BEYOND GEOMETRIC AUGMENTATION: HYBRID EQUIVARIANT LIDAR DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

LiDAR detectors rely heavily on geometric augmentation to learn how object states change with scene geometry. We argue that the burden of geometric robustness should not fall on augmentation alone: geometric structure should also be built into the detector. We investigate this principle through HybridFusion, which couples a steerable equivariant pathway with a non-equivariant stream. Field-wise gate conditioning connects the streams throughout the network, while cross-stream attention and residual merging combine their features in the detection head. On nuScenes, a detailed TransFusion-L study compares standard, predominantly equivariant, and hybrid detectors with and without global geometric augmentation, with balanced resampling disabled. HybridFusion improves mAP by 6.55 percentage points without global augmentation and by 4.83 points with augmented training, with large reductions in orientation and velocity errors. Augmentation remains valuable: it adds a further 9.27 mAP points to the hybrid model. Improvements in mAP and NDS also extend to augmented CenterPoint and SECOND. A non-equivariant two-stream control fails to recover the hybrid gains, and removing gate conditioning lowers both scores. These detection results support sharing the burden between architecture and augmentation. Hybrid equivariance builds transformation structure into feature processing, improving prediction beyond what the augmented standard detector achieves while retaining the benefits of augmented training.

1 INTRODUCTION

LiDAR-based 3D object detection is a core component of autonomous driving perception. A detector must recover object locations, dimensions, orientations, and velocities from sparse geometric measurements. Rotating or translating a scene changes these quantities in predictable ways, making transformation structure central to the task. Modern detectors exploit this structure through geometric augmentation: rotations, translations, scaling, and reflections transform point clouds and labels together during training (Yan et al., 2018; Yin et al., 2021; Bai et al., 2022). Removing these operations can sharply reduce detection accuracy (Zhang et al., 2020; Fang et al., 2021).

This dependence raises a broader design question: how much of the burden of geometric robustness should be placed on augmentation? Augmentation exposes the model to transformed scenes, but the detector must still learn how to process their geometry from those examples. Our premise is that the architecture should carry part of this burden. A detector should build useful transformation structure into its features and operations, alongside the geometric variation provided by augmented training data.

Equivariant networks offer a second route. They organize features into fields with prescribed transformation laws and constrain their processing to respect those laws (Cohen & Welling, 2016; Weiler & Cesa, 2019; Weiler et al., 2018). This is a natural fit for LiDAR, where the input, intermediate spatial features, and predicted object states all carry geometric meaning. The important question is how much this architectural structure contributes alongside augmentation. Does it recover some of the accuracy lost when augmentation is removed? Does it still help when the training data are already augmented? Answering both questions is necessary to understand its role in a practical detector.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Driving scenes also contain information that does not follow a simple rotation symmetry. Road layout, sensor placement, ego-vehicle coordinates, and class-dependent heading patterns can all matter for prediction. Prior work on soft and partial equivariance provides ways to combine geometric structure with this flexibility (Finzi et al., 2021; Wang et al., 2022; Romero & Lohit, 2022), while equivariant detectors demonstrate its relevance to 3D detection (Wu et al., 2023). We bring these ideas together to study the interaction between hybrid equivariance and geometric augmentation in sparse-to-BEV LiDAR detection.

We develop EquiFusion and HybridFusion to examine this interaction. EquiFusion introduces steerable processing in the sparse 3D encoder and planar BEV pathway, connected by restriction from $SO(3)$ to $SO(2)$. HybridFusion keeps this structured pathway and adds a non-equivariant stream. The streams interact throughout the network through FiLM conditioning of scalar activation gates, and in the detection head through one cross-stream attention layer followed by residual merging. This design preserves the organization of geometric fields while allowing unconstrained features to shape their activation and contribute directly to prediction.

Our nuScenes experiments reveal a clear difference between the two designs. The predominantly equivariant EquiFusion model underperforms the standard TransFusion-L baseline, whereas HybridFusion improves detection both with and without global geometric augmentation. Under augmented training, it adds 4.83 mAP and 9.92 NDS points and lowers orientation and velocity errors. Augmentation also improves the hybrid model substantially. The strongest result in this comparison comes from combining the two mechanisms.

Our central contribution is to show how hybrid geometric structure changes the benefits of augmentation in LiDAR detection. We make this relationship explicit through a six-configuration TransFusion-L comparison, then extend the augmented standard-versus-hybrid evaluation to CenterPoint and SECOND. Both additional detectors improve in mAP and NDS. Architectural controls sharpen the result: a second non-equivariant stream with similar fusion mechanisms does not recover the gains, and removing FiLM reduces hybrid performance. Together, these findings support the paper's central argument: the burden of geometric robustness should not rest solely on augmentation. Building geometric structure into the detector provides an additional source of accuracy, including when geometric augmentation is already part of the training pipeline.

2 Related work

LiDAR detection and augmentation. Voxel-based detectors process point clouds with sparse convolutions, while BEV representations support efficient localization and prediction (Zhou & Tuzel, 2018; Yan et al., 2018; Lang et al., 2019). CenterPoint uses center-based prediction and TransFusion uses object queries (Yin et al., 2021; Bai et al., 2022). Geometric augmentation substantially affects detector performance (Zhang et al., 2020; Fang et al., 2021). Our experiments connect this training choice to the detector's feature structure by measuring what hybrid processing adds in both augmented and unaugmented settings.

Equivariant detection. Group and steerable networks encode prescribed feature transformation laws (Cohen & Welling, 2016; Weiler & Cesa, 2019; Weiler et al., 2018). TED processes transformed point clouds using shared sparse convolutions and aligns and aggregates their features for proposal generation and refinement (Wu et al., 2023). Its cross-grid attention combines aligned instance features. TED reports strong detection without rotation/reflection augmentation and evaluates transformed inputs, establishing relevant precedent for studying architecture and augmentation together. Our approach uses steerable $SO(3)/SO(2)$ fields and interactions between equivariant and non-equivariant streams. This lets us examine a different architectural route to geometric structure and, in particular, the benefit it retains under augmented training.

Soft and hybrid equivariance. Residual Pathway Priors combine constrained and unconstrained layers with priors favoring structured solutions (Finzi et al., 2021). Approximate and partial equivariance likewise relax fixed symmetry assumptions (Wang et al., 2022; Romero & Lohit, 2022). HybridFusion shares the motivation of retaining geometric bias with non-equivariant flexibility. Its specific interactions are FiLM conditioning of field-wise activation gates throughout the network and cross-stream attention followed by residual merging in the detection head. We apply this com-

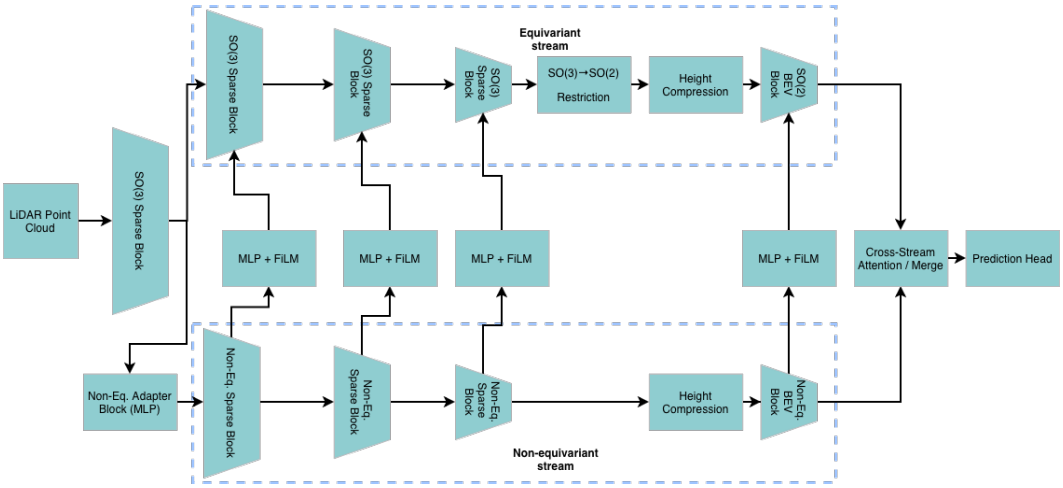


Figure 1: HybridFusion architecture overview. FiLM conditions one scalar activation gate per field throughout the network. The cross-stream attention/merge block denotes one cross-stream attention layer followed by residual merging within the detection head, before final prediction.

bination throughout a sparse-to-BEV detector and study its interaction with geometric augmentation. The two-stream and no-FiLM controls separate the evidence for the overall hybrid design from the contribution of its gate-conditioning mechanism.

3 Hybrid Equivariant Detection

Each detector voxelizes the point cloud, encodes sparse 3D features, projects them into BEV, and predicts object states. EquiFusion replaces the main feature processing with equivariant modules. HybridFusion adds a spatially aligned non-equivariant stream. The hybrid mechanisms are shared across TransFusion-L, CenterPoint, and SECOND, while their detector-specific prediction formulations are retained.

3.1 Structured fields in the sparse encoder

The sparse encoder uses real $SO(3)$ irreps of orders $0, 1, 2$ with a base multiplicity ratio of $4{:}2{:}1$. Before height aggregation, the representation is restricted to yaw rotations and projected into planar $SO(2)$ fields. The BEV pathway uses planar equivariant processing. The restriction, projection, and equivariant attention with modified IRPE are described in Appendix A.

A feature field associates a vector $f(x)$ with each occupied voxel location x. For a rotation R, its prescribed transformation is

$$[T_R f](x) = \rho(R) f(R^{-1}x). \tag{1}$$

The spatial argument moves with the scene, while $\rho(R)$ specifies how the components at that location transform. Scalar fields remain unchanged in value, vector fields rotate their components together, and higher-order fields encode more complex directional information. This differs from a standard feature tensor whose channels have no prescribed transformation law. An equivariant layer L satisfies $L(T_R f) = T_R L(f)$ in the idealized continuous formulation; discretization introduces the practical qualifications discussed below.

We use the base representation

$$\rho_{3\mathrm{D,base}} = 4\rho_0^{SO(3)} \oplus 2\rho_1^{SO(3)} \oplus \rho_2^{SO(3)}. \tag{2}$$

The three orders have real dimensions 1, 3, and 5, respectively, so one base copy contains 15 channels grouped into seven fields. Wider layers repeat this representation. The distinction between channels and fields is important: the five components of an order-two field form one geometric object, not five independently transformable scalar features.

The sparse encoder retains sparse spatial computation but uses steerable kernels. For input and output representations ρ_{in} and ρ_{out}, the kernel at relative offset r obeys

$$\Psi(Rr) = \rho_{\text{out}}(R)\Psi(r)\rho_{\text{in}}(R)^{-1}. \tag{3}$$

Kernels are expanded in a basis satisfying this constraint, with learned scalar coefficients (Weiler et al., 2018; Lin et al., 2021; Cesa et al., 2022). Consequently, the network learns filters within a structured family rather than learning an independent filter for every orientation. Gated nonlinearities preserve the coupling between components of each field. HybridFusion pairs this path with unconstrained features on aligned spatial locations, allowing information without a prescribed rotation law to enter through the conditioning and fusion mechanisms below.

3.2 From three-dimensional fields to planar BEV features

Height compression changes the relevant symmetry. A general three-dimensional rotation mixes height and horizontal position, so collapsing height cannot preserve the full $SO(3)$ action. Rotations about the vertical axis do preserve the BEV plane. We therefore first restrict the representation to this yaw subgroup, then map the restricted fields into the planar representation used by the BEV backbone.

In real coordinates, restriction decomposes the first three orders as

$$\rho_0^{SO(3)}\downarrow = \rho_0^{SO(2)}, \quad \rho_1^{SO(3)}\downarrow = \rho_0^{SO(2)} \oplus \rho_1^{SO(2)}, \quad \rho_2^{SO(3)}\downarrow = \rho_0^{SO(2)} \oplus \rho_1^{SO(2)} \oplus \rho_2^{SO(2)}. \tag{4}$$

Thus the 15-channel sparse base representation becomes $7\rho_0^{SO(2)} \oplus 3\rho_1^{SO(2)} \oplus \rho_2^{SO(2)}$ without losing channels at the restriction step. A planar equivariant projection then maps it to the BEV base type $4\rho_0^{SO(2)} \oplus 2\rho_1^{SO(2)} \oplus \rho_2^{SO(2)}$, which has 10 real channels because nonzero-order real $SO(2)$ fields are two-dimensional. Restriction and projection therefore serve distinct roles: the former changes the group under consideration, while the latter selects the feature type and width used downstream.

Height aggregation must commute with the planar representation for the structured path to retain its transformation law. This holds for summation or channel-consistent linear aggregation because yaw leaves the height coordinate unchanged. The resulting BEV fields feed planar steerable blocks. This transition adapts the feature structure to each stage: three-dimensional local geometry in the sparse encoder and planar geometry in BEV.

3.3 Equivariant attention and relative position

Attention within the structured path follows the HarmFormer principle of invariant attention weights acting on equivariant values (Karella et al., 2024). Let q_i, k_j, and v_j be equivariant projections of features at locations i and j. Query and key fields use the same orthogonal representation. Their inner product is invariant since $(\rho q_i)^T(\rho k_j) = q_i^T k_j$. Accordingly,

$$a_{ij} = \text{softmax}_j\left(\frac{q_i^T k_j + q_i^T r_{ij}}{\sqrt{d}}\right), \qquad o_i = \sum_j a_{ij} v_j, \tag{5}$$

is equivariant when the relative positional feature r_{ij} transforms under the query representation. Scalar weights determine how much to attend to each location without independently mixing field components.

We use a modified IRPE product encoding (Wu et al., 2021). For planar displacement Δ_{ij}, its distance determines a radial bin; the displacement is rescaled to the representative radius of that bin while retaining its direction. Evaluating an equivariant positional kernel at this offset produces r_{ij}. Distance binning is rotation invariant, and the directional encoding transforms with the displacement, making the positional term in Eq. 5 invariant. This construction preserves relative directional information rather than treating absolute coordinates as unrelated embeddings. Appendix A.7 provides the radial quantization and positional-kernel formulation.

3.4 Conditioning fields and merging the streams

FiLM on scalar activation gates. For field h_i, the non-equivariant stream predicts a scale γ_i and shift β_i for its scalar activation gate $\tilde{g}_i$:

$$g_i = \gamma_i \tilde{g}_i + \beta_i. \tag{6}$$

The resulting gate controls the entire field, using one scalar for all of its components. FiLM therefore acts directly on gating scalars, not through component-wise affine transformations of non-scalar features. It is applied throughout the architecture. Sharing a gate preserves component coupling, but conditioning it on unconstrained features intentionally relaxes exact equivariance. The modulation is initialized near identity using $\gamma = 1 + \tanh(\tilde{\gamma})$ and $\beta = \tanh(\tilde{\beta})$.

For example, an order-one three-dimensional field has three components but receives one modulated activation gate. Its components are jointly amplified or suppressed; FiLM does not independently shift its three coordinates. This makes the conditioning sensitive to the field organization. However, component sharing alone is insufficient to guarantee equivariance: the conditioning stream can change the gate when the input rotates. The model therefore retains structured filters and field organization while permitting input-dependent departures from the strict transformation law. Applying this mechanism throughout the network allows the unconstrained stream to influence intermediate feature selection before the final head-level combination.

Head-level attention and residual merging. One cross-stream attention layer in the detection head uses equivariant features as queries and non-equivariant features as keys and values:

$$Q = F^{\text{eq}} W_Q, \quad K = F^{\text{non}} W_K, \quad V = F^{\text{non}} W_V, \qquad C = \text{softmax}(QK^T/\sqrt{d})V. \tag{7}$$

Here feature matrices stack spatial or object features as rows, and d is the query/key dimension. Residual merging follows this attention layer in the head, before final prediction. This interaction is intentionally unconstrained. It is separate from the equivariant within-stream attention described in the appendix. The two forms of attention serve different purposes. Within-stream attention propagates structured information using invariant weights; cross-stream attention selects unconstrained information relevant to the structured queries. Residual merging subsequently combines the aligned head features before prediction. FiLM, attention, and merging are therefore complementary interaction points rather than interchangeable names for a single operation.

Scope of equivariance. Voxelization, finite grids, query assignment, decoding, and post-processing limit end-to-end equivariance. HybridFusion additionally relaxes it through conditioning and head-level fusion. Equivariance thus provides a structured feature-processing bias within the complete detector.

4 Experiments

4.1 Experimental setup

We evaluate the ten nuScenes detection classes on the validation split (Caesar et al., 2020). The standard and hybrid configurations share ten LiDAR sweeps, coordinate/intensity/timestamp features, velocity prediction, and balanced resampling disabled unless stated otherwise. Spatial preprocessing matches within each standard/hybrid pair. TransFusion-L and CenterPoint use 0.075 m horizontal voxels and a $[-54, 54]$ m horizontal range; SECOND uses 0.1 m horizontal voxels and a $[-51.2, 51.2]$ m range.

The detailed TransFusion-L study includes standard, predominantly equivariant, and hybrid models with and without global geometric augmentation. CenterPoint and SECOND compare standard and hybrid models under augmented training. Our experiments use one training seed per configuration. We report mAP, NDS, and nuScenes true-positive errors. Hybrid variants have 7.4–8.0% more parameters than their standard counterparts (Table 4). The two-stream control provides a separate check on the effect of adding a stream.

The configurations use 20 epochs, Adam OneCycle, batch size 8 per GPU, and weight decay 0.01. Learning rates and augmentation policies vary as listed in Appendix B. Hybrid configurations disable flipping but retain scaling. Disabling global augmentation retains ground-truth sampling. A separate balanced-resampling comparison evaluates hybrid TransFusion-L alongside the published OpenPCDet standard baseline.

The comparisons address three questions. First, does the hybrid design improve detection both with and without geometric augmentation, and does augmentation still benefit the hybrid model? Second,

Table 1: TransFusion-L results on nuScenes. Aug denotes the model-specific augmented settings described in Appendix B.

Method	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	mAP ↑	NDS ↑
TransFusion	0.3369	0.2674	0.6781	1.2685	0.2538	0.3984	0.4456
EquiFusion	0.3656	0.2743	0.7028	1.3935	0.2644	0.3529	0.4157
HybridFusion	0.3284	0.2654	**0.4146**	0.4444	0.1945	0.4639	0.5672
TransFusion-Aug	**0.3080**	**0.2595**	0.5369	1.1424	0.2792	0.5083	0.5157
EquiFusion-Aug	0.3355	0.2636	0.6328	1.2738	0.3238	0.4403	0.4646
HybridFusion-Aug	0.3178	0.2610	0.4587	**0.4053**	**0.1911**	**0.5566**	**0.6149**

do improvements extend to detector families with different prediction formulations? Third, can the gains be reproduced by a non-equivariant two-stream design, and what does field-wise conditioning add to the hybrid architecture? The six-row TransFusion-L comparison addresses the first question; the additional detectors and architectural controls address the latter two.

We report scores on a 0–1 scale and differences in percentage points. mAP summarizes class detection accuracy, while NDS combines detection accuracy with true-positive prediction quality. The error metrics separately measure translation (mATE), scale (mASE), orientation (mAOE), velocity (mAVE), and attributes (mAAE); lower is better. Reporting both aggregate scores and these errors helps distinguish improved detection from improvements to the geometry and motion of detected objects.

4.2 TransFusion-L: architecture and augmentation

Table 1 shows gains from hybrid processing in both training settings. Without global augmentation, HybridFusion improves mAP from 0.3984 to 0.4639 and NDS from 0.4456 to 0.5672: gains of 6.55 and 12.16 percentage points. Under augmented training, its mAP is 0.5566 compared with 0.5083 for TransFusion-L, and NDS is 0.6149 compared with 0.5157: gains of 4.83 and 9.92 points.

Augmentation remains beneficial. Augmentation improves standard TransFusion-L by 10.99 mAP and 7.01 NDS points, and HybridFusion by 9.27 mAP and 4.77 NDS points. Thus, architectural structure does not remove the value of augmented training. The hybrid model without global augmentation exceeds augmented TransFusion-L in NDS (0.5672 versus 0.5157), but remains below it in mAP (0.4639 versus 0.5083). Hybrid processing recovers substantial prediction quality without global augmentation, but combining the two produces the best overall detector in this comparison.

Geometry-sensitive errors. HybridFusion lowers orientation and velocity errors relative to the standard model in both settings. For augmented training, mAOE decreases from 0.5369 to 0.4587 and mAVE from 1.1424 to 0.4053. The improvement is not uniform across metrics: augmented HybridFusion has slightly higher translation and scale errors than augmented TransFusion-L. Within HybridFusion, augmentation improves overall mAP and NDS but increases mAOE from 0.4146 to 0.4587.

EquiFusion underperforms the standard and hybrid models in these settings. The hybrid construction succeeds where this predominantly equivariant variant does not, motivating the combination of structured and unconstrained processing.

How the augmentation benefit changes. The mAP gain from augmentation is 1.72 points smaller for HybridFusion than for standard TransFusion-L; the corresponding NDS gain is 2.24 points smaller. Equivalently, the hybrid advantage narrows under augmented training but remains positive. This suggests overlapping benefits, yet neither mechanism exhausts the benefit of the other. Their combination is therefore the important design choice. The complete augmentation recipes are documented in Appendix B.

Table 2: Class AP for the standard and hybrid TransFusion-L models. The two training settings correspond to Table 1. Values are retained from the detailed TransFusion-L evaluation.

	Without global augmentation		With augmentation	
Class	Standard	Hybrid	Standard	Hybrid
Car	0.7188	0.7837	0.7862	0.8412
Truck	0.3368	0.4173	0.4895	0.5066
Construction vehicle	0.0402	0.0837	0.1340	0.1840
Bus	0.4893	0.5074	0.6586	0.6302
Trailer	0.2577	0.2887	0.2942	0.3432
Barrier	0.5647	0.6056	0.6409	0.6474
Motorcycle	0.3295	0.3738	0.4520	0.5765
Bicycle	0.0931	0.2610	0.2569	0.3615
Pedestrian	0.6686	0.7613	0.7624	0.8260
Traffic cone	0.4854	0.5561	0.6079	0.6494

Table 3: Augmented nuScenes results for additional detectors. Each pair uses identical spatial preprocessing.

Detector	Variant	mAP	NDS
CenterPoint	Standard	0.5535	0.6319
CenterPoint	Hybrid	0.5674	0.6531
SECOND	Standard	0.4494	0.5667
SECOND	Hybrid	0.4638	0.5945

4.3 Class-level effects in TransFusion-L

Table 2 shows that the aggregate improvement is distributed across classes. Without global augmentation, HybridFusion improves AP for all ten classes, with the largest absolute gain for bicycles (16.79 points). Pedestrians and trucks improve by 9.27 and 8.05 points, respectively. With augmentation, HybridFusion improves nine of ten classes. Motorcycles and bicycles gain 12.45 and 10.46 points, while the bus class decreases by 2.84 points. The bus result is the main exception to an otherwise broad improvement.

Augmentation improves AP for every class within HybridFusion. Its largest absolute changes include motorcycles (20.27 points), buses (12.28 points), and construction vehicles (10.03 points). These class-level results reinforce the aggregate conclusion that architectural structure leaves a substantial role for augmented training. The large hybrid gains for motorcycles and bicycles are especially notable because they remain visible when comparing the two augmented models, not only when recovering accuracy from an unaugmented baseline.

4.4 Additional detectors

Table 3 reports additional detector comparisons. Hybrid processing improves CenterPoint from 0.5535 to 0.5674 mAP and from 0.6319 to 0.6531 NDS, increases of 1.39 and 2.12 percentage points. For SECOND, mAP increases from 0.4494 to 0.4638 and NDS from 0.5667 to 0.5945, gains of 1.44 and 2.78 percentage points. HybridFusion therefore improves both metrics across all three detector families under augmented training. Table 4 places these comparisons in the context of model capacity. The hybrid changes add approximately 0.67 million parameters to each detector, but the score changes differ across families: CenterPoint and SECOND show smaller gains than TransFusion-L. The improvement extends across detector families, while its magnitude depends on the baseline and training conditions. The same conditioning and fusion mechanisms work with each prediction formulation without requiring an identical detection head.

4.5 Architectural controls

Non-equivariant two-stream control. The control uses two non-equivariant streams with similar FiLM conditioning, cross-stream attention, residual merging, and augmented training. It tests

Table 4: Model size for standard and hybrid detector pairs. Counts are total parameters from model initialization.

Detector	Standard	Hybrid	Increase
TransFusion-L	8,306,030	8,974,222	8.04%
CenterPoint	8,941,590	9,608,822	7.46%
SECOND	9,035,004	9,702,236	7.38%

Table 5: Augmented TransFusion-L architectural controls. The two-stream baseline uses non-equivariant processing with similar fusion mechanisms. The no-FiLM variant retains the full hybrid training recipe.

Configuration	mAP	NDS
Standard single stream	0.5083	0.5157
Non-equivariant two stream	0.4876	0.4983
HybridFusion without FiLM	0.5417	0.5992
HybridFusion full	**0.5566**	**0.6149**

whether an additional stream with these interaction mechanisms can reproduce the hybrid improvement without equivariant feature processing. Table 5 reports 0.4876 mAP and 0.4983 NDS. Augmented HybridFusion obtains 0.5566 mAP and 0.6149 NDS, a difference of 6.90 mAP and 11.66 NDS percentage points. The two-stream baseline also falls below the augmented single-stream TransFusion-L result (0.5083 mAP, 0.5157 NDS), so adding a second unconstrained stream does not by itself reproduce the hybrid gains in these runs. The dual-stream model has 8.409M parameters compared with 8.974M for HybridFusion, a 6.72% difference.

Field-wise gate conditioning. The no-FiLM run obtains 0.5417 mAP and 0.5992 NDS. Relative to the full augmented HybridFusion result (0.5566 mAP, 0.6149 NDS), these are reductions of 1.49 mAP and 1.57 NDS percentage points. The no-FiLM run uses the same augmentation and other settings as the full augmented model, isolating the removal of FiLM.

Removing FiLM does not eliminate the hybrid advantage: the resulting model still exceeds the augmented standard baseline by 3.34 mAP and 8.35 NDS points. The full model improves further, suggesting that gate conditioning contributes to a design whose remaining structured path and head fusion are already useful. Attention and residual merging remain active in this ablation. Taken together, the controls identify two useful parts of the design: the hybrid model is already strong without FiLM, and field-wise conditioning makes it better. Simply adding a second stream with similar fusion mechanisms does not deliver the same benefit.

4.6 Balanced resampling

With balanced resampling, the hybrid result reaches 0.6548 mAP and 0.7024 NDS, compared with the published OpenPCDet TransFusion-L result of 0.6458 mAP and 0.6943 NDS (OpenPCDet Development Team, 2020). The differences are 0.90 mAP and 0.81 NDS percentage points. This comparison provides context under a stronger training recipe. Because the standard row is externally reported, comparison with our resampling-disabled runs does not isolate the causal effect of resampling alone. The core results and the published baseline therefore correspond to different training settings: the core study disables balanced resampling, whereas this comparison enables it. All of our detector pairs use the same ten-sweep input; the orientation and velocity comparisons do not mix single- and multi-sweep models. The hybrid model also improves on this stronger reference, although the margin is smaller than in the resampling-disabled study. This result extends the comparison to a widely used training setting while making the dependence of the gain on the reference recipe explicit.

Table 6: TransFusion-L results with balanced resampling. The standard row is quoted from OpenPCDet; the hybrid row is our experiment. Scores use the same 0–1 scale as the other tables.

Variant	Source	mAP	NDS
Standard	OpenPCDet	0.6458	0.6943
Hybrid	Our experiment	0.6548	0.7024

5 Discussion and limitations

The results support a shift in how we assign the burden of geometric robustness. Augmentation supplies transformed training examples; the architecture can also encode how features should respond to geometry. HybridFusion shows the value of giving the detector this structure directly. It improves the detailed TransFusion-L comparison in both training settings, and the strongest result comes from combining hybrid processing with augmentation. Its lower orientation and velocity errors show that the benefit extends beyond aggregate detection accuracy to the quality of predicted object states. The gains on CenterPoint and SECOND further show that the design is useful across different detection formulations.

The controls help explain what makes the combination effective. A second unconstrained stream with similar interactions performs worse than the standard baseline, while the hybrid model improves even without FiLM. Restoring FiLM then improves it further. This points to a useful interaction between structured features and learned conditioning, rather than a gain that follows automatically from adding a stream. EquiFusion's lower scores also favor the hybrid construction in this comparison.

Limitations. The three-detector study establishes the value of the hybrid design on nuScenes, with TransFusion-L providing the detailed augmentation analysis. The comparisons evaluate the complete hybrid design, with model sizes and training recipes documented for each detector. EquiFusion's lower performance is specific to the tested construction. Evaluating additional datasets, test-time transformations, and computational costs would further establish the design's scope.

The hybrid models use rotation, translation, and scaling augmentation, while random reflections are disabled. Future work will investigate how to incorporate reflection augmentation effectively into hybrid equivariant detectors, including its interaction with feature representations and cross-stream conditioning. We are also conducting balanced-resampling experiments across all model variants, extending the TransFusion-L results reported here. We plan to complete this evaluation before the camera-ready version and include the full results in that revision. The current CenterPoint, SECOND, and balanced-resampling experiments have not yet undergone full hyperparameter tuning. We plan to complete tuning for these settings before the camera-ready version and report the updated results alongside the final training configurations.

6 Conclusion

Geometric augmentation and hybrid equivariance improve LiDAR detection together. HybridFusion couples steerable fields with unconstrained features through gate conditioning and head-level fusion, improving TransFusion-L both with and without global augmentation. Augmentation drives its strongest performance, adding 9.27 mAP points, while the hybrid design adds 4.83 mAP points over the augmented standard model. Gains on CenterPoint and SECOND, together with the architectural controls, strengthen the case for this combination. These findings support a clear design principle: the burden of geometric robustness should not be placed solely on augmentation. Geometric structure should also be built into the detector, so that architecture and training data work together to improve geometric prediction.

AI USE STATEMENT

Generative AI assisted with manuscript preparation, literature comparison, and consistency checks. All experimental results were obtained from detector evaluations. The authors take responsibility for the paper's content.

ETHICS STATEMENT

This work studies perception for autonomous driving using nuScenes. Improved validation performance does not establish safety for deployment. Detection failures and performance differences across object classes remain relevant limitations when considering downstream use.

REPRODUCIBILITY STATEMENT

Appendix A describes the feature representations and hybrid interactions. Appendix B documents training, augmentation, data settings, and parameter counts. The main text distinguishes our experiments from the externally reported OpenPCDet baseline and identifies the changes made in each architectural control.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WE4qe9xlnQw.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Jin Fang, Xinxin Zuo, Dingfu Zhou, Shengze Jin, Sen Wang, and Liangjun Zhang. Lidar-aug: A general rendering-based augmentation framework for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4710–4720, 2021.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30037–30049, 2021.

Tomáš Karella, Adam Harmanec, Jan Kotera, Jan Blažek, and Filip Šroubek. Harmformer: Harmonic networks meet transformers for continuous roto-translation equivariance. *arXiv preprint arXiv:2411.03794*, 2024.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12697–12705, 2019.

Jiehong Lin, Hongyang Li, Ke Chen, Jiangbo Lu, and Kui Jia. Sparse steerable convolutions: An efficient learning of se (3)-equivariant features for estimation and tracking of object poses in 3d space. *Advances in Neural Information Processing Systems*, 34:16779–16790, 2021.

OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

David W Romero and Suhas Lohit. Learning partial equivariances from data. *Advances in Neural Information Processing Systems*, 35:36466–36478, 2022.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pp. 23078–23091. PMLR, 2022.

Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural information processing systems*, 31, 2018.

Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3d object detection for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2795–2802, 2023. doi: 10.1609/aaai.v37i3.25380.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10033–10041, 2021.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

Wenwei Zhang, Zhe Wang, and Chen Change Loy. Exploring data augmentation for multi-modality 3d object detection. *arXiv preprint arXiv:2012.12741*, 2020.

Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.