# AgentMix: Robust Camera-Only Collaborative Perception via BEV Feature Augmentation

Rishabh Bhatia[1], Till Beemelmanns[1], Alexey Nekrasov[2], Bastian Leibe[2]
[1]Institute for Automotive Engineering (ika), RWTH Aachen University
[2]Visual Computing Institute, RWTH Aachen University
rishabh.bhatia@rwth-aachen.de

## Abstract

*Camera-only intermediate-fusion collaborative perception in Vehicle-to-Everything (V2X) systems achieves high accuracy under clean conditions, but degrades severely when collaborators are corrupted by noise or adverse environmental conditions. This brittleness arises because standard fusion mechanisms cannot effectively distinguish between reliable and corrupted feature maps. Consequently, when agents exchange Bird's-Eye View (BEV) representations, local perception failures propagate across the network, contaminating the fused scene for all participants. Image-level augmentation alone is insufficient, as it acts before the view transform and multi-agent alignment steps where corruption becomes most harmful. We propose* ***AgentMix****, an effective training-time augmentation for the shared BEV features. AgentMix combines per-agent BEV translation perturbation with cross-agent BEV-CutMix feature mixing, exposing the fusion module to spatial misalignment and locally inconsistent collaborators during training, with no additional cost at inference. On OPV2V with 13 corruption types across five severity levels, AgentMix improves accuracy under corruption by up to* ***+50.8%*** *relative to standard training. Crucially, these gains are consistent across three different intermediate-fusion architectures and do not come at the expense of clean accuracy.*

## 1. Introduction

Camera-only collaborative perception lets connected vehicles exchange compact BEV features before fusion and detection [1, 12, 13, 18, 20]. This intermediate-fusion paradigm is inherently fragile: once one collaborating agent is corrupted, its BEV feature map distorts the fused scene for every participant.

RCP-Bench [2] made this *corruption propagation* failure mode visible under weather, blur, noise, and sensor-failure corruptions. The core issue is that the fusion module receives already-degraded BEV features with no mechanism to reject them. Image-level augmentation acts before BEV lifting, alignment, and multi-agent fusion. Moreover, unlike LiDAR corruption where input-level geometric remedies are natural, camera corruption shifts appearance statistics that become unstable BEV features after lifting and warping. Existing collaborative methods focus more on communication efficiency or pose robustness than on corruption-aware feature regularization [4, 6, 19]. The key question is therefore not whether the encoder preserves its representations under corruption, but whether the fusion stage can arbitrate between partially reliable agents.

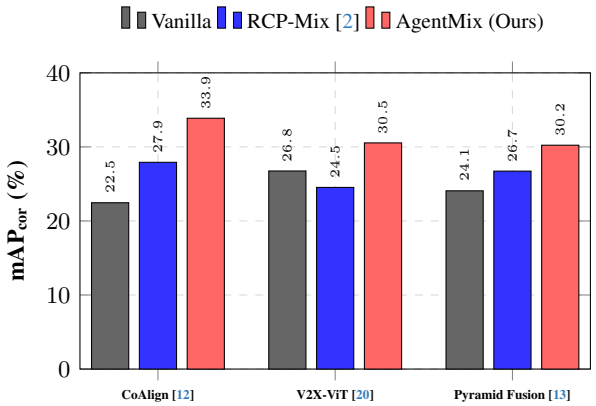


Figure 1. **AgentMix improves corrupted AP across all models.** CoAlign [12], V2X-ViT [20], and Pyramid Fusion [13] on OPV2V under the RCP-Bench protocol.

We propose **AgentMix**, a training-time augmentation applied directly to the shared BEV features before fusion. It comprises two complementary operations: (i) per-agent BEV translation perturbation, which trains the fusion module to handle small pose inconsistencies between agents; and (ii) BEV-CutMix [24], which pastes a randomly sampled rectangular patch from one agent's BEV features onto another agent's BEV map, forcing the fusion module to produce reliable detections even when parts of the shared

representation are spatially inconsistent or corrupted. As shown in Figure 1, AgentMix improves corrupted AP by up to **+50.8%** relative to standard training while maintaining clean accuracy across all three evaluated fusion architectures.

This work makes the following contributions: (1) We identify corruption propagation at the fusion-stage BEV representation as a major robustness bottleneck in the evaluated camera-only collaborative perception setting. (2) We propose AgentMix, a training-time augmentation that couples per-agent translation perturbation with cross-agent BEV-CutMix [24] to harden the shared representation. (3) We validate on OPV2V [21] under the RCP-Bench corruption protocol against three baselines: standard training (vanilla); an image-level *MixUp* [23] baseline; and RCP-Mix [2], the corruption-specific augmentation bundled with the benchmark. We implement MixUp without label mixing, as merging labels across agents with different viewpoints requires non-trivial spatial transformations that we leave to future work. AgentMix outperforms all three across CoAlign [12], V2X-ViT [20], and Pyramid Fusion [13], while preserving or improving clean accuracy.

## 2. Related Work

**Collaborative perception** approaches range from late fusion, where agents share final predictions, to intermediate fusion, where agents share and fuse intermediate BEV features before detection [1, 5, 10, 12, 13, 18, 20]. These methods are primarily designed and evaluated for clean conditions, and assume reliable collaborator features, an assumption that breaks under corruption. Alignment, communication pruning, and consensus modules [4, 6, 11, 19] improve nominal robustness but do not explicitly train the network to handle an unreliable collaborator.

**Corruption robustness** remedies still largely operate at the input level. In single-agent vision, corruption is benchmarked for image classification [3], and methods such as MixUp [23], CutMix [22], and test-time adaptation [16] were not introduced for BEV-level collaborative settings. For LiDAR-driven detection, domain generalization methods [8, 9, 17] broaden the training distribution but do not address the camera-only failure mode where corrupted BEV features propagate after lifting and warping. RCP-Bench [2] introduces a benchmark to measure this failure mode and proposes RCP-Mix [2], a global feature-statistic baseline inspired by style mixing [7]. AgentMix takes a different approach: it applies structural perturbations directly to the exchanged spatial representation, targeting the stage where corruption propagates.

## 3. Method

The design of AgentMix is motivated by the observation that in collaborative systems, perception failures are often agent-specific and spatially localized. While image-level augmentations can regularize individual encoders, they do not explicitly train the fusion module to handle corrupted or misaligned feature maps. By treating the agent as the fundamental unit of information and perturbing these units directly in the shared BEV space, we force the fusion stage to learn robust arbitration mechanisms that survive local feature inconsistencies.

### 3.1. Problem Formulation

Agent $i$ encodes its multi-view images $X_i$ into a BEV feature map via Lift-Splat-Shoot [14]:

$$F_i = \mathrm{LSS}(E(X_i)) \in \mathbb{R}^{C \times H_b \times W_b}. \tag{1}$$

The ego agent $i$ receives collaborator features $\{F_j\}_{j \neq i}$, aligns them via relative poses to obtain $\{\tilde{F}_j\}$, and fuses them with a learned fusion operator:

$$F_{\text{fused}}(p) = \textstyle\sum_j \alpha_j(p) \tilde{F}_j(p), \tag{2}$$

where $p$ denotes a spatial position in the shared lattice. Under corruption, a degraded collaborator can still transmit unstable features that are assigned a high weight $\alpha_j$ by the fusion module, distorting the fused representation. To tackle this overreliance on potentially unreliable collaborators, this work introduces a BEV-stage augmentation strategy.

### 3.2. AgentMix

To this end, AgentMix operates after BEV encoding and before fusion, as shown in Figure 2, augmenting the exchanged BEV tensor directly. Crucially, because this augmentation operates strictly on the intermediate feature representations, we can apply it without altering the ground-truth bounding box labels, treating AgentMix as a pure training-time robustness augmentation for feature-space misalignment.

**Agent-level translation.** Each aligned agent feature map $\tilde{F}_i$ is translated independently:

$$\tilde{F}'_i(x, y) = \tilde{F}_i(x - \Delta x,\, y - \Delta y), \quad \Delta x, \Delta y \sim \mathcal{U}(-d, d), \tag{3}$$

where $d$ is the maximum spatial shift. We implement these continuous sub-pixel shifts using bilinear interpolation to accurately simulate real-world spatial misalignment.

**BEV-CutMix [24].** For ego agent $i$ and donor $j \neq i$, a random rectangle $\mathcal{R}$ with target area ratio $\lambda$ is transferred from the translated donor feature $\tilde{F}'_j$ into the translated ego feature $\tilde{F}'_i$:

$$\tilde{F}''_i = M \odot \tilde{F}'_j + (\mathbf{1} - M) \odot \tilde{F}'_i, \tag{4}$$

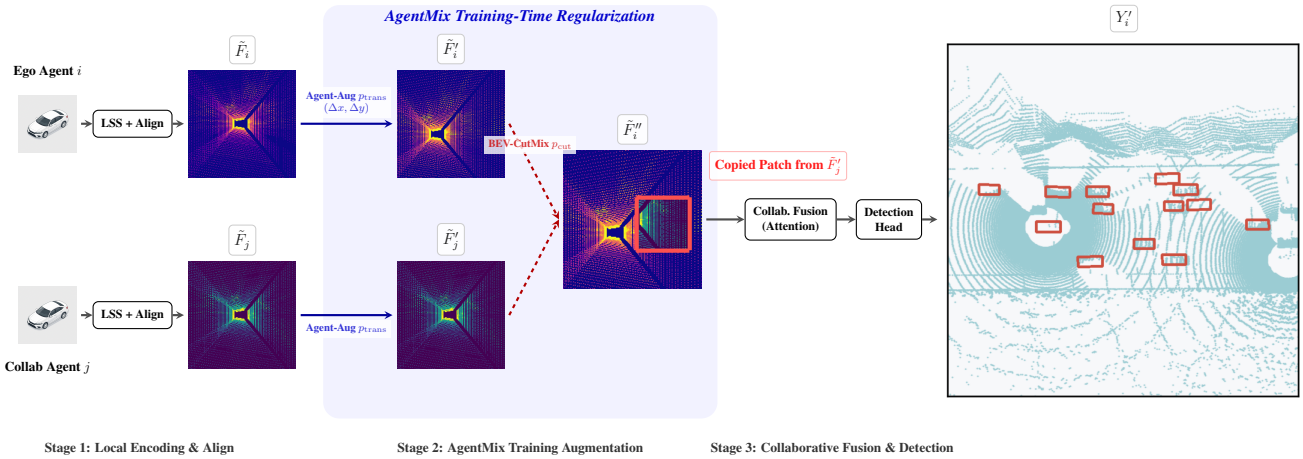


Figure 2. **AgentMix training pipeline.** After BEV encoding and alignment, each agent feature (both ego and collaborators) is independently translated with probability $p_{\text{trans}}$ to simulate small alignment errors. Cross-agent BEV-CutMix then stochastically copies local patches from a donor to the ego feature with effective probability $p_{\text{cut}}$. Both operations are training-only.

where $M \in \{0,1\}^{H_b \times W_b}$ is a binary spatial mask representing the rectangle $\mathcal{R}$. This preserves spatial structure while injecting local cross-agent disagreement, which is closer to practical corruption behavior than global mixing. The operation is deliberately local: in collaborative corruption settings, failures are often patchy. One region may be blurred or missing while the rest remains informative, so local patch exchange trains the network to arbitrate between conflicting collaborators without erasing scene context.

**Training policy.** AgentMix is applied stochastically during training via a global augmentation gate and independent per-operation sampling. This exposes the detector to a controlled mix of clean and perturbed collaborative features while leaving the base detector loss unchanged.

# 4. Experiments

## 4.1. Setup

We evaluate on OPV2V [21] with CoAlign [12], V2X-ViT [20], and Pyramid Fusion [13], which are strong, representative camera-only intermediate-fusion approaches with diverse fusion mechanisms. All use EfficientNet-B0 [15] and LSS [14]. All models are trained on clean data; corruptions are applied at test time via RCP-Bench [2] (13 types, 5 severity levels). We report clean AP@0.5 and $\text{mAP}_{\text{cor}}$ (AP@0.5 averaged over all corruption-severity pairs). For category-level analysis, we follow the corruption grouping used in RCP-Bench [2], aggregating the evaluated corruptions into Weather, Noise, Blur, Light, and Failure/Digital groups; we exclude temporal misalignment because it targets asynchronous feature transmission rather than the spatial corruption propagation we study. All models process camera images at an intermediate BEV feature resolution of $128\times128$ and $C=64$ channels. The input image size is $384\times512$ (resized from $600\times800$), and the perception range is $[-51.2, 51.2]$ m in both $x$ and $y$ axes with a $z$ range of $[-3.0, 1.0]$ m. All models are trained for 30 epochs. AgentMix is applied with a global augmentation gate $p_{\text{aug}}=0.5$; when activated, translation is sampled independently per agent with $p_{\text{trans}}=0.3$, and BEV-CutMix uses $p_{\text{cut}}=0.15$ (yielding an effective unconditional per-agent probability of $0.075$). Translation is bounded to $\pm3$ pixels, corresponding to approximately $2.4$ m in the BEV grid at $0.8$ m per cell, and BEV-CutMix uses $\lambda \in [0.2, 0.4]$.

Image-level interpolation inspired by MixUp [23] is applied with a mixing probability $p_{\text{mix}}=0.8$. AgentMix + MixUp combines this baseline with our AgentMix setting to isolate whether regularization benefits occur at the image input, feature statistics (RCP-Mix [2]), or the shared BEV representation level.

Table 1. **Cross-model comparison.** AgentMix: translation $\pm3$ px ($p_{\text{trans}}$=0.3) + BEV-CutMix ($\lambda \in [0.2, 0.4]$, $p_{\text{cut}}$=0.075). Best clean and corrupted AP among the compared settings per model are shown in bold.

| Setting | Model | Clean AP@0.5 ↑ | $\text{mAP}_{\text{cor}}$ ↑ |
|---|---|---|---|
| Baseline | CoAlign [12] | 69.94 | 22.46 |
| | V2X-ViT [20] | 66.74 | 26.75 |
| | Pyramid Fusion [13] | 71.75 | 24.07 |
| RCP-Mix [2] | CoAlign [12] | 69.79 | 27.92 |
| | V2X-ViT [20] | 66.77 | 24.53 |
| | Pyramid Fusion [13] | 70.76 | 26.73 |
| **AgentMix (Ours)** | CoAlign [12] | **71.04** | **33.87** |
| | V2X-ViT [20] | **71.56** | **30.54** |
| | Pyramid Fusion [13] | **71.94** | **30.23** |

## 4.2. Main Results

AgentMix is the strongest setting across all three evaluated architectures, as shown in Table 1. Compared with the vanilla baseline training, it improves $\text{mAP}_{\text{cor}}$ by +11.41

Table 2. **Configuration comparison across architectures.** Image-level interpolation inspired by MixUp [23]. *AgentMix + MixUp* applies that same image-only baseline together with our AgentMix setting.

| Model | Configuration | Clean AP@0.5 ↑ | $mAP_{cor}$ ↑ |
|---|---|---|---|
| CoAlign [12] | MixUp | **71.57** | 29.98 |
| | AgentMix + MixUp | 70.85 | 25.45 |
| | AgentMix (Ours) | 71.04 | **33.87** |
| V2X-ViT [20] | MixUp | 66.91 | 29.36 |
| | AgentMix + MixUp | 69.73 | **30.79** |
| | AgentMix (Ours) | **71.56** | 30.54 |
| Pyramid Fusion [13] | MixUp | 70.92 | 28.50 |
| | AgentMix + MixUp | 71.32 | **33.62** |
| | AgentMix (Ours) | **71.94** | 30.23 |

Table 3. **Component ablation of AgentMix on Pyramid Fusion [13].** Translation perturbation primarily stabilizes clean AP, while BEV-CutMix drives the corruption robustness gains. The full configuration achieves the best clean-robust balance.

| Configuration | Clean AP@0.5 ↑ | $mAP_{cor}$ ↑ |
|---|---|---|
| Vanilla | 71.75 | 24.07 |
| BEV-CutMix | 69.06 | **31.85** |
| Translation | **72.06** | 26.37 |
| AgentMix (Ours) | **71.94** | **30.23** |

for CoAlign (+50.8%), +3.79 for V2X-ViT (+14.2%), and +6.16 for Pyramid Fusion (+25.6%). Clean AP is maintained or improved in every case, confirming that AgentMix enhances robustness without sacrificing clean performance. RCP-Mix [2] improves CoAlign and Pyramid Fusion but hurts V2X-ViT, whereas AgentMix benefits all three. This suggests that, in this setting, image-level augmentation alone is not sufficient to match BEV-stage augmentation.

### 4.3. Ablations

The MixUp baseline is consistently stronger than the vanilla baseline but weaker than AgentMix, as shown in Table 2, indicating that input-level augmentation cannot substitute for BEV-stage augmentation. While the hybrid *AgentMix + MixUp* setting yields the highest result for Pyramid Fusion, it degrades CoAlign. Because AgentMix alone consistently improves all three architectures, it remains the most reliable cross-model default.

To isolate the components of our proposed method, we present an ablation on Pyramid Fusion (see Table 3). BEV-CutMix alone carries the majority of the robustness gain, but at a substantial clean-accuracy cost ($-1.7$ points). Applying translation only provides a mild robustness shift but actively improves clean AP. Full AgentMix successfully recovers clean AP while remaining far above the baseline on corrupted performance. This suggests that translation perturbation stabilizes collaborative fusion against minor spatial shifts, while BEV-CutMix acts as the primary regularizer against larger feature corruptions.

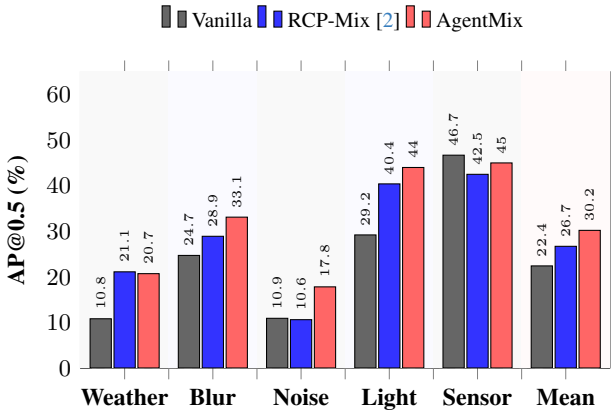


Figure 3. **Corruption-category analysis on Pyramid Fusion [13].** AgentMix is strongest on Light and Weather, with consistent gains on Blur.

### 4.4. Corruption Analysis

The gain is not uniform across corruption classes, as shown in Figure 3. The largest improvements appear in Light and Weather, followed by Blur. Translation perturbation exposes the fusion stage to small spatial inconsistencies, while BEV-CutMix discourages over-reliance on locally unreliable collaborators. Taken together, these trends support the view that, in this evaluated setting, collaborative robustness benefits from augmenting the exchanged representation rather than relying on input augmentation alone.

## 5. Conclusion

We show that, in the camera-only setting, a major robustness bottleneck is the exchanged BEV representation. AgentMix addresses this bottleneck with a simple training-time recipe: per-agent translation perturbation plus cross-agent BEV-CutMix. Across CoAlign, V2X-ViT, and Pyramid Fusion on OPV2V, AgentMix consistently improves corrupted performance over standard training, the MixUp baseline, and RCP-Mix while keeping clean performance stable or better. We hypothesize that translation discourages overfitting to exact collaborator alignment, while BEV-CutMix discourages over-reliance on any single agent, yielding a more conservative and stable fusion policy. All results are on OPV2V simulation; extending to real-world V2X datasets and multi-modal settings is an important future direction.

## References


[1] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Pro-*

*ceedings of the 4th ACM/IEEE Symposium on Edge Computing (SEC)*, pages 88–100. ACM, 2019. 1, 2

[2] Shihang Du, Sanqing Qu, Tianhang Wang, Xudong Zhang, Yunwei Zhu, Jian Mao, Fan Lu, Qiao Lin, and Guang Chen. RCP-Bench: Benchmarking robustness for collaborative perception under diverse corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11908–11918, 2025. 1, 2, 3, 4

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[4] Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. Multi-agent collaborative perception via motion-aware robust communication network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15301–15310, 2024. 1, 2

[5] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 4874–4886. Curran Associates, Inc., 2022. 2

[6] Zhe Huang, Shuo Wang, Yongcai Wang, Wanting Li, Deying Li, and Lei Wang. Roco: Robust cooperative perception by iterative object matching and pose adjustment. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pages 7867–7876, Melbourne, VIC, Australia, 2024. Association for Computing Machinery. 1, 2

[7] Philip T. G. Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P. Breckon, and Boguslaw Obara. Style augmentation: Data augmentation via style randomization. In *CVPR Workshops*, 2019. 2

[8] Baolu Li, Jinlong Li, Xinyu Liu, Runsheng Xu, Zhengzhong Tu, Jiacheng Guo, Xiaopeng Li, and Hongkai Yu. V2X-DGW: Domain generalization for multi-agent perception under adverse weather conditions. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 2

[9] Baolu Li, Zongzhe Xu, Jinlong Li, Xinyu Liu, Jianwu Fang, Xiaopeng Li, and Hongkai Yu. V2X-DG: Domain generalization for vehicle-to-everything cooperative perception. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025. First work on domain generalization in V2X cooperative perception. 2

[10] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 17394–17406. Curran Associates, Inc., 2021. 2

[11] Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*. Curran Associates, Inc., 2023. 2

[12] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3D object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4812–4818, London, UK, 2023. 1, 2, 3, 4

[13] Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Siheng Chen, and Yanfeng Wang. An extensible framework for open heterogeneous collaborative perception. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3, 4

[14] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision – ECCV 2020: 16th European Conference on Computer Vision*, pages 194–210. Springer International Publishing, 2020. 2, 3

[15] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 3

[16] Dequan Wang, Evan Shelhamer, Jiaping Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 2

[17] Sichao Wang, Ming Yuan, Chuang Zhang, Qing Xu, Lei He, and Jianqiang Wang. V2X-DGPE: Addressing domain gaps and pose errors for robust collaborative 3D object detection. *arXiv preprint arXiv:2501.02363*, 2025. Accepted for IEEE Intelligent Vehicles Symposium (IV) 2025. 2

[18] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision – ECCV 2020: 16th European Conference on Computer Vision*, pages 605–621. Springer International Publishing, 2020. 1, 2

[19] Sizhe Wei, Yuxi Wei, Yue Hu, Yifan Lu, Yiqi Zhong, Siheng Chen, and Ya Zhang. Asynchrony-robust collaborative perception via bird's eye view flow. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 16096–16109. Curran Associates, Inc., 2023. 1, 2

[20] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *Computer Vision – ECCV 2022: 17th European Conference on Computer Vision*, pages 107–124. Springer Nature Switzerland, 2022. 1, 2, 3, 4

[21] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2583–2589, 2022. 2, 3

[22] Sangdoo Yun, Dongyoon Han, Sang Joon Oh, Junsuk Elm, Geonmo Chun, and Youngjae Choe. CutMix: Regularization strategy to train strong classifiers. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032. IEEE, 2019. 2

[23] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018. 2, 3, 4

[24] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Lightweight spatial embedding for vision-based 3d occupancy prediction. *arXiv preprint arXiv:2412.05976*,

2024. Introduces BEV-CutMix augmentation for occupancy prediction. 1, 2