampling rate = 24 kHz). All STARSS23 systems are trained on the real dev-train split only.

Training and evaluation. We adopted the Multi-ACCDOA output format with $K = 3$ tracks for each of C sound classes and trained models on 5 s segments using the ADPIT loss [16] and a detection threshold of 0.5. We trained models for 100 epochs on TAU2021 and for 200 epochs for STARSS23. We used a longer training regime for STARSS23 to avoid underfitting.

Evaluation metrics. Following the conventional evaluation of SELD systems, we report the macro-averaged, location-dependent error rate (ER_{20°) and F-score (F_{20°). Here, a prediction counts as a true positive only if the DOA falls within 20° of the target. We furthermore report the class-dependent localization error (LE_{CD}) in degrees and localization recall (LR_{CD}) [18]. We use the $\mathcal{E}_S$ to quantify overall SELD performance:

$$\mathcal{E}_S = \frac{1}{4} [ER_{20^\circ} + (1 - F_{20^\circ}) + LE_{CD}/180^\circ + (1 - LR_{CD})]. \quad (10)$$

3.2. Model space and training parameters

EquiSELD. We optimized EQUISELD with AdamW [19] ($\text{lr} = 3 \times 10^{-4}$, $\beta = (0.9, 0.95)$, and weight decay 0.05). We applied gradient clipping at 1.0 and used a linear warmup over 2000 steps followed by cosine decay down to 1% of the peak learning rate.

EquiSELD-SO(3). To quantify the potential benefit of $O(3)$ equivariance of EQUISELD in comparison with $SO(3)$ equivariance, we introduce an $SO(3)$ -equivariant version of the model by mixing a third, axial-vector channel — the cross product $\mathbf{v}_a \times \mathbf{v}_r$ — into the vector stream. As the product of two polar vectors, this feature transforms as $\mathbf{v} \mapsto \det(Q) Q\mathbf{v}$. This is identical to a polar vector under proper rotations, but sign-flipped under reflection. It therefore reduces the guaranteed symmetry from $O(3)$ to $SO(3)$. The variant is otherwise identical to EquiSELD in architecture, parameter count, and training recipe.

AttnSELD. To compare equivariant and non-equivariant parameterizations at matched model size, we construct ATTNSELD. This architecture is similar to EQUISELD but releases the constraints that guarantee equivariance through four modifications: (i) vector channel mixings gain learned 3×3 spatial mixing matrices and per-channel biases; (ii) attention logits and the feed-forward block read raw three-dimensional vector components rather than their norms; (iii) vector normalization permits per-component affine transformations; and (iv) inducing points and the pooling head carry learnable, non-zero vector parts. Since exposing raw components increases the input dimensions of the scalar projections, we adjust the vector width to $d_v = 12$ ($d_v = 32$ in EQUISELD) resulting in nearly 2.25 M parameters as in EQUISELD.

AttnSELD+. To assess to what extent the non-equivariant version of EQUISELD can learn equivariance through $O(3)$ augmentation, we introduce ATTNSELD+. Our augmentation implementation follows the channel-first [7]. We sample an element from $O(3)$ for each instance in the batch and apply on-the-fly augmentations. It is continuous over the $O(3)$ elements.

SELDNet. We adopt the SELDNet variant that served as the official DCASE 2023 Task 3 baseline [4]. The CRNN of [20], extended with two multi-head self-attention layers and a Multi-ACCDOA output [16]. This model first extracts a concatenated mel-spectrogram and normalized active intensity vector (IV) using a CNN front-end, then processes the resulting frame-wise features with a GRU followed by multi-head self-attention. We train two variants: the conventional recipe with 0.75 M parameters, and a larger parameter-matched version with 2.25 M parameters (obtained by increasing the GRU hidden dimension from 128 to 240). All SELDNet variants follow the published training recipe using the Adam optimizer [21] and a learning rate of 10^{-3} . We also introduce **SELDNet+**, which follows the same data augmentation schema as the ATTNSELD+.

CGNet-STs and CGNet-STs-MultiACCDOA. CGNet-STs is an existing SELD network that incorporates equivariance to $SO(3)$ rotations, time translations, and scale changes through Clebsch–Gordan product networks [9]. We re-implemented the architecture from the original paper and verified functional equivalence against the authors’ reference code². However, the original implementation outputs separate DOA and SED predictions, which prevents the model from handling overlapping sound events of the same class. We therefore adapted CGNet-STs to the Multi-ACCDOA framework by widening its final layer to emit $K \times C$ first-order direction vectors (where K is the number of tracks and C is the number of classes) and widening the GRU readout to produce matching logits $z_{k,c}$. The final Multi-ACCDOA vector is then formed as $\mathbf{a}_{k,c} = \sigma(z_{k,c}) \mathbf{u}_{k,c} / \|\mathbf{u}_{k,c}\|$. As we found that the Multi-ACCDOA version performs better on the synthetic scenes than the original implementation (Section 4), we continued with this version for real-world STARSS23 dataset.

4. RESULTS

4.1. SELD Results

On the simulated scenes with measured RIRs (TAU2021), EquiSELD achieves a summary SELD score of $\mathcal{E}_S = 0.40$, outperforming all versions of SELDNet (lowest $\mathcal{E}_S = 0.47$) and CGNet-STs (lowest $\mathcal{E}_S = 0.51$, Table 1). Crucially, EQUISELD has $2.9 \times$ fewer parameters and uses approximately $19 \times$ less GPU hours than CGNet-STs to achieve this performance, showcasing the effectiveness of equivariant parametrization via the equivariant transformers

²<https://github.com/nttrd-mdlab/group-equiv-seld>

Table 1. SELD performance on STARSS23 and TAU2021, with training cost on TAU2021 (single A100). For STARSS23, “Det.” is the number of the 13 classes the model detects and the parenthesised LE is the class-averaged localization error restricted to those classes. We show the jackknife estimates of the $\mathcal{E}_S$ under the test-set recordings.

Model	Par.	GPU h↓	STARSS23						TAU2021					
			ER↓	F↑	LE (det.)↓	LR↑	Det.↑	$\mathcal{E}_S \downarrow$	ER↓	F↑	LE↓	LR↑	$\mathcal{E}_S \downarrow$	
SELDNet	0.75M	0.16	0.74	13.7	55.7° (33.1°)	26.5	11	0.66 [0.61, 0.69]	0.68	27.5	35.8°	42.8	0.55 [0.53, 0.56]	
SELDNet	2.25M	0.18	0.75	16.1	51.9° (28.6°)	29.5	11	0.65 [0.62, 0.68]	0.68	29.8	22.3°	46.6	0.51 [0.50, 0.52]	
SELDNet ⁺	2.25M	0.83	0.57	25.4	68.2° (18.5°)	30.0	9	0.60 [0.57, 0.63]	0.61	40.5	32.3°	49.4	0.47 [0.46, 0.49]	
CGNet-STS-MultiACCDOA	6.53M	14.5	0.80	11.3	78.3° (33.1°)	24.5	9	0.72 [0.68, 0.74]	0.73	30.1	25.3°	52.0	0.51 [0.50, 0.53]	
CGNet-STS	6.28M	12.1	—	—	—	—	—	—	0.79	26.2	31.9°	55.1	0.54 [0.52, 0.55]	
EquiSELD	2.25M	0.77	0.72	24.9	34.3° (22.1°)	39.7	12	0.56 [0.51, 0.59]	0.64	49.8	16.4°	63.9	0.40 [0.38, 0.41]	
EquiSELD-SO(3)	2.25M	0.77	0.69	21.4	31.7° (19.4°)	30.9	12	0.58 [0.50, 0.61]	0.60	51.3	15.9°	63.9	0.38 [0.37, 0.40]	
AttnSELD	2.25M	0.77	0.97	13.6	52.6° (29.5°)	30.6	11	0.70 [0.68, 0.74]	0.71	42.6	17.9°	57.6	0.45 [0.44, 0.47]	
AttnSELD ⁺	2.25M	0.85	0.82	25.1	33.6° (21.5°)	44.9	12	0.58 [0.51, 0.61]	0.67	42.2	18.8°	56.7	0.45 [0.43, 0.46]	

framework. Finally, on the more challenging and limited real-world scenes (STARSS23), EquiSELD achieves a summary SELD score of $\mathcal{E}_S = 0.56$, once more surpassing all versions of SELDNet (lowest $\mathcal{E}_S = 0.60$) and CGNet-STS ($\mathcal{E}_S = 0.72$).

Rotation vs. rotation and reflection equivariance. EQUiSELD-SO(3) achieves comparable summary SELD scores as EQUiSELD under $O(3)$ equivariance for both the synthetic and real-world scenes (Table 1, see overlapping jackknife confidence intervals). These results indicate that the gain in effectiveness of $O(3)$ equivariance (rotation and reflection equivariance) is comparable to the gain in effectiveness of $SO(3)$ equivariance (rotation equivariance only).

EquiSELD without equivariance constraints. ATTENSELD performs substantially worse on the synthetic scenes (TAU2021, $\mathcal{E}_S = 0.45$) as well as the real-world scenes (STARSS23, $\mathcal{E}_S = 0.70$, Table 1). These findings highlight the impact of the equivariance constraints on model performance. Training ATTENSELD⁺ with $O(3)$ augmentation improved performance on the real-world scenes at a small increase of the training cost, achieving a similar summary score as EQUiSELD ($\mathcal{E}_S = 0.58$, Table 1). However, performance of ATTENSELD⁺ did not improve on the synthetic scenes ($\mathcal{E}_S = 0.45$, Table 1).

Undetected classes. On STARSS23, LE_{CD} assigns 180° to undetected classes, coupling class coverage with localization precision. SELDNet⁺ and CGNet-STS each detect nine of 13 classes, yielding overall errors of 68.2° and 78.3°, respectively, versus 18.5° and 33.1° over detected classes. These restricted averages involve different class subsets and are therefore not directly comparable.

4.2. Orientation Robustness

We evaluate the spatial equivariance of each model on a sweep of 74 elements of $O(3)$: 12 Fibonacci-sphere directions with 3 in-plane rolls each, plus the identity, paired with their parity partners $-Q$. On test-fold recordings comparing $f(Qx)$ against $Qf(x)$, EquiSELD achieves exact equivariance (maximum deviation for EQUiSELD = 2.8×10^{-7}). Table 2 shows that SELDNet’s scores range from 0.65 to 0.82 and AttnSELD’s range from 0.70 to 0.86. Variants with $O(3)$ augmentations exhibit smaller residual variation as well, but EquiSELD attains this without augmentations and gives the best absolute performance.

4.3. Data Scaling

To compare how the effectiveness of inherent network equivariance scales with data quantity, we depict performance on TAU2021 as the training set increases from one to four folds (100–400 min, Fig-

Table 2. SELD score over $O(3)$ elements on STARSS23.

Model	Mean	Std	Min	Max
SELDNet	0.76	± 0.03	0.65	0.82
SELDNet ⁺	0.60	± 0.01	0.60	0.62
AttnSELD	0.75	± 0.04	0.70	0.86
AttnSELD ⁺	0.58	± 0.01	0.57	0.59
EquiSELD (ours)	0.56	± 0.00	0.56	0.56

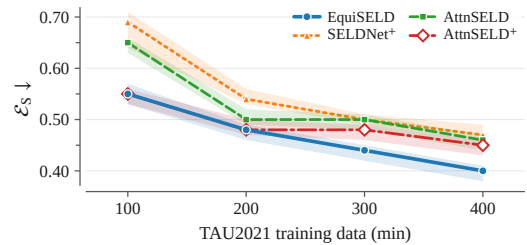


Fig. 2. Data scaling. Shaded areas are 95% confidence intervals.

ure 2). EquiSELD improves monotonically from $\mathcal{E}_S = 0.55$ to $\mathcal{E}_S = 0.40$, while SELDNet+, ATTENSELD and ATTENSELD+ show less improvement. This graph shows that for simulated scenes with measured RIRs, EquiSELD outperforms ATTENSELD+, especially at larger training-set sizes.

5. CONCLUSION

We proposed EQUiSELD, an $O(3)$ -equivariant attention network that encodes the rotation and reflection symmetry. The inherent equivariance of this low-parameter network requires no $O(3)$ augmentation to build rotation and reflection robustness. Experimental results show that EQUiSELD outperforms a prior equivariant SELD network and the SELD baseline system on both synthetic and real-world scenes. Moreover, EQUiSELD has $2.9\times$ fewer parameters and requires $19\times$ less training time than the preceding equivariant network. Finally, we show that the benefit of inherent equivariance scales with data size, consistent with findings on scaling equivariant transformers more broadly [22]. Future work will explore softening this hard constraint — through soft-equivariance penalties — to exploit the elevation priors that strict $O(3)$ and $SO(3)$ symmetry cannot, and extending the network to higher-order Ambisonics.

6. ACKNOWLEDGMENTS

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-14624.

7. REFERENCES

- [1] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, “Sound event detection in the dcase 2017 challenge,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 6, pp. 992–1006, June 2019.
- [2] Jun-Wei Yeow, Ee-Leng Tan, Santi Peksi, and Woon-Seng Gan, “Environmental acoustic intelligence through sound event localization and detection: a review,” *npj Acoustics*, vol. 1, no. 1, pp. 31, Dec 2025.
- [3] Jinbo Hu, Yin Cao, Ming Wu, Fang Kang, Feiran Yang, Wenwu Wang, Mark D. Plumbley, and Jun Yang, “PSELD-Nets: Pre-trained neural networks on a large-scale synthetic dataset for sound event localization and detection,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2845–2860, 2025.
- [4] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji, “STARSS23: an audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023, NIPS ’23, Curran Associates Inc.
- [5] Franz Zotter and Matthias Frank, *XY, MS, and First-Order Ambisonics*, pp. 1–22, Springer International Publishing, Cham, 2019.
- [6] Ryotaro Sato, Kenta Niwa, and Kazunori Kobayashi, “Ambisonic signal processing dnns guaranteeing rotation, scale and time translation equivariance,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 1449–1462, Mar. 2021.
- [7] Luca Mazzon, Yuma Koizumi, Masahiro Yasuda, and Noboru Harada, “First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 154–158.
- [8] Qing Wang, Ya Jiang, Shi Cheng, Maocheng Hu, Zhaoxu Nian, Pengfei Hu, Zeyan Liu, Yuxuan Dong, Mingqi Cai, Jun Du, and Chin-Hui Lee, “The NERC-SLIP system for sound event localization and detection of dcase2023 challenge,” *Tech. Rep.*, DCASE2023 Challenge, June 2023.
- [9] Risi Kondor, Zhen Lin, and Shubhendu Trivedi, “Cleb-sch–gordan nets: a fully fourier space spherical convolutional neural network,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, vol. 31, Curran Associates, Inc.
- [10] Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen, “A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection,” in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 125–129.
- [11] Yi-Lun Liao and Tess Smidt, “Equiformer: Equivariant graph attention transformer for 3d atomistic graphs,” in *International Conference on Learning Representations*, 2023.
- [12] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling, “SE(3)-transformers: 3d roto-translation equivariant attention networks,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020, NIPS ’20, Curran Associates Inc.
- [13] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh, “Set Transformer: A framework for attention-based permutation-invariant neural networks,” in *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov, Eds. 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, PMLR.
- [14] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenc, Andrea Tagliasacchi, and Leonidas J. Guibas, “Vector neurons: A general framework for so(3)-equivariant networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 12200–12209.
- [15] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror, “Learning from protein structure with geometric vector perceptrons,” in *International Conference on Learning Representations*, 2021.
- [16] Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji, “Multi-ACDDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 316–320.
- [17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [18] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen, “Overview and evaluation of sound event localization and detection in dcase 2019,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.
- [19] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.
- [20] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.
- [21] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2017.
- [22] Johann Brehmer, Sönke Behrends, Pim De Haan, and Taco Cohen, “Does equivariance matter at scale?,” *Transactions on Machine Learning Research*, 2025.