# Sync or Sink? The Robustness of Sensor Fusion against Temporal Misalignment

Daniel Kuhse*, Nils Hölscher*, Mario Günzel*, Harun Teper*, Georg von der Brüggen,
Jian-Jia Chen*†, Ching-Chi Lin*
*TU Dortmund University, †Lamarr Institute for Machine Learning and Artificial Intelligence
Email: {daniel.kuhse, nils.hoelscher, mario.guenzel, harun.teper, georg.von-der-brueggen,
jian-jia.chen, chingchi.lin}@tu-dortmund.de

***Abstract*—Sensor fusion is the process of combining data from multiple sensors for acquiring a more accurate and comprehensive understanding of the observed environment. However, temporal misalignments between sensors can lead to incorrect fusion results, while the temporal robustness of sensor fusion algorithms is still a relatively unexplored research topic.**

**To address this gap, we define three types of temporal robustness for sensor fusion: *reference-point-based*, *strong sample-point-based*, and *weak sample-point-based* temporal robustness. These definitions provide a framework to quantitatively evaluate the temporal robustness of sensor fusion functions. We also investigate the case where only a part of the sensors are misaligned. Furthermore, we consider potential probabilistic aspects for the proposed definitions.**

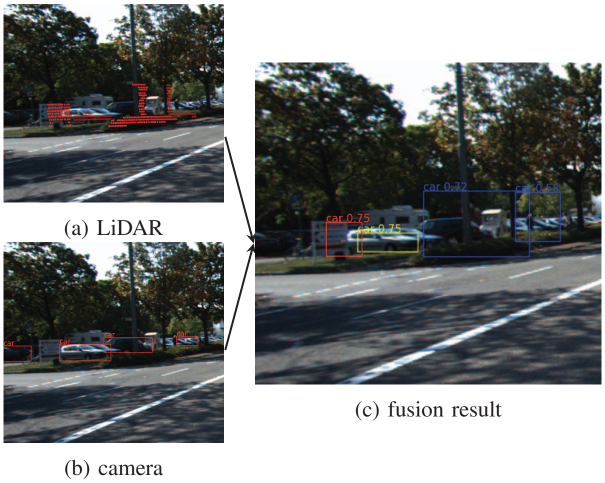
**We assess the temporal robustness of a state-of-the-art fusion method in the context of 3D object detection, where camera and LiDAR data are fused. Our empirical evaluation shows that the examined fusion methods exhibit moderate robustness against temporal misalignment of images, but are especially sensitive to LiDAR misalignment. Our findings call attention to the necessity of providing robustness guarantees for sensor fusion functions against temporal misalignment.**

***Index Terms*—Sensor Fusion, Temporal Misalignment, Robustness, Timing Analysis.**

## I. Introduction

Accurately representing the environment is crucial for many application domains, such as autonomous vehicles [7] and robotics [6], where applications rely on sensor data to perceive the environment and interact with it. Since individual sensors are designed only for specific purposes, data from multiple sensors must be fused to obtain a complete picture of the environment or system. For instance, while a single camera is not designed for capturing the 3D positions of objects, it is possible to estimate the locations of objects in a 3D space with high correctness by combining camera data with the depth information provided by a LiDAR sensor.

*Sensor fusion* is a critical process that combines data from multiple sensors to create a more accurate and comprehensive representation of the observed environment or system. Three types of sensor fusion are distinguished, depending on where in the data processing pipeline the data is fused: *early fusion*, which involves combining the sampled data directly after collection; *late fusion*, which processes the data before combining them; and *intermediate fusion*, which fuses the calculated features of the data during processing [7].



(a) LiDAR

(b) camera

(c) fusion result

Figure 1: Example of temporally-misaligned sensor fusion: (a) a point cloud from a LiDAR sensor overlaid on the temporally-misaligned image; (b) a camera image with detected bounding boxes; and (c) the fusion result, where the misclassified red bounding box is caused by the misaligned of the two sensors.

Ensuring the performance and reliability of the sensor-fusion function is essential for the system safety. Inaccuracies in the sensor-fusion function may lead to incorrect decisions being made by the system, which can have severe consequences in safety-critical domains, such as autonomous driving. For instance, inaccurate position estimates can result in collisions with other vehicles or pedestrians, thereby jeopardizing the safety of the passengers and other road users. Hence, it is crucial to employ sensor fusion techniques that are both precise and robust to minimize the risk of such incidents.

*Robustness* is a metric for estimating the correctness and reliability of a process. It refers to the property of a function being insensitive or resilient to small input perturbations. Robustness has been studied in different areas such as robustness towards noise [6], [32] or adversarial attacks [42], [51], [53]. However, robustness in the context of sensor fusion is comparatively understudied in the literature.

In this paper, we focus on the robustness of sensor fu-

2642-7346/24/$31.00 ©2024 IEEE
DOI 10.1109/RTAS61025.2024.00018

sion against temporal misalignment. *Temporal misalignment* refers to the situation where the measurements from different sensors to be fused are not aligned, i.e., the sensors are not being sampled at the exact same point in time. Temporal misalignment can be caused by a number of factors which are common in real-world scenarios, including different sampling rates, different latencies, and other time synchronization issues such as sensor clock drift or different data processing times. Figure 1 shows an example of fusing the results from a LiDAR sensor and a camera with temporal misalignment.

Temporal misalignment is generally acknowledged as problematic [1]. There are many prior works on minimizing temporal misalignment, with temporal calibration being a highly active research domain [36], [41], [43]. Methods to address temporal misalignment range from ensuring accurate and synchronized time stamps for sensors [39], to filtering out misaligned results, and employing post-processing methods like the Kalman filter to correct the misalignment.

While temporal misalignment cannot be completely eliminated, these methods are generally assumed to minimize it to a level where the impact becomes negligible [1] — implicitly assuming that the sensor-fusion function is not impacted by the remaining temporal misalignment. It would therefore be desirable to have guarantees that the sensor fusion function is robust against this remaining temporal misalignment if the performance of the sensor fusion function is safety critical.

Such a robustness guarantee for a fusion function could be applied to time-sensitive systems in two ways:

- Given a threshold for the maximum possible temporal misalignment, determine a bound for the maximum error the fusion function can produce for this temporal misalignment.
- Given a threshold for the maximum tolerable error, determine a bound for the maximum temporal misalignment that the fusion function can tolerate while ensuring that the error does not exceed the threshold.

For existing systems, either empirical measurements of the temporal misalignment or timing analysis can be used to evaluate how much the sensor fusion function is potentially affected by the possible temporal misalignment. During the development of systems, the robustness guarantee and domain specific requirements for the maximum tolerable error could be used to derive the maximum tolerable temporal misalignment, which could then be used as a constraint during the design of the system. We emphasize that robustness guarantees do not replace methods to mitigate temporal misalignment, but rather complement them, with stronger mitigations of temporal misalignment allowing for stronger robustness guarantees.

**Contributions:** In this work, we focus on the temporal robustness of sensor fusion. To the best of our knowledge, no formal definition of robustness guarantees for sensor fusion in the presence of temporal misalignment have yet been provided. Our contributions are as follows.

- In Section IV, we define three types of temporal robustness for sensor fusion: *reference-point-based*, *strong sample-based*, and *weak sample-based* temporal robustness, and discuss their implications. We also investigate the case where only some of the sensors are misaligned for *reference-point-based* temporal robustness.
- In Section V, we consider potential probabilistic aspects for our proposed definitions. That is, if the misalignment is probabilistic, we can define the robustness as the probability of the fusion result being correct.
- In Section VI, we devise a method for determining the misalignments threshold, i.e., the maximum possible temporal misalignment among sampling data, for a given system incorporating sensor fusion. To achieve this, we utilize results from end-to-end timing analysis.
- In Section VIII, we demonstrate how our defined notions of robustness can be evaluated by assessing the temporal robustness of a state-of-the-art 3D object detection method involving the fusion of camera and LiDAR data on synthetic and real-world datasets. Our empirical evaluation shows that the examined fusion methods exhibit moderate robustness against temporal misalignment in images, but especially sensitive to LiDAR misalignment.

## II. Background

We introduce the concept of robustness in Section II-A and discuss related work in Section II-B.

### A. Robustness

*Robustness* refers to the property of a function to be insensitive or resilient to small input perturbations. Specifically, a robust function or system is able to produce reliable outputs in the presence of noise and other types of perturbations. The focus of this work is the robustness of sensor fusion (where sensor data from multiple sources is combined) to temporal misalignment (that is, the fused sensors are not sampled at the same point in time). We call a function that is robust to temporal misalignment *temporally robust*.

While we study sensor fusion in the abstract, many modern sensor fusion functions use machine learning (ML) techniques — to preprocess the initial sensor data or during the fusion function itself [19], [54], [60]. Hence, robustness has been studied extensively in the context of ML. We thus orient our definitions of robustness towards the existing ML definitions. Robustness for sensor fusion functions using ML is also of special interest as neural networks have been observed to be relatively fragile [24], [48]. Yet, our definitions are not limited to ML and can be applied to any sensor fusion function.

In machine learning, robustness is usually studied in the context of classifiers, where a classifier is considered to be robust if it is able to produce the same classification result for similar inputs. For instance, a classifier for images is considered to be robust if it is able to classify an image correctly even if the image is slightly perturbed. A typical formulation for robustness of a classifier [2], [10] is as follows.

**Definition 1** (Robustness of a Classifier.)**.** Given a classifier $f$ and a distance metric $d$, the classifier $f$ is $\delta$-robust to an input $x$, if for all $x'$ such that $d(x, x') \leq \delta$, $f(x') = f(x)$.

The robustness for classifiers can be extended to general regression tasks. Contrary to classification, regression typically has a continuous and possibly multidimensional output space. Hence, instead of applying the notion of equality in the output space, a specified threshold $\epsilon$ allows a small perturbation in the output space. The extended definition is as follows.

**Definition 2** (Robustness of a Regression Function.)**.** Given a regression function $f$ and distance metrics $d_x$ and $d_y$ for the input and output space, respectively, the function $f$ is $(\epsilon, \delta)$-robust to an input $x$ if for all $x'$ such that $d_x(x, x') \leq \delta$, $d_y(f(x),\, f(x')) \leq \epsilon$ holds.

### *B. Related Work*

Prior work on sensor-fusion robustness can generally be categorized into: **(i)** finding adversarial examples to prove that the fusion function is not robust against adversarial perturbations, or **(ii)** determining methods that improve the robustness of the fusion function against adversarial perturbations.

Adversarial attacks on sensor fusion aim to find examples where small perturbations cause large errors such as misclassifications. Popular gradient-based methods formulate an optimization problem where the loss has to be maximized [6], [51], [53]. This approach allows to show that adversarial perturbations can cause large errors in the fusion result, even if a sensor is only used as an auxiliary sensor to improve the performance of another sensor [32]. To the best of our knowledge, typical adversarial attacks do not target to invoke temporal misalignment. Therefore, these methods to determine robustness against adversarial attacks seem less relevant to robustness against temporal misalignment.

Typical methods to improve the robustness of neural networks have been successfully applied to neural network based sensor fusion methods. A common approach to improve robustness against adversarial examples is to train them with adversarial examples [24], [35]. Similarly, training data can be augmented for robustness against sensor failure; examples include taking out data points to simulate missing data [38] or adding noise to the data [18]. An alternative to augmenting the data is to modify the loss function as done by Kim et al. [32] who propose a loss function that improves robustness against single-sensor noise by maximizing over all possible single channel perturbations. Architectural changes may also improve the network robustness [9], [32]. An important caveat is that these approaches often result in a trade-off between performance and robustness, as the methods that improve robustness often also reduce performance [50]. One classical, non-neural network specific approach is to preprocess the data before feeding it into the fusion function to reduce noise and other perturbations, such as interpolating to fill in missing data [40].

In recent years, verification for neural networks has become a topic of interest with robustness as a common target, having the goal of gaining guarantees instead of just empirical evidence of robustness [10], [34]. These approaches use formal methods to prove that there is no perturbation that can cause the network to misclassify an input. Examples of such methods include SMT solvers [28], [31], mixed integer programming [49], and abstract interpretation [22]. While the methods have been applied to several different domains, to the best of our knowledge they have not been applied to sensor fusion yet [29], [55], [58]. One major hurdle for verification is that current methods generally do not scale well to the large size of neural networks used in practice, which might prove especially problematic for sensor fusion. We also note that the vast majority of robustness verification work focuses on classification tasks, though it has also been applied to regression (e.g., for object detection [8]).

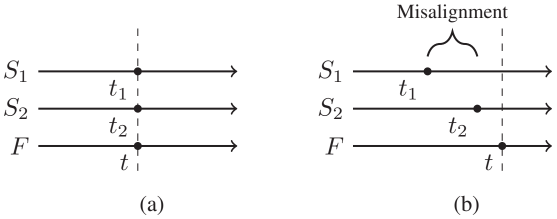


Figure 2: (a) Perfectly aligned fusion. (b) Misaligned fusion.

De Silva et al. [13] study the calibration of sensors and their robustness against resolution and spatio-temporal misalignment to address the uncertainty of sensor values, though they mainly focus on spatial misalignment and calibration. In regard to temporal misalignment, they consider a case of combining multiple LiDAR frames into an occupancy grid map. They note that a larger time window causes a larger spread of values, though they do not study this formally.

## III. System Model

We investigate a system with $n$ sensors denoted as $S_1, S_2, \ldots, S_n$. At any time $t \in \mathbb{R}$, the sensor data for $S_i$ is represented by $S_i(t)$. The sensor data may comprise either the raw data sample or an estimate derived from the data.

Sensor fusion combines data samples from multiple sensors, possibly after additional pre-processing. We formalize sensor fusion as a function $F$ that takes the sensor data as inputs and produces an output that represents the fused result. Specifically, given sensor data $S_1(t), S_2(t), \ldots, S_n(t)$ at time $t$, the result of the sensor fusion is expressed as:

$$y = F(S_1(t), S_2(t), \ldots, S_n(t)) \qquad (1)$$

More generally, a fusion might use multiple data points from a sensor that can not be characterized by a single time point. One example would be a series of images from a camera whose time stamps are not perfectly periodic. We focus on cases where the data from sensors can be viewed as having a single time point, but believe that our definitions can be extended.

It is worth noting that, in this work, we adopt an abstract approach to handling the sensor data and fusion output. Specifically, since our focus is on the robustness of sensor fusion, we do not provide a precise definition of the sensor data and fusion output. The specific nature of the data and output should be tailored to the application. For instance, in the context of object detection, the fusion output could refer to the estimated location of all the detected objects.

We take the temporal misalignment of sensors into consideration during data fusion. **Temporal misalignment** occurs when the sensor data used in a fusion function is sampled at different time points. This may be caused by a number common real-world scenarios, including different sampling rates, latency, and other time synchronization issues, such as sensor clock drift or different data processing times. Figure 2 shows the sensor fusion function with and without misalignment.

Without loss of generality, we assume that sensor $S_i$ samples its data at time $t_i$. By considering the *temporal misalignments*, the result of the sensor fusion is expressed as:

$$\hat{y} = F(S_1(t_1), S_2(t_2), \ldots, S_n(t_n)) \tag{2}$$

A loss function $L$ is applied to measure the difference between the ideal fusion result $y$ and the actual fusion result $\hat{y}$. An ideal fusion result $y$ is the expected output without temporal misalignment among sensors. That is, when all sensors are sampled at the same time point $t$.

The loss function is highly application-dependent and must be chosen appropriately, as it affects the evaluation of robustness and the comparison between methods. For instance, if the fusion result is a classification label, cross-entropy might be chosen as the loss function. If the fusion result is a continuous value, the loss function might be some distance metric, the mean squared error or mean absolute error.

We bound the temporal misalignment of a sensor-fusion function by a *misalignment threshold* $\Delta$. We denote the largest and smallest time point among $t_1, \ldots, t_n$ in one fusion operation as $t_{max}$ and $t_{min}$. $\Delta$ specifies the maximum distance allowed between $t_{max}$ and $t_{min}$, i.e., $|t_{max} - t_{min}| \leq \Delta$. Many systems with sensor fusion [30] use such thresholds to determine if the samples are too far apart to be fused.

The misalignment threshold $\Delta$ of a sensor fusion function can be determined in different ways. For instance, it can be determined based on the data age, which will be discussed in Section VI. Alternatively, it can be determined empirically by measuring the maximum possible temporal misalignment, or by choosing the maximum allowed temporal misalignment based on the error tolerance. In this work, we consider the *maximum* threshold under all relevant scenarios to provide a conservative robustness guarantee.

We assume that all data samples are reliable, i.e., not affected by sensor noise or other sources of errors. While these errors can indeed influence temporal robustness, we focus on isolated robustness against misalignment in the definitions for simplicity. Still, our subsequent definitions for temporal robustness work in the presence of other error sources.

## IV. Robustness Against Temporal Misalignment

A sensor fusion function is temporally robust if limited temporal misalignment does not significantly affect the fusion result. Due to a lack of formal study for temporal robustness in the literature, we establish formal definitions in this section.

The central idea of all definitions is to determine if, given a set of data samples $S_i(t_i)$, a fusion operation $F$ can **tolerate** temporal misalignment. To this end, we compare the loss between the ideal fusion result $y$ and the actual fusion result $\hat{y}$ with a specified error threshold $\epsilon$. Specifically,

$$L(F(S_1(t), \ldots, S_n(t)), F(S_1(t_1), \ldots, S_n(t_n))) \leq \epsilon \tag{3}$$

or (in short) $L(y, \hat{y}) \leq \epsilon$.

A key question however is how to determine the *reference time point* $t$. We consider two perspectives. *Reference-point-based temporal robustness* (see Section IV-A) assumes that the reference time point $t$ is given and that the time points $t_1, \ldots, t_n$ are variable. *Sample-point-based temporal robustness* (see Section IV-B) assumes $t_1, \ldots, t_n$ to be given and that $t$ is variable. Section IV-C discusses the relation between these definitions. Probabilistic variants are discussed in Section V.

### A. Reference-Point-Based Temporal Robustness

For *reference-point-based* temporal robustness (or short, reference-based robustness), the robustness of a sensor fusion function is measured based on reference points. The reference point $t$ is where the data would be sampled if there was no temporal misalignment. Given $t$ and a misalignment threshold $\Delta$, we assume that the misaligned sensor data $S_i(t_i)$ is sampled within a range of $\Delta$ around $t$, i.e., from $[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$.

For a given reference point $t$, a fusion function $F$ is then considered to be *reference-point-based* robust if it can tolerate all possible misalignments within the given interval. The formal definition is as follows.

**Definition 3** (Reference-Point-Based Temporal Robustness)**.** A sensor fusion function $F$ is *reference-point-based* temporally robust for reference point $t$ under the misalignment threshold $\Delta$ and error threshold $\epsilon$, if the difference between the ideal fusion result $y$ and all possible actual fusion result $\hat{y}$ is bounded by $\epsilon$ according to the loss function $L$. Formally, $L(y, \hat{y}) \leq \epsilon$ holds for all $t_1, \ldots, t_n \in [t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$.

The above definition only consider a single fixed reference point $t$. If $F$ is robust for every possible reference point $t$, we call it *global* reference-point-based temporal robustness.

*Reference-point-based temporal robustness* is temporal robustness from the perspective of the reference point $t$. It measures not only temporal robustness among sensors, but also the robustness against sensors being temporally misaligned with the reference point $t$. Therefore, even if the sensors are perfectly aligned among themselves, i.e., $t_i = t_j$ for all pairs of sensors $i$ and $j$, the fusion result may still be considered as non-robust if $t_i$ is too far away from $t$, leading to a large error in the fusion result. Sample-point-based temporal robustness, which is discussed in the next section, avoids this.

We derive a special case of reference-based temporal robustness that gives insight into what parts of the fusion function are sensitive to misalignment.

*Single-source temporal robustness* is a variant where only one of the sensors (w.l.o.g. $S_1$) is misaligned, or only one of the sensors is sensitive to temporal misalignment. Since sensors have different characteristics, they may be affected differently by temporal misalignment. For example, for 3D object detection, misalignment for the camera has a much smaller impact on the fusion result than misalignment for the LiDAR sensor, as shown in the evaluation in Section VIII. Thus, it is reasonable to consider only one sensor which is sensitive to temporal misalignment while evaluating the robustness of the sensor fusion function to reduce the complexity. Since single-source temporal robustness is a special case of reference-point-based temporal robustness, we can derive its definition from Definition 3 with $t = t_i$ for all $i = 2, \ldots, n$.

**Definition 4** (Single-Source Temporal Robustness)**.** We call the fusion function *single-source temporally robust*, if

$$\begin{aligned} &L(F(S_1(t), S_2(t) \ldots, S_n(t)), \\ &\quad F(S_1(t_1), S_2(t), \ldots, S_n(t))) \leq \epsilon \end{aligned} \tag{4}$$

holds for all $t_1 \in [t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$.

It is important to distinguish between single-source robustness for each individual sensor and overall robustness against temporal misalignment for all sensors. While single-source robustness for each sensor is desirable, it does not ensure robustness when dealing with temporal misalignment across multiple sensors simultaneously. An illustrative example is a fusion function where one sensor serves as a redundancy for another. The fusion function may exhibit robustness against temporal misalignment for each sensor individually. However, this does not guarantee robustness when both sensors are considered together. The fusion function might be arbitrarily robust for each sensor individually but not for both sensors simultaneously. This distinction emphasizes the need to evaluate the fusion function's performance in handling temporal misalignment across multiple sensors jointly, rather than focusing solely on the robustness of individual sensors. Still, single-source robustness is a prerequisite for robustness across multiple sensors and provides insights into the factors that can potentially compromise a fusion function's robustness.

*Multi-source temporal robustness* generalizes single-source temporal robustness from one misaligned sensor to a set of misaligned sensors (w.l.o.g. $S_1, \ldots, S_m$).

**Definition 5** (Multi-Source Temporal Robustness)**.** We call the fusion function *multi-source temporally robust*, if

$$\begin{aligned} &L(F(S_1(t), \ldots, S_n(t)), \\ &\quad F(S_1(t_1), \ldots, S_m(t_m), S_{m+1}(t), \ldots, S_n(t))) \leq \epsilon \end{aligned} \tag{5}$$

holds for all $t_1, \ldots, t_m \in [t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$.

Multi-source temporal robustness can be of interest for multi-sensor system whose fusion function operates on groups of sensors, such as one that has multiple stages. For example, in the context of autonomous driving, a sensor fusion function may involve a first stage fusing the data from a camera and a LiDAR sensor as part of object detection, and a second stage fusing the object detection results with IMU data. Here, studying the temporal robustness of the camera and LiDAR sensors might provide insight into which stage of the sensor fusion function is most sensitive to temporal misalignment.

### B. Sample-Point-Based Temporal Robustness

Whereas reference-point-based temporal robustness focuses on the misalignment between the reference point $t$ and the time points $t_1, \ldots, t_n$, *sample-point-based* temporal robustness (or short, sample-based robustness) considers the misalignment between the time points $t_1, \ldots, t_n$ themselves. Instead of assuming the reference point as given, we assume that the time points $t_1, \ldots, t_n$ with $t_{max} - t_{min} \leq \Delta$ are given and that the reference point $t$ is in the interval $[t_{min}, t_{max}]$.

For a given set of sampling points, we focus on two apparent questions: 1) Is there *at least one* temporally robust reference point $t$ in the interval, or (in a stricter version) 2) are *all* possible reference points in the interval temporal robust. The two resulting scenarios are called *weak* and *strong* sample-based robustness, respectively, and defined subsequently.

**Definition 6** (Weak Sample-Based Temporal Robustness)**.** A sensor-fusion function $F$ is *weak sample-based temporally robust* for a set of sample points $t_1, \ldots, t_n$ under the error threshold $\epsilon$, if there **exists** a reference time point $t$ in $[t_{\min}, t_{\max}]$ such that $L(y, \hat{y}) \leq \epsilon$.

**Definition 7** (Strong Sample-Based Temporal Robustness)**.** A sensor fusion function $F$ is *strong sample-based temporally robust* for the sample points $t_1, \ldots, t_n$ under the error threshold $\epsilon$, if $L(y, \hat{y}) \leq \epsilon$ holds for **all** time points $t$ in $[t_{\min}, t_{\max}]$.

Similar to reference-based temporal robustness, we can extend these definitions to be global for a given $\Delta$, by considering weak or strong sample-based temporal robustness for all possible sample points $t_1, \ldots, t_n$ with $t_{\max} - t_{\min} \leq \Delta$.

### C. Relation Between Robustness Definitions

The relation between weak and strong sample-based temporal robustness is clear, with the latter implying the former. However, the connection between reference- and sample-based temporal robustness is less straightforward. In this section, we explain their correlation using the misalignment threshold.

The misalignment threshold $\Delta$ holds different interpretations for reference-based and strong sample-based temporal robustness. For global sample-based temporal robustness, $\Delta$ specifies the maximum distance permitted between any two time points $t_i$ and $t_j$ for all sensors $i$ and $j$. For reference-based temporal robustness, $\Delta$ restricts the maximum permissible deviation of a time point $t_i$ from the reference point $t$.

Global $2\Delta$-reference-based temporal robustness implies $\Delta$-global strong-sample-based robustness: if $t_{\max} - t_{\min} \leq \Delta$ and $t \in [t_{\min}, t_{\max}]$ holds for $t, t_1, \ldots, t_n$, then $t_1, \ldots, t_n \in [t - \Delta, t + \Delta]$ also holds. Therefore, $F$ tolerates temporal misalignment for such $t, t_1, \ldots, t_n$, and is therefore $\Delta$-global strong-sample-based robust. However, $\Delta$-reference-point-based temporal robustness is not sufficient to imply $\Delta$-

global strong-sample-based robustness: Given the time points $t_1 = t$ and $t_2 = t_1 + \Delta$, then $t_2$ falls outside the $[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$ interval, but would be covered under sample-based robustness. No form of sample-based temporal robustness implies reference-based robustness, as reference-based temporal robustness allows $t$ to lie outside the interval of sample points.

The proposed definitions thus provide a precise framework for evaluating the temporal robustness of a sensor-fusion function. It gives different, interconnected robustness definitions with different levels of strictness, as $2\Delta$-reference-point-based temporal robustness encompasses $\Delta$-strong robustness, which, in turn, encompasses $\Delta$-weak temporal robustness.

## V. Probabilistic Robustness Guarantees

In Section IV-A and Section IV-B, we consider hard guarantees for temporal robustness. That is, the losses of the fusion results caused by temporal misalignments is bounded by the error threshold $\epsilon$ under all circumstances. Certain misalignments however might come up only very rarely and in most applications it is not necessary that faults are impossible, but rather that they are sufficiently rare. In this section we therefore consider probabilistic guarantees that take into account the probability of temporal misalignments.

In general, if the misalignment, and thus also the fusion result $\hat{y}$ itself, is probabilistic, then $L(y, \hat{y})$ is a random variable. Therefore, tolerance can be described in a probabilistic manner as $\mathbb{P}(L(y, \hat{y}) \leq \varepsilon)$ or $\mathbb{E}(L(y, \hat{y}))$, being either how likely it is to exceed the threshold or how large the expected error is. Using this we can then define probabilistic variants of our previously established robustness definitions.

For reference-based robustness, the key idea is that while we once again pick a reference point, now the $\Delta$ interval around it is not a hard constraint but rather a probability distribution.

**Definition 8** (Probabilistic Reference-Point-Based Temporal Robustness)**.** If the misalignment threshold $\Delta$ is a random variable, then we say that the fusion function $F$ is *probabilistic reference-point-based temporally robust* if

$$\mathbb{P}(\forall t_1, \ldots, t_n \in [t - \Delta, t + \Delta], L(y, \hat{y}) \leq \epsilon) \geq p \tag{6}$$

for a given probability threshold $p$.

We can extend this definition to be global, by either considering it for all possible reference time points $t$ or by considering the reference time point to also be a random variable, with the former being more strict than the latter.

Slight modifications are necessary to consider the expected loss. Instead of determining if all possible sample-point combinations in $\Delta$ are within the error threshold $\epsilon$, we consider the *worst-case* loss across all possible sample point combinations.

**Definition 9** (Probabilistic Reference-Point-Based Temporal Robustness with Bounded Expected Loss)**.** If $\Delta$ is a random variable, we say that $F$ is *probabilistic reference-point-based temporally robust with bounded expected loss*, if

$$\mathbb{E}\Big(\max_{t_1, \ldots, t_n \in [t-\Delta, t+\Delta]} L(y, \hat{y})\Big) \leq \epsilon. \tag{7}$$

For probabilistic sample-based temporal robustness the idea is that we use random variables $T_1, \ldots, T_n$ for the sample points and denote $\hat{Y} = F(S_1(T_1), \ldots, S_n(T_n))$. One practical choice for $T_1, \ldots, T_n$ is to have the distance between $T_1$ and $T_i$ follow a probability distribution over $\Delta$.

**Definition 10** (Probabilistic Sample-Point-Based Temporal Robustness)**.** We say that the fusion function $F$ is *weak probabilistic sample-point-based temporally robust*, if

$$\mathbb{P}(\exists t \in [T_{min}, T_{max}], L(y, \hat{Y}) \leq \epsilon) \geq p \tag{8}$$

and we say that the fusion function $F$ is *strong probabilistic sample-point-based temporally robust*, if

$$\mathbb{P}(\forall t \in [T_{min}, T_{max}], L(y, \hat{Y}) \leq \epsilon) \geq p, \tag{9}$$

for a given probability threshold $p$.

To bound the expected loss, the same modification as for reference-point-based temporal robustness can be applied. However, in addition to the worst-case loss across all possible reference points, the best-case is also considered, which is respectively the smallest error threshold for which strong and weak sample-based temporal robustness hold.

**Definition 11** (Probabilistic Sample-Point-Based Temporal Robustness With Bounded Expected Loss)**.** We say $F$ is *weak probabilistic sample-point-based temporally robust with bounded expected loss*, if

$$\mathbb{E}\Big(\min_{t \in [T_{min}, T_{max}]} L(y, \hat{Y})\Big) \leq \epsilon, \tag{10}$$

and we say $F$ is *strong probabilistic sample-point-based temporally robust with bounded expected loss*, if

$$\mathbb{E}\Big(\max_{t \in [T_{min}, T_{max}]} L(y, \hat{Y})\Big) \leq \epsilon. \tag{11}$$

So far, we focused on the probabilistic misalignment, but it might also be desirable to include other forms of probability, such as the probability of a certain sample occurring. For instance, in the context of autonomous driving, one might incorporate a probabilistic movement model for objects in the environment to assign lower weights to scenarios with unlikely movement when estimating robustness. Our probabilistic definitions consider the fusion function to be a random variable, so it is straightforward to incorporate other forms of probability into the definitions, but outside the scope of this paper.

The presented probabilistic notions of robustness against temporal misalignment provide a more nuanced perspective on the performance of sensor fusion functions in the presence of timing misalignments. They allow a less conservative characterization of robustness, considering the varying probabilities of different timing misalignments and the acceptable level of risk or performance degradation in the application. The application domain would determine whether an error probability bound, an expected loss bound or even a combination of both notions is more suitable. It allows for a more fine-grained characterization of robustness, which can be tailored to the specific requirements and priorities of the application.

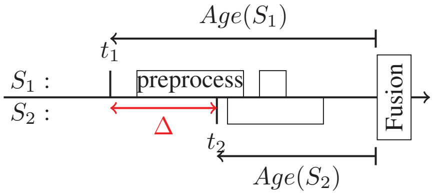


Figure 3: Quantifying the maximum misalignment $\Delta$ between sensor samples based on the data age.

## VI. Estimating the Maximum Misalignment

We established various levels of robustness against temporal misalignment, characterized by the misalignment threshold $\Delta$. This threshold specifies the maximum allowable distance between sensor samples and is determined by the underlying system. To ensure that robustness is tailored to a particular system, it becomes essential to quantify the value of $\Delta$. In this section, we determine the maximum time interval between sensor samplings utilized in a sensor fusion function.

Typically, sensors follow a sequence of steps where they sample data, undergo preprocessing through a series of tasks, and then wait for the sensor fusion function to collect the processed data. This workflow is depicted in Figure 3. In order to quantify the misalignment threshold $\Delta$, it is crucial to determine the age of the data when it is utilized by the sensor fusion function. Specifically, we need to quantify the value of $Age(S_i)$ shown in Figure 3 in order to calculate $\Delta$.

Such *data age* is specified in AUTOSAR [3], and its maximal value is analyzed in the literature [4], [5], [12], [17], [20], [21], [25], [26], [33], [44], [46]. The relation between data age and the difference between data points is discussed by Günzel et al. [27], who show that the time difference between data points is upper bounded by $\max_{i \neq j}(\delta_i - \rho_j)$, where $\delta_i$ is an upper bound and $\rho_i$ is a lower bound on the data age. As a result, given lower bound $\rho_i$ and upper bound $\delta_i$ on the data age of data from each sensor $S_i$ used in the sensor fusion, we can bound the misalignment threshold by

$$\Delta \leq \max_{i \neq j}(\delta_i - \rho_j). \tag{12}$$

## VII. Real-World Sensor Fusion Application

In this section, we take a closer look into a specific application of sensor fusion: 3D object detection, which involves integrating data from cameras and LiDAR sensors to precisely identify and locate objects in three-dimensional space. Cameras capture high-resolution RGB images, which offer detailed insights into the appearance and texture of objects. Still, they lack the capability to directly measure depth. On the other hand, LiDAR sensors employ laser beams to measure the distance between the sensor and objects, generating depth information in the form of point clouds. However, LiDAR data typically has lower spatial resolution and does not provide color or texture details. By fusing data from both modalities, the complementary strengths of camera and LiDAR data can be utilized for robust and accurate 3D object detection.

Depending on when the data is fused, there are three types of sensor fusion techniques:

- **Early Fusion**: Camera and LiDAR data are merged into a single representation, e.g., virtual point cloud [56], [57], immediately after being collected. The fused data is later used as input for subsequent object detection algorithms.
- **Intermediate Fusion**: Features are extracted from the camera and LiDAR data separately, using dedicated 2D and 3D backbone networks, and the extracted features are then combined and pass through a detection algorithm, which generates the final 3D object detection results [47].
- **Late Fusion:** Camera and LiDAR data are processed independently using separate object detection algorithms. The results are subsequently fused in a final step to generate the comprehensive 3D object detections [56].

We summarize three relevant approaches for 3D object detection using sensor fusion. One is a simple default method provided in the Autoware Universe framework [30], [37], the other is part of the popular MMDetection3D framework [11], and the third is the current state of the art for sensor-fusion-based 3D object detection on the KITTI dataset [23]. We have evaluated the temporal robustness of the last one.

### *1) ROI Projection Fusion:*

*ROI-Cluster fusion* and *ROI-Detected-Object fusion* are two non-deep-learning projection-based fusion methods used in the perception module of the ROS 2-based Autoware Universe framework [30], [37]. Both are **late-fusion** methods utilized for combining 2D bounding boxes (Regions of Interest) with clusters or 3D bounding boxes, respectively. The fusion function involves projecting these bounding boxes into the image plane and evaluating the overlap to determine the validity and class of the clusters or bounding boxes. Euclidean clustering is used on the voxelized point cloud to facilitate the fusion function. For 2D object detection, YOLOv7 [52], the latest iteration of the well-known YOLO object detection network, is employed, while SECOND [59] is applied for 3D object detection.

### *2) MVX-Net:*

The *MVX-Net* is a multi-modal deep-fusion detection network [47] that is integrated into the MMDetection3D framework [11], which is part of the popular OpenMMLab project. MVX-Net combines the Faster-RCNN [45] as image-detection backbone with the VoxelNet [61] as 3D-detection backbone. MVX-Net is an **early-fusion** approach, where the points from the point cloud are projected onto the image plane and appended with the data from the 2D backbone. This fused data is then passed into the VoxelNet backbone for comprehensive 3D object detection.

### *3) VirConv:*

*VirConv-L* and *VirConv-T* [56] are **early-** and **late-fusion** detection networks, respectively, adapted from Voxel-RCNN [14] utilizing the VirConvNet backbone. They operate on virtual points, which are pixels with depth information, resulting in a dense-data representation. To address this density, a technique called *virtual sparse convolution* is applied, discarding nearby points to reduce redundancy. This process enhances performance and robustness against data noise. The two variants, *VirConv-L* and *VirConv-T*, differ in their data fusion approaches. *VirConv-L* is an **early-**

**fusion** method and combines the point cloud and virtual points through projection, followed by feature extraction and object detection using the VirConvNet architecture. *VirConv-T* adopts a **late-fusion** style, where the final detections are fused, using VoxelNet as the 3D backbone and VirConvNet as the 2D backbone. This approach leverages the strengths of both networks to achieve accurate and comprehensive object detection in three-dimensional space.

## VIII. Evaluation

In this section, we conducted a comprehensive assessment of temporal robustness using synthetic and real-world datasets. First, we introduce the datasets in Section VIII-A. To emphasize the influence of temporal misalignment on the performance of sensor fusion models, we performed a preliminary evaluation in Section VIII-B. In Section VIII-C, we outline the methods we applied for evaluating the robustness of the fusion models. Subsequently, we present the fusion results under different temporal robustness metrics. To be more specific, we examined single-source temporal robustness in Section VIII-D, evaluated the strong and weak sample-based temporal robustness in Section VIII-E, and showcased the results of reference-point-based temporal robustness in Section VIII-F We also explored probabilistic temporal robustness in Section VIII-G. In Section VIII-H, we summarize the results and discuss the practical implications for real-world applications.

### A. Datasets and Fusion Models

We evaluated the robustness of sensor fusion against temporal misalignment using two datasets: *KITTI* and a *KITTI-CARLA* adaption. *KITTI* [23] is a widely-used benchmark for 3D object detection, providing a large dataset of real-world data. The dataset consists of samples captured by a car driving through a city, with a sampling rate of 10Hz. Each sample includes a LiDAR point cloud, a camera image, and a set of 3D bounding boxes. By taking data from different sensors in adjacent frames, we can simulate temporal misalignment.

Still, the KITTI dataset has limitations when evaluating the robustness of sensor fusion against temporal misalignment. With a sampling rate of 10Hz, the granularity of temporal misalignment is limited to a minimum misalignment of 100 milliseconds. This large time gap does not represent the range of temporal misalignments that occur in real-world scenarios.

Hence, we introduce an adapted *KITTI-CARLA* dataset [15], which is a synthetic dataset generated using the CARLA simulator [16]. The *KITTI-CARLA* dataset has a similar sensor configuration as the original KITTI dataset, but allows us to increase the sampling rate to 100Hz. This higher sampling rate allows a finer granularity simulation. The dataset consists of 35000 frames. To increase the variety of the dataset, we restart the simulation every 750 frames, using the first 500 as a warm-up period and only recording the last 250.

It is important to note that the CARLA simulator is not a perfect representation of the real world. While CARLA provides high-fidelity graphics, they are not photorealistic, and the scene complexity in CARLA, compared to KITTI, is generally lower. Additionally, the sensors in CARLA are not subject to the same level of noise as real-world sensors. Therefore, we utilized the synthetic dataset primarily to study the relative performance of different fusion methods under different temporal misalignment scenarios. The results obtained from the synthetic dataset should be considered more optimistic compared to real-world sensor inputs.

| Model | Misalignment | Easy | Med. | Hard |
|---|---|---|---|---|
| **VirConv-L** | None | 89.94% | 86.42% | 85.58% |
| | Image | 89.80% | 85.52% | 78.84% |
| | Lidar | 1.25% | 1.70% | 1.50% |
| **VirConv-T** | None | 90.08% | 87.93% | 86.74% |
| | Image | 89.83% | 87.24% | 84.86% |
| | Lidar | 1.76% | 1.79% | 1.44% |

Table I: Average Precision of 3D detection on KITTI with different temporal misalignments

| Model | Misalignment | Easy | Med. | Hard |
|---|---|---|---|---|
| **VirConv-L** | None | 95.34% | 88.13% | 85.36% |
| | Image | 93.08% | 85.92% | 83.17% |
| | Lidar | 0.76% | 1.10% | 0.90% |
| **VirConv-T** | None | 94.93% | 90.41% | 88.12% |
| | Image | 92.50% | 88.22% | 85.81% |
| | Lidar | 0.94% | 1.10% | 0.90% |

Table II: Average Precision of 3D detection for objects within $40m$ on KITTI with different temporal misalignments

We focus on the results obtained from VirConv-T as it represents the state-of-the-art fusion model for sensor-fusion-based object detection in our evaluations. In a preliminary evaluation, we measured the temporal robustness for all the models introduced in Section VII. The results suggest that these models exhibit similar trends in terms of robustness.

### B. Impact of Temporal Misalignment

Most existing sensor-fusion research focuses on improving the accuracy of fusion outputs. However, accuracy alone does not guarantee robustness of a fusion model. In this section, we demonstrate the impact of temporal misalignment on the performance of the fusion models.

To evaluate the performance of VirConv-L and VirConv-T under single-source temporal misalignment, we utilized the KITTI detection benchmark. Specifically, we considered three scenarios: no misalignment (None), only the camera shifted by 100ms to cause a misalignment (Image), and only the LiDAR shifted by 100ms to cause a misalignment (LiDAR). We evaluated the Average Precision (AP) for detecting all objects and the AP for objects within a 40m range. The input cases were categorized into three groups: *Easy*, *Medium*, and *Hard*, as defined in the KITTI 3D detection benchmark.

Table I and Table II show the average precision for detecting all objects and detecting objects within 40m, respectively. Even without temporal misalignments, the average precision of both models is not perfect (100%) and decreases as the scenarios become more challenging. Notably, the accuracy for detecting objects within 40m is generally higher, since objects in closer proximity to the sensor are easier to detect.

The results reveal that temporal misalignment from different sensors have varying effects on accuracy. For instance, misalignment in LiDAR significantly decreases accuracy, while misalignment in the camera has only a minor impact. This indicates that these models are more sensitive to misalignments in LiDAR data compared to camera data. Additionally, we hypothesize that objects closer to the sensor are slightly more susceptible to temporal misalignment due to their larger relative movement on the image plane.

The preliminary results highlight the importance of considering the robustness of fusion models under temporal misalignment, which we subsequently evaluate more comprehensively.

### C. Evaluation Metrics

To evaluate the robustness of fusion models against temporal misalignments, we employed loss functions to quantify the impact on the performance of a fusion model. We consider three metrics for evaluating the loss, considering their interpretability for object-detection tasks: *F1-Score*, the *Average Intersection-over-Union (IoU)*, and the *Euclidean distance*.

- **F1-Score** is a balanced measure of the correct object detections and missed detections. It is the harmonic mean of precision and recall. Precision is the ratio of true positives to the total number of predictions, while recall is the ratio of true positives to all actual reference objects. The F1 score ranges from $0$ (worst) to $1$ (best) with $1$ denoting perfect precision and recall. Typically, F1 score is calculated across the entire dataset, but adapting it to detections for a single pair of aligned and misaligned samples does not significantly alter the meaning.
- **Intersection-over-Union (IoU)** offers insights into the overlap between predicted and reference bounding boxes. It calculates the ratio of the overlapping (intersecting) area of the two bounding boxes to the area of their union. Similar to the F1-Score, a value of $1$ indicates complete overlap, while $0$ represents no overlap. We averaged the IoU over each pair of predicted and reference bounding box. A detection score threshold of $0.5$ is used to filter out low-confidence detections.
- **Euclidean distance** is the straight-line distance between two points. In our case, the center of the predicted bounding box and the center of the reference bounding box. A larger distance indicates worse performance. We use the median Euclidean distance over all pairs.

Note that F1-Score and IoU cannot be used as loss functions directly. Therefore, we converted them into loss functions by subtracting them from 1. As these metrics are typically maximized in the literature, we report them as such instead of as losses. That is, the greater the score is, the smaller the loss is, thus indicating better robustness.

To evaluate the robustness of a fusion method for a single case under a specific robustness definition, we determine the loss that corresponds to the smallest threshold $\epsilon$ for which the method would be considered robust. We evaluate global robustness for the Euclidean distance, which then corresponds to the maximum distance a detection is off by.

| $\Delta$ | **F1** | **IoU** | **Euclid. (m)** | **max. Euclid. (m)** |
|---|---|---|---|---|
| | | **Image misaligned** | | |
| 10ms | 0.93 ± 0.17 | 0.95 ± 0.12 | 0.03 ± 0.03 | 0.25 |
| 20ms | 0.90 ± 0.19 | 0.94 ± 0.15 | 0.04 ± 0.04 | 0.27 |
| 30ms | 0.88 ± 0.21 | 0.93 ± 0.16 | 0.05 ± 0.04 | 0.30 |
| 40ms | 0.86 ± 0.22 | 0.92 ± 0.17 | 0.05 ± 0.04 | 0.31 |
| 50ms | 0.85 ± 0.23 | 0.92 ± 0.18 | 0.05 ± 0.05 | 0.35 |
| 60ms | 0.84 ± 0.24 | 0.91 ± 0.19 | 0.06 ± 0.05 | 0.35 |
| | | **LiDAR misaligned** | | |
| 10ms | 0.88 ± 0.22 | 0.90 ± 0.17 | 0.08 ± 0.06 | 0.37 |
| 20ms | 0.82 ± 0.26 | 0.86 ± 0.20 | 0.13 ± 0.09 | 0.49 |
| 30ms | 0.78 ± 0.28 | 0.83 ± 0.21 | 0.18 ± 0.13 | 0.59 |
| 40ms | 0.75 ± 0.29 | 0.80 ± 0.22 | 0.23 ± 0.16 | 0.70 |
| 50ms | 0.73 ± 0.30 | 0.77 ± 0.22 | 0.28 ± 0.20 | 0.86 |
| 60ms | 0.71 ± 0.30 | 0.74 ± 0.23 | 0.33 ± 0.23 | 0.96 |

Table III: Single-source temporal robustness of VirConv-T on the synthetic KITTI-CARLA dataset

The fusion method is not robust globally for F1 score and IoU under any definition for all misalignment thresholds $\Delta$ we can test, with the error threshold being $1$, as there is always at least one case where the detection completely fails. To not just look at the corner case and give more insight into the robustness of each method, we also report the *average* loss and its standard deviation for each loss metric.

### D. Single-Source Temporal Robustness

We evaluated the robustness of fusion methods under single-source temporal misalignment for a discrete time system. We conducted two sets of evaluations. First, we used VirConv-T as the fusion method on the *KITTI-CARLA* dataset, and measured the robustness under different misalignment thresholds $\Delta$. We then compared the results of using VirConv-T and VirConv-L as the fusion method on the *KITTI* dataset. For both sets of evaluations, we considered two scenarios — misalignment in the image data and misalignment in the LiDAR data — and report the maximum or average losses.

Table III shows the results for each metric under single-source misalignment on the synthetic KITTI-CARLA dataset, where either the image data or the LiDAR data was progressively misaligned from 10ms up to 60ms. When only the image data is misaligned, the F1 score remains relatively high, above $0.84$, showing only a small change across misalignment thresholds. Similarly, the IoU remains high and mean and maximum Euclidean distance remain low. These findings indicate that the fusion method exhibits a certain level of robustness in the presence of image data misalignment, as it maintains a relatively accurate prediction with minimal changes to the bounding box. However, there may be some false positives and false negatives compared to the reference result.

In contrast, when only the LiDAR data is misaligned, a larger impact is observed on the F1 score. The Euclidean distance shows a substantial increase for misaligned LiDAR, and these values notably escalate with the degree of misalignment. Note that even for small misalignments, the maximum Euclidean distance reaches approximately $0.37$m rising up to $0.96$m at 60ms. These results indicate that the fusion method is relatively less robust to misaligned LiDAR data, as it leads

| $\Delta$ | F1 | IoU | Euclid. (m) | max. Euclid. (m) |
|---|---|---|---|---|
| | | **VirConv-L** | | |
| Image 100ms | 0.72 | 0.83 ± 0.27 | 0.09 ± 0.06 | 0.75 |
| LiDAR 100ms | 0.44 | 0.50 ± 0.31 | 0.50 ± 0.39 | 1.66 |
| | | **VirConv-T** | | |
| Image 100ms | 0.82 | 0.88 ± 0.27 | 0.05 ± 0.04 | 0.63 |
| LiDAR 100ms | 0.44 | 0.52 ± 0.31 | 0.50 ± 0.39 | 1.70 |

Table IV: Single-source temporal robustness of VirConv-L and VirConv-T on the KITTI dataset

| $\Delta$ | F1 | IoU | Euclid. (m) | max. Euclid. (m) |
|---|---|---|---|---|
| 10ms | 0.85 ± 0.23 | 0.89 ± 0.19 | 0.08 ± 0.06 | 0.52 |
| 20ms | 0.83 ± 0.25 | 0.87 ± 0.20 | 0.11 ± 0.08 | 0.70 |
| 30ms | 0.81 ± 0.26 | 0.85 ± 0.21 | 0.13 ± 0.10 | 0.89 |
| 40ms | 0.79 ± 0.27 | 0.84 ± 0.22 | 0.16 ± 0.13 | 0.99 |
| 50ms | 0.77 ± 0.28 | 0.82 ± 0.22 | 0.18 ± 0.15 | 1.07 |
| 60ms | 0.76 ± 0.28 | 0.81 ± 0.22 | 0.21 ± 0.18 | 1.11 |

Table V: Strong sample-based temporal robustness metrics of VirConv-T on the KITTI-CARLA dataset

| $\Delta$ | F1 | IoU | Euclid. (m) | max. Euclid. (m) |
|---|---|---|---|---|
| 10ms | 0.93 ± 0.19 | 0.96 ± 0.10 | 0.03 ± 0.03 | 0.25 |
| 20ms | 0.93 ± 0.21 | 0.96 ± 0.10 | 0.03 ± 0.03 | 0.30 |
| 30ms | 0.93 ± 0.21 | 0.96 ± 0.09 | 0.03 ± 0.03 | 0.35 |
| 40ms | 0.93 ± 0.22 | 0.96 ± 0.09 | 0.03 ± 0.03 | 0.36 |
| 50ms | 0.92 ± 0.23 | 0.96 ± 0.09 | 0.03 ± 0.03 | 0.36 |
| 60ms | 0.92 ± 0.23 | 0.96 ± 0.09 | 0.03 ± 0.03 | 0.36 |

Table VI: Weak sample-based temporal robustness metrics of VirConv-T on the KITTI-CARLA dataset

| $\Delta$ | F1 | IoU | Euclid. | max. Euclid. |
|---|---|---|---|---|
| 20ms | 0.75 ± 0.28 | 0.85 ± 0.22 | 0.12 ± 0.08 | 0.61 |
| 40ms | 0.66 ± 0.30 | 0.79 ± 0.25 | 0.18 ± 0.11 | 0.78 |
| 60ms | 0.60 ± 0.31 | 0.75 ± 0.27 | 0.24 ± 0.15 | 0.85 |

Table VII: Reference-point-based temporal robustness metrics of VirConv-T on the KITTI-CARLA dataset

to larger increases in both false positives and false negatives, as well as notable deviations in the predicted bounding box.

Table IV shows the results for VirConv-L and VirConv-T under single-source misalignment on the KITTI dataset. We observe a similar trend as for the KITTI-CARLA dataset, indicating a higher sensitivity to LiDAR-data temporal misalignment. However, VirConv-T, which we evaluate on both, exhibits lower robustness on the KITTI dataset compared to the KITTI-CARLA dataset. For the KITTI dataset, VirConv-T shows slightly better robustness against image temporal misalignment than VirConv-L. This suggests that late fusion may be more resilient to image temporal misalignment than early fusion. However, for LiDAR temporal misalignment, both methods exhibit equal sensitivity.

In summary, our single-source robustness results indicate that the fusion methods demonstrate moderate robustness against image temporal misalignment but lack robustness against LiDAR temporal misalignment.

### *E. Strong and Weak Sample-Based Temporal Robustness*

We evaluated strong and weak sample-based temporal robustness of two sensors under different maximum latencies for a discrete time system. We leverage the KITTI-CARLA dataset, which provides the necessary sample granularity.

To assess both types of sample-based temporal robustness, we consider all possible pairs of samples in the dataset and calculate the evaluation metrics for each potential reference point within the specified time interval. For strong and weak sample-based temporal robustness, we determine the loss associated with the worst and best reference point, respectively.

Table V and Table VI shows the average losses of VirConv-T on the synthetic KITTI-CARLA dataset under strong and weak sample-based temporal robustness evaluations, respectively. We observe that the strong sample-based temporal robustness of VirConv-T exhibits limited resilience and with larger misalignments, the losses in F1 score, IoU, and Euclidean distance significantly increase. We conclude that the fusion method is not temporally robust under the strong sample-based temporal robustness definition.

Considering the weak sample-based temporal robustness results from Table VI, the localization aspect exhibits a high level of robustness against temporal misalignment, with an average distance of a negligible 3cm and a maximum distance of less than 36cm across all $\Delta$ values. The IoU shows that the bounding boxes are generally rather well aligned with deviations of less than $5\%$, though the standard deviation indicates that there are sporadic outliers. The F1 score indicates occasional discrepancies between the detected objects in the fused result and all reference points, which could be problematic when high accuracy is required.

### *F. Reference-Point-Based robustness*

We evaluated the reference-point-based temporal robustness of VirConv-T for a discrete time system. Due to limited granularity, we used multiples of 20ms for $\Delta$. Table VII shows the results for the KITTI-CARLA dataset. Similar to strong and weak sample-based temporal robustness, we calculated averages to quantify the impact of temporal misalignment. We observe that VirConv-T is even less robust when considering reference-point-based temporal robustness rather than strong sample-based temporal robustness. We notice a harsher drop in the F1 score and IoU as $\Delta$ increases. The localization error is slightly worse, though the maximum error is actually smaller, potentially because the maximum time difference between reference point and sample points is only half the interval.

### *G. Probabilistic robustness*

We evaluate the probabilistic variants of strong sample-based, weak sample-based, and reference-point-based temporal robustness defined in Section V for a discrete time system. Due to the need to explore misalignment variability and the limited sample granularity available in the KITTI dataset, our evaluation is conducted only on the KITTI-CARLA dataset, which provides the necessary sample granularity.

Recall that the probabilistic robustness definitions require a distribution of misalignments to be specified. We considered two different distributions of misalignments in our analysis.

| Type | Distribution | F1 | IoU | Euclid. |
|---|---|---|---|---|
| Weak | Long-tailed | 0.95 ± 0.17 | 0.96 ± 0.09 | 0.03 ± 0.03 |
|  | Uniform | 0.95 ± 0.19 | 0.96 ± 0.08 | 0.03 ± 0.03 |
| Strong | Long-tailed | 0.87 ± 0.22 | 0.90 ± 0.17 | 0.09 ± 0.08 |
|  | Uniform | 0.81 ± 0.25 | 0.85 ± 0.20 | 0.16 ± 0.14 |
| Reference | Long-tailed | 0.74 ± 0.29 | 0.84 ± 0.23 | 0.13 ± 0.10 |
|  | Uniform | 0.67 ± 0.31 | 0.80 ± 0.25 | 0.18 ± 0.13 |

Table VIII: Probabilistic robustness with expected loss metrics of VirConv-T on the KITTI-CARLA dataset

First, a *discrete uniform distribution* where misalignments of 10ms, 20ms, 30ms, 40ms, 50ms and 60ms are equally prevalent. The second distribution is a *long-tailed distribution* with misalignments of 10ms, 20ms, 30ms, 50ms and 60ms having probabilities of occurrence at 80%, 10%, 5%, 3%, 1% and 1%, respectively. We believe the second distribution to be a more realistic reflection of real-world scenarios according to the *Pareto Principle*, or 80-20 rule. Still, we have included the uniform distribution for comparative analysis.

*1) Probabilistic Strong and Weak Sample-Based Temporal Robustness:* First, we evaluated probabilistic robustness with an expected loss for weak and strong sample-based temporal robustness. We sampled 35000 pairs of sensor samples using the corresponding distribution. For each sample pair, we evaluated the maximum and minimum loss for all possible reference points to determine the smallest, valid error bound for strong and weak sample-based temporal robustness, respectively.

Table VIII shows the probabilistic robustness results under the two specified distributions. We observe that while the expected errors for weak sample-based temporal robustness remain relatively consistent across both distributions, the expected errors for strong sample-based temporal robustness are significantly higher when using the uniform distribution.

This finding aligns with our observations in Section VIII-E, where weak sample-based temporal robustness tends to remain stable across various $\Delta$ values, while strong sample-based temporal robustness experiences a more substantial decline.

We also evaluated probabilistic temporal robustness without an expected loss, using the same set of sample pairs as for expected loss. However, instead of determining the average loss, we determine the $95\%$ quantile and $85\%$ quantile for each respective metric. The results can be found in table IX.

| Type | Distribution | p | F1 | IoU | Euclid. |
|---|---|---|---|---|---|
| Weak | Long-tailed | 0.85 | 1.00 | 0.95 | 0.05 |
|  |  | 0.95 | 0.67 | 0.93 | 0.08 |
|  | Uniform | 0.85 | 1.00 | 0.95 | 0.05 |
|  |  | 0.95 | 0.67 | 0.93 | 0.08 |
| Strong | Long-tailed | 0.85 | 0.67 | 0.88 | 0.16 |
|  |  | 0.95 | 0.50 | 0.79 | 0.25 |
|  | Uniform | 0.85 | 0.67 | 0.80 | 0.29 |
|  |  | 0.95 | 0.00 | 0.56 | 0.44 |
| Reference | Long-tailed | 0.85 | 0.50 | 0.84 | 0.21 |
|  |  | 0.95 | 0.00 | 0.00 | 0.30 |
|  | Uniform | 0.85 | 0.40 | 0.79 | 0.31 |
|  |  | 0.95 | 0.00 | 0.00 | 0.41 |

Table IX: Probabilistic robustness with expected loss metrics of VirConv-T on the KITTI-CARLA dataset

We observe that under weak sample-based robustness, the errors are in general quite small for both distributions and for both $p = 0.85$ and $p = 0.95$. However, there is a noteworthy exception in the F1 score, which drops to $0.67$ for $p = 0.95$ under both distributions, indicating rare cases of false positives and false negatives even under weak sample-based robustness. For strong sample-based temporal robustness, the long-tailed distribution exhibits significantly lower errors, already at $p = 0.85$ and even more noticeable for $p = 0.95$. We furthermore note that in the case of $p = 0.95$, the F1 score drops to $0.00$ for the uniform distribution, indicating that there is a $5\%$ chance that there is a reference point with no overlap in detections.

*2) Probabilistic Reference-Point-Based Robustness:* We evaluated both variants of probabilistic reference-point-based robustness, using a similar approach as for probabilistic sample-based robustness. We look only at the $\Delta$ values of 20ms, 40ms, and 60ms, with a uniform distribution and also probabilities of 0.91, 0.8, and 0.02, respectively, to match the long-tailed distribution in the previous section. The results are shown in Table VIII and Table IX. As one would expect, we observe that the uniform distribution produces similar values to non-probabilistic reference-based temporal robustness for $\Delta = 20$ms, while the long-tailed distribution produces values similar to the non-probabilistic results for 10ms. We generally observe the same trend as for strong sample-based temporal robustness, but with slightly larger errors overall.

### H. Summary and Discussion

The examined fusion methods exhibit moderate single-source temporal robustness for images, but a lack of robustness for the LiDAR sensor. Under both strong sample-based and reference-point-based temporal robustness the examined methods are not temporally robust. Under weak sample-based robustness, the localization (Euclidean distance) aspect and overlap (IoU) is highly temporally robust, while the detection (F1-Score) aspect is a bit less robust. We caveat this by noting that we only determined this for the synthesized KITTI-CARLA dataset, which might give overly optimistic results.

We have not conducted a separate analysis of multi-source robustness in our evaluation. For a setup with two sensors, it is equivalent to either single-source robustness (if one sensor remains fixed to the reference time point) or reference-point-based temporal robustness (if both sensors have the freedom to move independently). For more than two sensors, the same scheme could be applied to evaluate multi-source robustness by keeping a subset of the sensors fixed to the reference time point and shifting the remaining sensors.

The results of such robustness assessments serve as valuable tools for estimating potential errors in systems with temporal misalignments. For example, consider the sensor fusion methods used in Autoware [30], which rely on timestamps to fuse temporally proximate data. In their default configuration, a maximum time difference of 50ms is set as the threshold, representing the maximum allowable misalignment for data fusion. Our evaluation reveals that even with a misalignment under 50ms, such sensor fusion methods can still yield significant errors and may not be considered robust enough.

Another practical application would be that given the maximum tolerable error, the evaluation results can help to determine the maximum misalignment that the system can tolerate. This information could be utilized to establish precise thresholds in timestamp-based systems, or to guide the design of systems under a latency-based analysis.

## IX. Conclusion

We provide a comprehensive investigation into the temporal robustness of sensor-fusion algorithms — a research area relatively unexplored in the literature, with a lack of formal definitions. We propose three definitions of temporal robustness for sensor fusion: reference-point-based, strong sample-based, and weak sample-based temporal robustness. They offer different levels of restrictions, capturing distinct intuitive notions of temporal robustness, enabling a quantitative evaluation of the robustness of sensor fusion functions.

In addition, we explore the special case of single-source temporal robustness to assess the relative significance of alignment for each sensor. Furthermore, we develop probabilistic variants that account for the probabilistic distributions of sensor deviation. These variants are designed to address scenarios where certain types of temporal misalignments are more likely to occur than others. By considering these probabilistic factors, we aim to provide robustness measures that are tailored to specific misalignment patterns, ensuring that our evaluations are not overly conservative in scenarios where general robustness measures may fall short.

We evaluated the robustness of several fusion methods against temporal misalignment in the context of 3D object detection, specifically focusing on scenarios involving camera and LiDAR data. We evaluated a real-world dataset, KITTI, and a synthetic dataset generated using the CARLA simulator to investigate smaller temporal misalignments. The results indicate that none of the examined fusion methods are completely temporally robust under any definition, though under weak robustness the localization is highly robust. Furthermore, we find that misalignment of LiDAR data has a larger impact compared to temporal misalignment of the image sensor.

For this work, we focused on the case that all data from a sensor can be characterized by a single time point $t_i$, and the fusion model in Equations 1 and 2 is based on this assumption. However, some fusion functions use data from the same sensor at different timestamps. While some time series data can be characterized only by its latest time stamp, this does not hold in general. A generalization of our approach to fully cover fusions like Kalman filters and SLAM algorithms will be part of future work.

Our work aims to provide a fresh perspective on temporal misalignment and a first step to go beyond solely minimizing its impact and instead offering robustness guarantees in the face of it. We believe that our findings and proposed definitions pave the way for the development of more resilient and dependable fusion algorithms.

## Acknowledgements

This result is part of a project (PropRT) that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 865170). This work has received funding by the German Federal Ministry of Education and Research (BMBF) in the course of the 6GEM research hub under grant number 16KISK038.

## References

[1] M. Aeberhard and N. Kaempchen. High-level sensor data fusion architecture for vehicle surround environment perception. In *Proc. 8th Int. Workshop Intell. Transp*, volume 665, 2011.

[2] A. Albarghouthi et al. Introduction to neural network verification. *Foundations and Trends® in Programming Languages*, 7(1–2):1–157, 2021.

[3] AUTOSAR. Specification of timing extensions (AUTOSAR CP R21-11). https://www.autosar.org/fileadmin/user_upload/standards/classic/21-11/AUTOSAR_TPS_TimingExtensions.pdf, 2021. Accessed: 2022-10-18.

[4] M. Becker, D. Dasari, S. Mubeen, M. Behnam, and T. Nolte. Mechaniser-a timing analysis and synthesis tool for multi-rate effect chains with job-level dependencies. In *Workshop on Analysis Tools and Methodologies for Embedded and Real-time Systems (WATERS)*, 2016.

[5] M. Becker, D. Dasari, S. Mubeen, M. Behnam, and T. Nolte. Synthesizing job-level dependencies for automotive multi-rate effect chains. In *International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pages 159–169, 2016.

[6] M. Bednarek, P. Kicki, and K. Walas. On robustness of multi-modal fusion—robotics perspective. *Electronics*, 9(7):1152, 2020.

[7] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021.

[8] P.-y. Chiang, M. Curry, A. Abdelkader, A. Kumar, J. Dickerson, and T. Goldstein. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33:1275–1286, 2020.

[9] J.-H. Choi and J.-S. Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019.

[10] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[11] M. Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[12] A. Davare, Q. Zhu, M. D. Natale, C. Pinello, S. Kanajan, and A. L. Sangiovanni-Vincentelli. Period optimization for hard real-time distributed automotive systems. In *Design Automation Conference, DAC*, pages 278–283, 2007.

[13] V. De Silva, J. Roche, and A. Kondoz. Robust fusion of lidar and wide-angle camera data for autonomous mobile robots. *Sensors*, 18(8):2730, 2018.

[14] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021.

[15] J.-E. Deschaud. Kitti-carla: a kitti-like dataset generated by carla simulator. *arXiv preprint arXiv:2109.00892*, 2021.

[16] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[17] M. Dürr, G. von der Brüggen, K.-H. Chen, and J.-J. Chen. End-to-end timing analysis of sporadic cause-effect chains in distributed systems. *ACM Trans. Embedded Comput. Syst. (Special Issue for CASES)*, 18(5s):58:1–58:24, 2019.

[18] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[19] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, 20(15):4220, 2020.

[20] N. Feiertag, K. Richter, J. Nordlander, and J. Jonsson. A compositional framework for end-to-end path delay calculation of automotive systems under different path semantics. In *Workshop on Compositional Theory and Technology for Real-Time Embedded Systems*, 2009.

[21] J. Forget, F. Boniol, and C. Pagetti. Verifying end-to-end real-time constraints on multi-periodic models. In *ETFA*, pages 1–8, 2017.

[22] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2018.

[23] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[25] M. Günzel, K.-H. Chen, N. Ueter, G. v. der Brüggen, M. Dürr, and J.-J. Chen. Compositional timing analysis of asynchronized distributed cause-effect chains. *ACM Transactions on Embedded Computing Systems*, 2023.

[26] M. Günzel, K.-H. Chen, N. Ueter, G. von der Brüggen, M. Dürr, and J.-J. Chen. Timing analysis of asynchronized distributed cause-effect chains. In *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 40–52, 2021.

[27] M. Günzel, N. Ueter, K.-H. Chen, G. von der Brüggen, J. Shi, and J.-J. Chen. End-to-end processing chain analysis. In *RTSS Industrial Challenge*, 2021.

[28] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, pages 3–29. Springer, 2017.

[29] R. Jia, A. Raghunathan, K. Göksel, and P. Liang. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*, 2019.

[30] S. Kato, S. Tokunaga, Y. Maruyama, S. Maeda, M. Hirabayashi, Y. Kitsukawa, A. Monrroy, T. Ando, Y. Fujii, and T. Azumi. Autoware on board: Enabling autonomous vehicles with embedded systems. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*, pages 287–296. IEEE, 2018.

[31] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, pages 97–117. Springer, 2017.

[32] T. Kim and J. Ghosh. On single source robustness in deep fusion models. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] T. Kloda, A. Bertout, and Y. Sorel. Latency analysis for data chains of real-time periodic tasks. In *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, pages 360–367, 2018.

[34] L. Li, T. Xie, and B. Li. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[36] E. Mair, M. Fleps, M. Suppa, and D. Burschka. Spatio-temporal initialization for imu to camera registration. In *2011 IEEE International Conference on Robotics and Biomimetics*, pages 557–564. IEEE, 2011.

[37] Y. Maruyama, S. Kato, and T. Azumi. Exploring the performance of ros2. In *Proceedings of the 13th International Conference on Embedded Software*, pages 1–10, 2016.

[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[39] J.-O. Nilsson and P. Händel. Time synchronization and temporal ordering of asynchronous sensor measurements of a multi-sensor navigation system. In *IEEE/ION Position, Location and Navigation Symposium*, pages 897–902. IEEE, 2010.

[40] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[41] C. Park, P. Moghadam, S. Kim, S. Sridharan, and C. Fookes. Spatiotemporal camera-lidar calibration: A targetless and structureless approach. *IEEE Robotics and Automation Letters*, 5(2):1556–1563, 2020.

[42] W. Park, N. Liu, Q. A. Chen, and Z. M. Mao. Sensor adversarial traits: Analyzing robustness of 3d object detection sensor fusion models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 484–488. IEEE, 2021.

[43] T. Qin and S. Shen. Online temporal calibration for monocular visual-inertial systems. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3662–3669. IEEE, 2018.

[44] A. Rajeev, S. Mohalik, M. G. Dixit, D. B. Chokshi, and S. Ramesh. Schedulability and end-to-end latency in distributed ecu networks: formal modeling and precise estimation. In *International Conference on Embedded Software*, pages 129–138, 2010.

[45] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[46] J. Schlatow, M. Möstl, S. Tobuschat, T. Ishigooka, and R. Ernst. Data-age analysis and optimisation for cause-effect chains in automotive control systems. In *IEEE International Symposium on Industrial Embedded Systems (SIES)*, pages 1–9, 2018.

[47] V. A. Sindagi, Y. Zhou, and O. Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019.

[48] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[49] V. Tjeng, K. Xiao, and R. Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*, 2017.

[50] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[51] J. Tu, H. Li, X. Yan, M. Ren, Y. Chen, M. Liang, E. Bitar, E. Yumer, and R. Urtasun. Exploring adversarial robustness of multi-sensor perception systems in self driving. *arXiv preprint arXiv:2101.06784*, 2021.

[52] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[53] S. Wang, T. Wu, A. Chakrabarti, and Y. Vorobeychik. Adversarial robustness of deep sensor fusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2387–2396, 2022.

[54] Z. Wang, Y. Wu, and Q. Niu. Multi-sensor fusion in automated driving: A survey. *Ieee Access*, 8:2847–2868, 2019.

[55] L. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, L. Daniel, D. Boning, and I. Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

[56] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang. Virtual sparse convolution for multimodal 3d object detection. *arXiv preprint arXiv:2303.02314*, 2023.

[57] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5418–5427, 2022.

[58] K. Xu, Z. Shi, H. Zhang, Y. Wang, K.-W. Chang, M. Huang, B. Kailkhura, X. Lin, and C.-J. Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33:1129–1141, 2020.

[59] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[60] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6):2140, 2021.

[61] Y. Zhou and O. 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*, pages 4490–4499, 2018.