

ObjectTransforms for Uncertainty Quantification and Reduction in Vision-Based Perception for Autonomous Vehicles

Nishad Sahu¹ Shounak Sural¹ Aditya Satish Patil² Ragunathan (Raj) Rajkumar¹

¹Carnegie Mellon University, Pittsburgh, PA, USA ²University Of Minnesota, Twin Cities, MN, USA

nsahu@andrew.cmu.edu ssural@andrew.cmu.edu patil255@umn.edu rajkumar@andrew.cmu.edu

Abstract

Reliable perception is fundamental for safety-critical decision-making in autonomous driving. Yet, vision-based object detector neural-networks remain vulnerable to uncertainty arising from issues such as data bias and distributional shifts. In this paper, we introduce **ObjectTransforms**, a technique for quantifying and reducing uncertainty in vision-based object detection through object-specific transformations at both training and inference times. At training time, **ObjectTransforms** perform color-space perturbations on individual objects, improving robustness to lighting and color variations. **ObjectTransforms** also uses diffusion models to generate realistic, diverse pedestrian instances. At inference time, object perturbations are applied to detected objects and the variance of detection scores are used to quantify predictive uncertainty in real-time. This uncertainty signal is then used to filter out false positives and also recover false negatives, improving the overall precision-recall curve. Experiments with YOLOv8 on the NuImages 10K dataset demonstrate that our method yields notable accuracy improvements and uncertainty reduction across all object classes during training, while predicting desirably higher uncertainty values for false positives as compared to true positives during inference. Our results highlight the potential of **ObjectTransforms** as a lightweight yet effective mechanism for reducing and quantifying uncertainty in vision-based perception during training and inference respectively.

1. Introduction

Autonomous vehicles (AVs) are poised to play an important role in increasing transportation safety, but reliable vision-based perception remains a major challenge for the wide-scale deployment of AVs due to the non-negligible occurrence of false positives and false negatives. False negatives, such as pedestrians being missed completely in low-light conditions, threaten the safety of all road users, while false positives can trigger very unsafe and/or unsettling maneuvers like phantom braking. Addressing these challenges re-

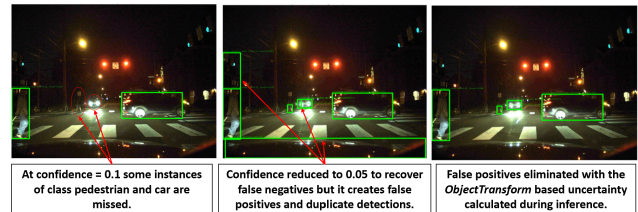


Figure 1. **Night-time crosswalk event captured from a real AV.** This is out of distribution data not used during training and validation of the YOLOv8 model. [Left] baseline detector at conf = 0.1 misses a pedestrian and a car. [Middle] Lowering to conf = 0.05 recovers these but introduces false positives. [Right] Our **ObjectTransforms**-based uncertainty filtering suppresses the false positives while retaining the true positives.

quires methods to both quantify and reduce predictive uncertainty. A confidence score in a neural network represents how confident the network is about its output, whereas uncertainty indicates how reliable the network is in correctly detecting that output. Uncertainty arises due to noisy data, insufficient training data, an improper model and/or model weights [7]. With valid uncertainty estimates, reliable sensor fusion can be done with other on-board sensors like lidars and radars, potentially augmented by V2X communication for safer downstream decision-making in AVs.

In this paper, we propose **ObjectTransforms**, a technique that applies object-specific augmentations at both training and inference stages to quantify and reduce uncertainty. At training time, **ObjectTransforms** applies targeted object perturbations in the color space and diffusion-based pedestrian transformations to improve variability in the training dataset. At inference time, **ObjectTransforms** performed controlled color perturbations to quantify predictive uncertainty, and enable the filtering of false positives and reducing false negatives. Our contributions are threefold: (i) A novel theoretical formulation of uncertainty quantification as a violation of transformation invariance; (ii) Using **ObjectTransforms** for increasing accuracy during training; and (iii) Using **ObjectTransforms** at inference time to improve the overall area under the precision-recall curve.

Table 1. Notation used in the theoretical framework in section 3

Symbol	Meaning
X	Input image
o	Object instance in the image
$Y \in \{0, 1\}$	Ground-truth label: object present/absent
$\theta \in \Theta$	Transformation parameters.
$q(\theta)$	Sampling distribution of transformations
T_θ	Object-specific transformation with parameter θ
$o_\theta = T_\theta(o)$	Object after applying transformation T_θ
$S(o_\theta) \in [0, 1]$	Detector confidence score for o_θ
τ	Detection threshold
A_τ	Detection event: $\{S(X_\theta) \geq \tau\}$
μ	Ideal probability of A_τ under transformation invariance
Z_τ	Binary variable: 1 if detection event A_τ occurs
B_θ	Event that a transformation with parameter θ is applied
C	Scene context (geometry, pose, background, illumination)
$\mathbb{E}_\theta[\cdot]$	Expectation over transformations θ drawn from $q(\theta)$
$U(C)$	Uncertainty score (in a given context C)
$U_{class}(C)$	Empirical variance of scores across transformations
$U_{bbox}(C)$	Localization uncertainty (variance of box parameters)

2. Related Work

Neural networks can sometimes produce incorrect outputs with high confidence, reducing reliability and increasing uncertainty. The quantification of network uncertainty in itself is a challenge. Uncertainty quantification in neural networks is commonly performed using Bayesian methods such as Monte Carlo dropout and deep ensembles [4, 8, 9]. Estimation of bounding box localization through Gaussian methods has also been explored [2]. Data augmentation is widely used to improve robustness in object detection. Techniques such as HSV jittering, CutMix [13], and RandAugment [3] apply global transformations. However, global perturbations do not capture object-specific variability, leaving detectors sensitive to challenging scenarios like camouflaged pedestrians or lighting-induced appearance shifts. Recently, there has been a trend towards object-level augmentation in medical imaging and natural images [14], but the area is still underexplored in the context of autonomous driving. Diffusion models [6] enable realistic data synthesis, offering opportunities for targeted augmentation. Test-Time Augmentation (TTA) [10] provides uncertainty estimates, yet remains limited to image-level perturbations. Our work proposes object-specific transformations to quantify and reduce uncertainty. To the best of our knowledge, this is the first work exploring object-specific test-time augmentations in the context of AVs for quantifying and reducing uncertainty.

3. Uncertainty As a Violation of Transformation Invariance via *ObjectTransforms*

This section presents a theoretical framework for quantifying uncertainty in vision-based 2D object detection tasks using an invariance measure. Let X denote an input image containing an object instance o with ground-truth label $Y \in \{0, 1\}$. We apply a object-specific transformation T_θ to the object o , where θ is sampled from a distribution $q(\theta)$. This operation of applying a transformation, denoted by B_θ , produces a perturbed image X_θ which con-

tains $o_\theta = T_\theta(o)$. A detector outputs a confidence score $S(o_\theta) \in [0, 1]$, and for a threshold τ we define the *detection event* $A_\tau := \{S(o_\theta) \geq \tau\}$.

Transformation-Invariance Hypothesis. For a fixed scene Context C (such as geometry, background, illumination outside the object mask), we postulate that the probability of detection should be independent of object-level transformations:

$$\Pr(A_\tau | B_\theta, C) = \Pr(A_\tau | C) = \mu, \quad \forall \theta \in \Theta. \quad (1)$$

Equation (1) formalizes our intuition that reliable detectors must not rely on superficial changes in color, texture or appearance of a single object instance.

Uncertainty as a Violation of Transformation Invariance. To reason about invariance more clearly, let Z_τ be a binary random variable corresponding to the detection event A_τ that equals 1 if the detector detects a transformed object o_θ in X_θ i.e. $S(o_\theta) \geq \tau$ and 0 otherwise. If the detector is perfectly transformation invariant, Z_τ will not change across different transformations θ , and its variance will be zero. By the *law of total variance* [12], the variability of Z_τ across transformations can be decomposed as

$$\text{Var}(Z_\tau | C) = \underbrace{\mathbb{E}_\theta[\text{Var}(Z_\tau | B_\theta, C)]}_{\text{Noise}} + \underbrace{\text{Var}_\theta(\Pr(A_\tau | B_\theta, C))}_{\text{Effect of transformations}}. \quad (2)$$

The *Noise* term captures randomness internal to the detector (e.g., dropout or stochastic inference). The *Effect of transformations* term measures how much the detection probability μ changes when we apply different transformations to the same object. If the detector is transformation invariant, probability is the same for all θ , and the second term vanishes. Thus, variance across transformations acts as a direct measure of uncertainty. In other words, when predictions are transformation invariant, their variance across transformations ought to vanish. This motivates a practical definition of uncertainty as the variance. Detectors have both a classification confidence score as well as bounding box-coordinates for predictions. This motivates a practical definition of uncertainty as the variance of either classification scores or bounding-box coordinates:

$$U_{\text{class}}(C) = \text{Var}_\theta(S(o_\theta)), \quad U_{\text{bbox}}(C) = \frac{1}{4} \sum_{d \in \{x, y, w, h\}} \text{Var}_\theta(d). \quad (3)$$

where, $U_{\text{class}}(C)$ captures the instability in classification confidence, while $U_{\text{bbox}}(C)$ captures instability in localization. We then combine them into a weighted sum.

$$U(C) = \omega_1 U_{\text{bbox}}(C) + \omega_2 U_{\text{class}}(C), \quad \omega_1 + \omega_2 = 1, \quad (4)$$

where the weights ω_1 and ω_2 can be tuned using a calibration or validation set.

ObjectTransforms. In this paper, we instantiate T_θ as HSV perturbations and diffusion-based pedestrian augmentations, as concrete cases to our approach. In general, any object specific transformations such as object transforms in

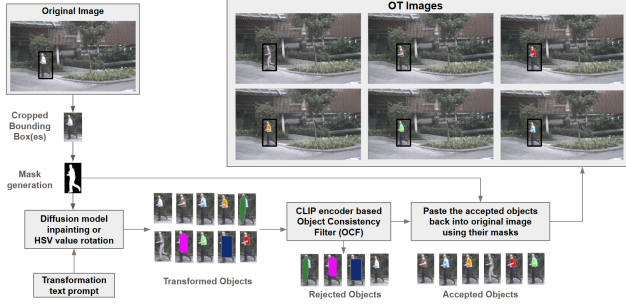


Figure 2. The diffusion-based pedestrian augmentation pipeline.

Table 2. Detection Performance (mAP50-95) on NuImg10k

Class	NuImg 10K	NuImg+ HSV (entire image)	<i>ObjectTransforms</i>
Overall	0.384	0.383	0.407
Pedestrians	0.298	0.292	0.314
Barriers	0.359	0.371	0.397
Cones	0.35	0.344	0.382
Vehicles	0.528	0.524	0.533

Note: mAP50-95 is a strict metric resulting in values that might seem modest (≈ 0.4). However, in practice, a lower value of IoU is typically used during deployment. For reference, YOLOv8 trained with *ObjectTransforms* data achieves an mAP50 of over 0.6. While we use a 10K subset of nuImages for controlled experiments, full-scale training on the entire nuImages dataset ($> 90K$ data points) is likely to push these metrics to much higher values.

color space, addition of noise, crop, flip and rotations. can be applied. The unified uncertainty metric $U(C)$ thus quantifies overall invariance-based uncertainty. This allows us to filter false positives (high $U(C)$) while recovering stable low-confidence true positives.

4. Our Methodology

ObjectTransforms are applied during training time using two methods:

(1) **Object-specific HSV transformations:** *ObjectTransforms* instead apply object-specific HSV modifications: each mask of an object is randomly perturbed in hue, saturation or value and then reinserted at its original position into the image. For note: conventional HSV jittering applies global shifts across an entire image. This morphing significantly enriches object appearance variability while preserving the scene context. It also increases the range of robustness of the model across various illumination and camouflage conditions. HSV transformations are good for classes which are inanimate objects with well-defined shapes such as vehicles, barriers and cones.

(2) **Diffusion-based pedestrian augmentations:** HSV transformation in pedestrians may introduce change, to skin and hair colors which may not be realistic. In contrast, *ObjectTransforms* takes a different approach to pedestrians by generating synthetic pedestrian samples with a diffusion model. The masks are first in-painted [11]. To maintain semantic consistency, an *Object Consistency Filter (OCF)* uses CLIP embeddings such that only pedestrian images

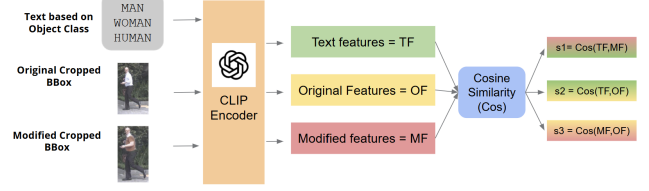


Figure 3. The Object Consistency Filter (OCF) using a CLIP encoder filters out incorrect outputs from diffusion-model inpainting

generated with high similarity to reference text embeddings are retained. Finally, images that are socially or ethically unacceptable are filtered out. Our diffusion model-based augmentation and filtering pipelines are illustrated in Figures 2 and 3 respectively.

At inference time, *ObjectTransforms* apply a controlled set of HSV perturbations to detected objects and rerun the detector. The variance in confidence scores across perturbations serves as an explicit uncertainty estimate: reliable detections maintain stable confidence scores, whereas those of ambiguous cases fluctuate considerably. This outcome enables filtering of unstable false positives that typically have higher uncertainty. Having the capability to filter false positives above an uncertainty threshold helps decrease the detection confidence threshold and yield fewer false negatives. To recover the false negatives we can decrease the confidence threshold. This may lead to increase in false positives which we can filter with the uncertainty threshold. In general, the area under the precision recall curves can improve notably with *ObjectTransforms*.

5. Experiments and Results

We conduct two sets of experiments to evaluate the effectiveness of *ObjectTransforms* in reducing uncertainty. The first assesses the effectiveness of HSV transformations and the second that of diffusion-based pedestrian augmentation. Across both sets, we use the *Yolov8x* network for 2D object detection. Our baseline dataset is the *nuImages* 10K dataset which contains 6,999 training, 1,515 validation and 1,484 test images. We maintain this partition in our experiments.

Experiment Setup 1: We detect four classes:- pedestrians, vehicles, barriers and cones. We generate 97778 images (about 14 transformations of each image in the base dataset) using different object randomized HSV transformations across different object classes and their object instances in each training image. The *Yolov8x* network is trained on (a) the base dataset, (b) the base dataset with image-level HSV augmentations and (c) the *ObjectTransforms* dataset, each for 100 epochs. As shown in Table 2, the *ObjectTransforms* dataset results in significant gains in mAP50-95 scores relative to the base dataset and the base dataset with image-level HSV augmentations.

Next, we evaluate the inference-time uncertainty U across all detections using Monte Carlo dropouts for the

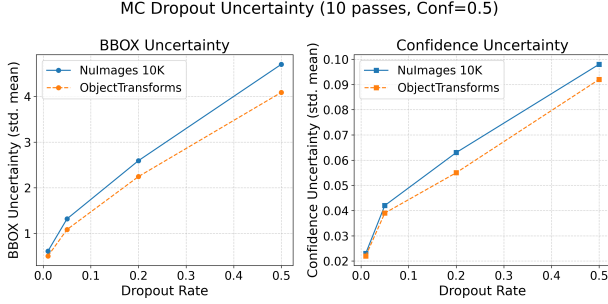


Figure 4. MC Dropout Uncertainty on NuImages [1] Test Dataset (10 passes, Conf=0.5).

Table 3. Comparison of TP and FP with and without *ObjectTransforms* (OT) during inference.

Metric	Conf.=0.25		Conf.=0.01		
	Without OT	With OT	Without OT	With OT	
		$U_{th}=0.146$		$U_{th}=0.146$	$U_{th}=0.19$
TP	3349	3186	4818	3719	4156
FP	938	640	1864	685	908
TP/FP	3.57	4.98	2.59	5.43	4.58

Table 4. Mean uncertainty scores for True Positives (TP) and False Positives (FP) obtained using the *ObjectTransforms* inference-stage uncertainty quantification on the test set.

Metric	TP Mean	FP Mean	Separation (FP/TP)
x -uncertainty	4.02×10^{-6}	2.34×10^{-5}	5.82
y -uncertainty	2.98×10^{-6}	2.74×10^{-5}	9.20
w -uncertainty	7.25×10^{-6}	6.75×10^{-5}	9.31
h -uncertainty	9.44×10^{-6}	1.07×10^{-4}	11.36
Conf. uncertainty	6.26×10^{-3}	2.60×10^{-2}	4.16

model trained on the base dataset and the *ObjectTransforms* dataset. where x, y represent the center of a bounding box, h, w are its height and width respectively and $\text{Var}(\cdot)$ is the statistical variance. Figure 4 summarizes the results: the model trained with *ObjectTransforms* yields lower bounding box uncertainty ($U_{\text{bbox}}(C)$) and confidence uncertainty ($U_S(C)$). There is a consistent reduction in uncertainty of up to 20% and the relative performance improvement increases with higher dropout rates.

We next perform uncertainty quantification using *ObjectTransforms* at inference time on our 1484 test images. The results are summarized in Table 4. The uncertainty scores across all the parameters x, y, w, h and confidence are much lower for the true positives compared to those of the false positives. Such substantive separation between the uncertainty of TPs and FPs helps in distinguishing between them, and hence isolate and highlight the FPs. For our test dataset, using the grid search technique, we found 0.25 and 0.75 to be good values of ω_1 and ω_2 respectively. With these values and a threshold of $U_{th} = 0.146$, the framework preserves 95% TPs and eliminates about 32% FPs at a detection confidence threshold of 0.25. To recover false negatives, as shown in Table 3, when we reduce the confidence

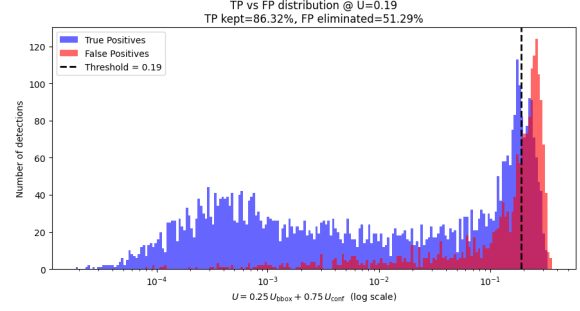


Figure 5. Distribution of TP and FP along the uncertainty score. A threshold of $U = 0.19$ for $\text{conf} = 0.01$ (refer to Table 3)

threshold to 0.01 and $U_{th}=0.146$ as before, we get a much higher TP with a decrease in FP as compared to Conf=0.25 with no *ObjectTransforms*. Furthermore, using $U_{th} = 0.19$ results in a more significant increase in TP and a slight reduction in FP. These results are presented in Table 3 and are also graphically illustrated in Figure 5. In practice, to find values of ω_1, ω_2 and U_{th} , we can use a calibration set (similar to conformal learning [5]).

Real-time Feasibility: With inference time uncertainty calculation using *ObjectTransforms* we get 5 fps with the Yolov8 Extra Large model with a GPU usage of 2.5GB on a Nvidia L4 GPU. We expect to significantly improve the frame rate with lighter versions like *nano*.

Experiment Setup 2: We detect only one class: pedestrians. The *ObjectTransforms* dataset is generated using diffusion-based inpainting [11] on the pedestrian instances in the base dataset and the OCF shown in Figure 3. The size of the *ObjectTransforms* dataset is 2072 images. We further performed a comparable analysis for this step and obtained similar results. Page length considerations limit an extended discussion.

6. Concluding Remarks

We introduced *ObjectTransforms*, a technique that applies object-specific transformations in both training and inference stages to reduce and quantify vision-based uncertainty, particularly for use in autonomous vehicle perception. The approach was evaluated using YOLOv8 on the nuImages 10K dataset. By performing color-wheel perturbations and diffusion-based pedestrian transformations, *ObjectTransforms* improves mAP50-95 during training. Inference-time uncertainty estimates further enable significant improvements in filtering of false positives and recovery of false negatives. Together, these contributions highlight the promise of object-level transforms as a lightweight yet effective approach to quantify and reduce uncertainty for safer vision-based perception. In the future, we plan to study how object detection in low-visibility conditions can be improved.

References

- [1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *CVPR*, 2020. [4](#)
- [2] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving, 2019. [2](#)
- [3] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18613–18624, 2020. [2](#)
- [4] Yarín Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016. [2](#)
- [5] Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. *arXiv preprint arXiv:1301.7375*, 2013. [4](#)
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. [2](#)
- [7] Alex Kendall and Yarín Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. [1](#)
- [8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [2](#)
- [9] Zongyao Lyu, Nolan B. Gutierrez, and William J. Beksi. An uncertainty estimation framework for probabilistic object detection, 2021. [2](#)
- [10] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation, 2021. [2](#)
- [11] Navneet Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18313–18324, 2025. [3](#), [4](#)
- [12] Kirk M Wolter and Kirk M Wolter. *Introduction to variance estimation*. Springer, 2007. [2](#)
- [13] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. [2](#)
- [14] Jiawei Zhang, Yanchun Zhang, and Xiaowei Xu. Objectaug: object-level data augmentation for semantic image segmentation. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021. [2](#)