OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Detecting Human Gesture Anomalies to Enhance Road Safety of Autonomous Vehicles Using Machine Learning Models

Yash Sharma Sooreea
yash.sooreea@gmail.com

ABSTRACT

This research paper uses Machine Learning (ML) models to detect human gesture anomalies to enhance Autonomous Vehicle safety. Specifically, it addresses the following research question: Can ML models trained on normal traffic control gestures reliably detect anomalous gestures and generalize across datasets? We hypothesize that ML models trained on normal traffic control gestures can detect abnormal poses and can transfer this recognition ability across datasets. This study investigates this hypothesis through three experiments: (1) anomaly detection on a Traffic Control Gestures (TCG) dataset using an autoencoder and reconstruction error as the anomaly proxy, (2) supervised classification of known pose categories to test the model's ability to understand semantically meaningful information, and (3) cross-dataset generalization by converting a Chinese Traffic Police (CTP) gestures dataset to the European TCG-compatible pose feature for evaluation and generalization. Statistical analysis of the highest performing autoencoder confirmed a statistically significant separation of the reconstruction errors between the normal and Out-of-Vocabulary (anomalous) frames (Mann-Whitney $U=5{,}938{,}675{,}636.5$; $p<0.001$). The MLP model using MeanPose plus standard deviation features achieved 87.1% accuracy and an F1 score of 0.821 for semantic pose information. In the cross-dataset generalization, the model regained anomaly separability between CTP "inactive" or normal, and "active" or active poses used as proxy anomalies (anomalous poses), achieving an ROC-AUC of 0.727. Together, these results suggest that the model demonstrated a clear separability between pose types, an understanding of pose gesture information, and a meaningful cross-dataset adaptation after calibration.

Keywords: Anomaly Detection, Autoencoder, Autonomous Vehicles, Human Gestures, Machine Learning

INTRODUCTION

Human traffic control gestures are a specific unsolved perception problem. Each year, both in the United States and globally, vehicle crashes cause 1.19 million fatalities and 50 million injuries worldwide; in the U.S. alone, they caused more than 42,000 deaths in 2022 (Safety – Waymo, n.d.). Despite the advancement in autonomous driving technology, 100% Autonomous Vehicle (AV) adoption is unlikely anytime soon because of technological, regulatory, and operational challenges (Alonso and Jordan, World Economic Forum, 2025). Even if there were full AV adoption, there are numerous real-world unpredictable issues, such as pedestrians and construction zones, that necessitate anomaly detection.

Traffic direction and hand-signaling by authorized humans is a common road safety mechanism across various jurisdictions and environments. In the U.K., The Highway Code provides the road safety and vehicle rules, which contain hand signals by authorized persons such as police officers and traffic officers that must be obeyed (The Highway Code - Signals by Authorised Persons - Guidance - GOV.UK, n.d.). In the U.S., the Federal Highway Administration's Manual on Uniform Traffic Control Devices describes the standardized road safety procedures that must be followed in work zones, including how to signal stop, slow, and proceed using STOP/SLOW paddles, flags, and other devices (Federal Highway Administration, 2023).

This matters for AVs because an AV is a vehicle under the control of an Automated Driving System (ADS) rather than a human driver, and it is expected to perform a Dynamic Driving Task (DDT) regardless of environmental conditions, including responding to signals and human gestures (Garrett et al., 2024). An ADS represents the hardware and software that are capable of doing an entire dynamic driving task, irrespective of the Operational Design Domain (ODD) (Garrett et al., 2024). So, an AV that cannot reliably interpret and prioritize road safety directions from authorized humans poses a risk not only to humans giving directions but also to the vehicle itself.

Although there is a growing literature on human gesture recognition (pedestrians, road users, and cyclists), the understanding of the interaction between AVs and traffic control gestures is still sparse. In 2020, Wiederer et al. (2020) developed a new dataset, the Traffic Control Gesture (TCG) dataset, that contains European traffic control skeleton data to model road traffic control gesture recognition. He et al. (2019) used a vision-based pose estimation network (Convolutional Pose Machines) and temporal modeling to create a Chinese traffic police gesture dataset. Xiong et al. (2021) noted that since traffic police gestures are primarily upper body, pose features that focus on hand gestures (instead of lower limb joints) might be more meaningful in traffic control gesture experiments. Mishra et al. (2021) pointed out that since an authorized traffic controller has the highest priority in directing road traffic (in South Korea), AVs should have a human-level understanding of traffic gestures, especially traffic control hand gestures in irregular situations.

However, existing literature doesn't cover a pose-based anomaly detection system that can recognize anomalous poses regardless of the pose estimation pipeline. To bridge the gap between this problem and

the lack of literature surrounding this domain, this research study introduces an autoencoder-based approach that not only shows meaningful separation between and detection of normal gestures and gestures outside of a known vocabulary, otherwise categorized as anomalous in this study, but also the ability to regain separability across datasets and pose estimation pipelines.

Our approach uses an autoencoder to detect anomalies, defined as "Out-of-Vocabulary" (OoV) gestures relative to a known set (e.g., stop, go, right, left). OoV gestures were gestures provided in the TCG dataset that were separate from “normal” poses. OoV gestures were not included as part of the known set of gestures (e.g., stop, go, right, left). As such, they are “Out-of-Vocabulary.” Our datasets contain human gestures based on skeleton poses. The use of human skeleton poses from video sequences to detect gesture anomalies is particularly effective in terms of preservation of human anonymity, the need for low computational requirements, and the fact that the poses are invariant to various domains and noise (Hirschorn & Avidan, 2022; Delić, Grcić, & Šegvić, 2025).

An autoencoder is a type of unsupervised neural network designed to encode the input data into a compressed representation and then decode it so that the reconstructed data is similar to the original data (Bank et al., 2023). The difference between the original data and the reconstructed data represents the reconstruction error. A high reconstruction error indicates an anomaly. By detecting anomalies, we can enhance the road safety guardrails in AV technology development, an area deemed to be of high priority for regulatory bodies (Alonso & Jordan, WEF, 2025).

The need for AVs to safely handle anomalies (human variability and unexpected gestures) in order to have strong safety guardrails leads to the following research hypothesis: Machine Learning (ML) models trained on normal traffic control gestures can detect abnormal poses and can transfer the normal behavior recognition ability across pose estimation pipelines. This study investigates this overarching research hypothesis through three experiments: (1) anomaly detection on a Traffic Control Gestures (TCG) dataset using an autoencoder and reconstruction error as the anomaly proxy, (2) supervised classification of known pose categories to test the model’s ability to understand semantically meaningful information, which is used as a baseline diagnostic for the model, and (3) cross-dataset generalization by converting a Chinese Traffic Police (CTP) gestures dataset to the European TCG-compatible pose feature format for evaluation and generalization.

METHODS

Dataset and Pose Representation

Two human-pose datasets, the Traffic Control Gestures (TCG) dataset and the Chinese Traffic Police Pose (CTP) dataset, were used across three experiments in this study. The TCG dataset, acquired from researchers at Mercedes-Benz AG, Germany, contained 3D pose skeletons represented by 17 joints. Each joint contains 3 coordinates (x, y, z) which correspond to a known human joint, distinguishable by a

provided joint dictionary mapping. For subsequent machine learning analysis, each pose was flattened into a 51-dimensional vector.

The TCG dataset contained 1,493,472 frames across 550 videos. Annotations were provided across segments (e.g., frames 1-1200) for each video that included both the state label (clear, go, stop, inactive) and pose label (both-static, normal-pose, out-of-vocabulary). Out-of-vocabulary (OoV) frames are poses that should not fall within the set of known traffic poses and are treated as anomalous in this study. Figure 1 shows a few examples of the 3D skeleton pose representation from this dataset.

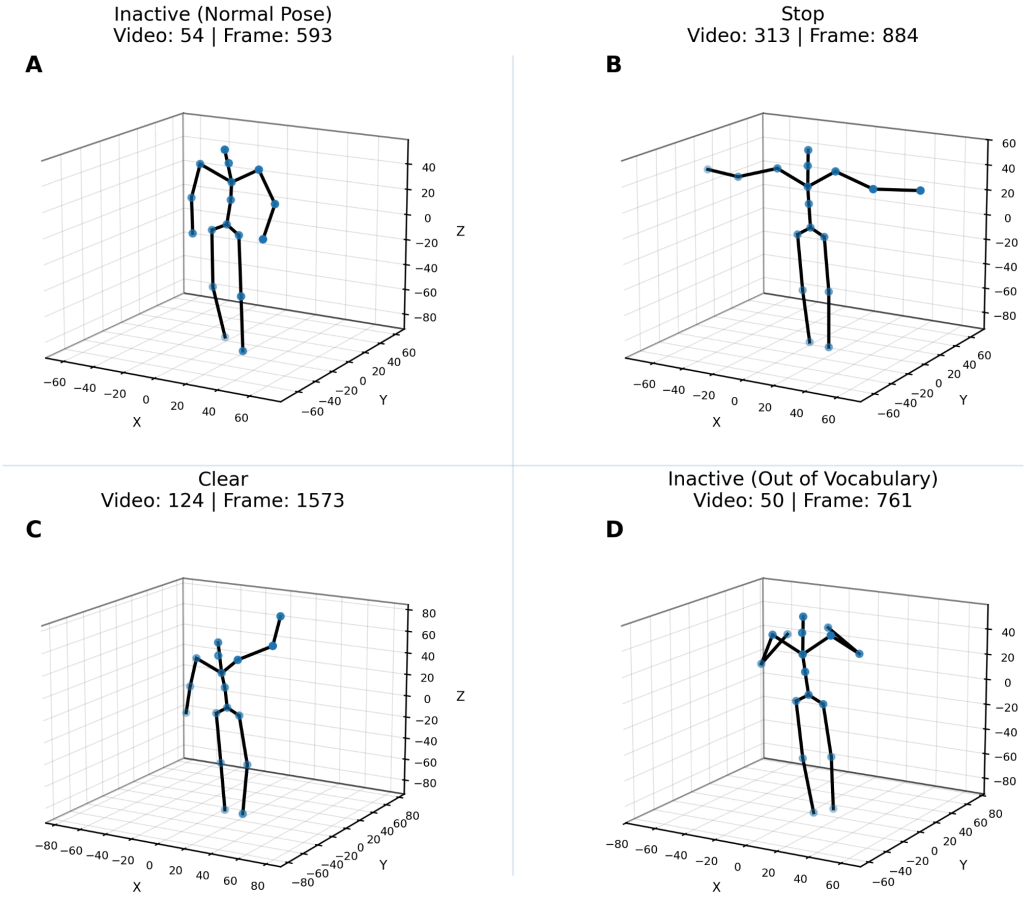

Figure 1: 3D Skeleton Pose Representations for Normal and Out-of-Vocabulary Poses. Shown are 4 different pose types found in the TCG dataset: (A) Inactive (Normal Pose), (B) Stop, (C) Clear, and (D) Inactive (Out-of-Vocabulary).

The CTP dataset was used for cross-dataset generalization. Within the repository, there were two CTPGesture sub-datasets (v1 and v2). This study utilized CTPGesture v2, which contained 15 (m4v)

videos in addition to frame-level gesture annotations, frame-level body orientation annotations, and VIBE-based SMPL pose parameters per frame.

In this study, the term "anomaly" is not a universal label for unsafe traffic gestures, but rather, a term employed within the context of the used datasets. In Experiment 1, anomalies describe OoV poses from the TCG dataset. These poses are dataset-labeled anomalies because they are explicitly identified as separate poses from safe, normal traffic-control poses. In Experiment 3, because the CTP dataset does not contain explicit OoV or unsafe-gesture labels, "active" CTP poses are processed as proxy anomalies relative to "inactive" poses. This is used to test whether the TCG-trained autoencoder can successfully regain separability across a new pose-estimation pipeline, but not to claim that the active CTP gestures are inherently unsafe traffic gestures.

Experiment 1: Autoencoder-Based Anomaly Detection

Experiment 1 was conducted by building a series of models on a loaded and pre-processed training set from the TCG dataset. An autoencoder was trained to detect anomalous traffic poses using reconstruction error. To accomplish this, the TCG dataset was split at the video sequence level to prevent frames from the same video from appearing in both training and testing. Experiment 1 has 757,690 normal test frames and 93,635 OoV test frames. Sequences containing no OoV frames were identified as normal-only sequences, and 80% of those normal-only sequences were randomly selected for training. A subset of the training videos was held out for validation. The test set contained the remaining normal-only videos and all videos containing OoV frames, which ensured that OoV examples were not used during training. This helps avoid data leakage, where the model could memorize representations of similar frames, inflating final performance. Pose coordinates were standardized via StandardScaler prior to model training.

The ML model utilized in this experiment was the autoencoder (see Figure 2 for an illustration of the autoencoder architecture). An autoencoder is a type of unsupervised neural network designed to encode input data into a compressed representation and then decode it, resulting in reconstructed data that is similar to the original data (Bank et al., 2023). The inputs for this model were the 51-dimensional pose vectors. The model learned to recreate normal poses. When a pose differs from normal training examples, the reconstruction is worse, producing a larger error.

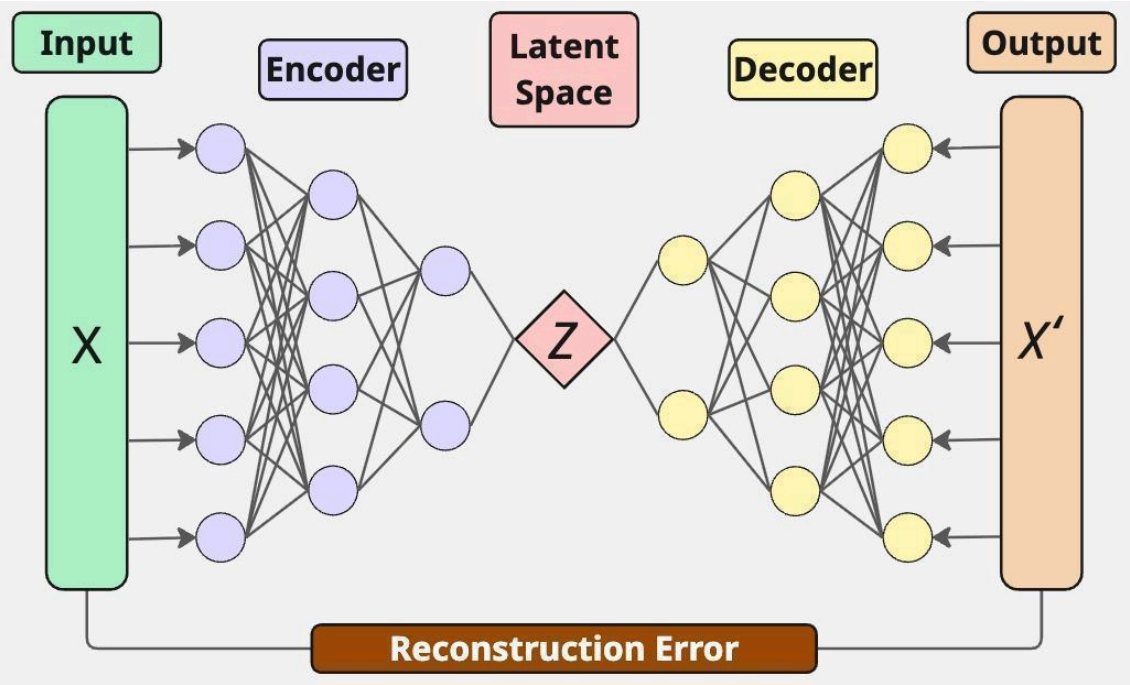

Figure 2: Autoencoder Architecture. The autoencoder is a type of unsupervised neural network that encodes input data X into a compressed dimensional representation or bottleneck (Latent Space Z), which constrains the data, and then a decoder reconstructs the data to create output X'. The difference between the input and output is the Reconstructed Error.

To identify an effective architecture, 15 autoencoder models were trained using a randomized grid search over several hyperparameters, including embedding dimension, hidden layer sizes, dropout rate, learning rate, and L2 regularization strength. The best performing model hyperparameters are summarized in Table 1.

Hyperparameter	**Value**
Encoding Dimension	32
Hidden Layer Sizes	(64, 32)
Dropout Rate	0.0
Learning Rate	0.001
L2 Regularization	0.0

Table 1: Hyperparameter Values for Best-Performing Autoencoder Model.

To train the autoencoder, the Adam optimizer was used with a batch size of 256, which ran for 40 epochs. Early stopping was monitored on validation loss, with a patience of 5 epochs. It then restored the best validation loss weights. Mean Squared Error (MSE) was used as the loss function for model training, and a separate scaled training validation dataset was used for validation. The best autoencoder was selected based on its performance metrics, sorted by the highest ROC-AUC and PR-AUC scores.

Additionally, this study implemented two anomaly detection approaches, Principal Component Analysis (PCA) and Isolation Forest (ISO), as baselines. PCA identifies anomalies after dimensionality reduction (Maćkiewicz and Ratajczak), while ISO searches for observations that differ from the majority (Liu et al.). All models were evaluated using Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and Precision–Recall Area Under the Curve (PR-AUC). The autoencoder was also evaluated with the reconstruction error, defined by calculating the mean squared error between the original and reconstructed frames.

For threshold-dependent metrics, the anomaly threshold was selected using only normal training reconstruction errors, which were set at the 95th percentile of reconstruction errors from the normal-only training frames. Held-out test frames with reconstruction error above this threshold were classified as anomalous, while frames at or below this threshold were classified as normal. This threshold was chosen before evaluating the test set, so test labels were not used to select the threshold. This method was utilized for Experiment 1. In Experiment 3, threshold-dependent metrics were calculated after calibration using the same percentile-based thresholding approach on the train set, with inactive CTP frames representing the normal calibration distribution. Experiment 2 did not require a reconstruction-error threshold because it was a supervised classification task.

The normal and OoV reconstruction errors were compared using the Mann-Whitney Test to determine if there was a significant difference between the two groups. Additionally, Cliff's Delta was used to measure the degree to which OoV reconstruction errors were higher than the normal reconstruction errors. To calculate the Confidence Interval (CI), 1000 bootstrap samples were generated by resampling held-out test frames with replacement.

Experiment 2: Supervised Pose Classification

The goal of Experiment 2 was to evaluate whether the frame-level pose features used in this study encoded semantically meaningful information about known traffic control gestures. By training a supervised classifier on non–out-of-vocabulary (non-OoV) pose data, this experiment assessed the separability of known gesture categories using the same pose representation employed in the anomaly detection experiment. The TCG data was once again split into 80% training and 20% testing datasets, which was performed using GroupShuffleSplit at the video-level to avoid data leakage. Experiment 2 had 72,399 training windows and 19,390 test windows, with 440 train sequences and 110 test sequences, and was conducted by building baseline models that validated the results from Experiment 1. This was done by creating a window-based extraction with the average pose (MeanPose) and standard deviation

(MeanPose + STD) as features across two models: Logistic Regression and Multi-layer Perceptron (MLP). These two models were tested and evaluated with accuracy and F1 score.

To classify poses, this experiment used short, overlapping sliding windows containing several frames to make a pose prediction rather than classifying individual images. Each sliding window spanned 50 frames, with a stride of 15 frames. The overlap of windows ensured that frame-level information on the borders of each sequence was not lost. For each window, the input into the model was calculated by averaging the pose vectors of all frames within that window. Additionally, we tested adding the standard deviation to the mean pose as a set of additional features, resulting in 102-dimensional pose vectors, with the first 51 points representing the average pose within that window and the second 51 points representing the standard deviation of those averages.

Experiment 3: Cross-Dataset Generalization

Experiment 3 examined whether the model trained on normal traffic control gestures to detect abnormal or OoV gestures in Experiment 1 could generalize across datasets and geographies. This experiment used the best-performing Autoencoder that was trained on the TCG dataset and applied it to the CTP dataset. The objective was to demonstrate whether the Autoencoder could identify anomalies in the CTP dataset. Experiment 3 had 12 training videos and 3 testing videos for CTP calibration. A three-stage process was used in Experiment 3 to evaluate cross-generalization.

The first stage involved running the TCG Autoencoder directly on the CTP dataset. The CTP pose joints were mapped to the closest TCG pose joints, allowing the autoencoder to process the new data. In order to align the pose estimation pipelines, the joints in the VIBE-generated SMPL skeletons from the CTP dataset were mapped onto the TCG joint structure by centering and rescaling the CTP pose coordinates to match the TCG representation. Then the TCG StandardScaler from the Python scikit-learn package was applied to the CTP pose coordinates. Reconstruction error was calculated for each frame to evaluate the model's performance.

The second stage consisted of normalizing the CTP pose joints using a dataset-specific StandardScaler to account for differences in the underlying distribution. The TCG Autoencoder was run again, and reconstruction error was recalculated to determine if normalization contributed to performance.

In the third stage, the Autoencoder was fine-tuned using the CTP frames with "inactive" poses. Although the CTP dataset lacked explicit OoV frames, we aimed to evaluate whether the autoencoder could distinguish between inactive (normal poses) and active poses used as proxy anomalies. The CTP dataset was split at the video level using GroupShuffleSplit, with 80% of videos used for calibration and 20% held out for testing. This meant that 12 videos were used for fine-tuning and 3 videos were held out for final testing. The fine-tuning used only inactive frames from the training videos. A StandardScaler was fit only on these inactive training frames, then applied to the held-out CTP test frames. The autoencoder was initialized with the TCG-trained weights and fine-tuned using Adam with a learning rate of 1e-4, MSE loss, batch size 256, and a maximum of 15 epochs. Ten percent of the inactive training frames were used

as a validation split for monitoring validation loss. Early stopping monitored validation loss with a patience of 2 epochs and restored the best validation-loss weights. Final evaluation was performed only on the three held-out CTP test videos. Reconstruction error was recalculated to assess whether anomaly separation improved after calibration.

RESULTS

Autoencoder Reconstruction Error Separates Normal and OoV Poses

The goal of Experiment 1 was to build an autoencoder trained only on normal pose frames, then evaluate whether it assigns higher reconstruction error (mean squared error (MSE)) to out-of-vocabulary (OoV) poses.

Of the 15 autoencoders trained, the highest performing one achieved a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.916 (95% CI: 0.914-0.917) and a Precision–Recall Area Under the Curve (PR-AUC) of 0.690 (95% CI: 0.687-0.693). For comparison, two baseline models, Principal Component Analysis (PCA) and Isolation Forest (ISO), were tested. PCA achieved a ROC-AUC score of 0.927 (95% CI: 0.926-0.928) and a PR-AUC score of 0.740 (95% CI: 0.737-0.743), while the other baseline model, ISO, achieved a ROC-AUC score of 0.538 (95% CI: 0.536-0.540) and a PR-AUC score of 0.153 (95% CI: 0.151-0.154) (Figure 3). Although PCA slightly outperformed the autoencoder on the TCG data, we chose to utilize the autoencoder for this research because of its capability to be fine-tuned and its ability to support nonlinear relationships, which is especially useful for more complex hand gestures.

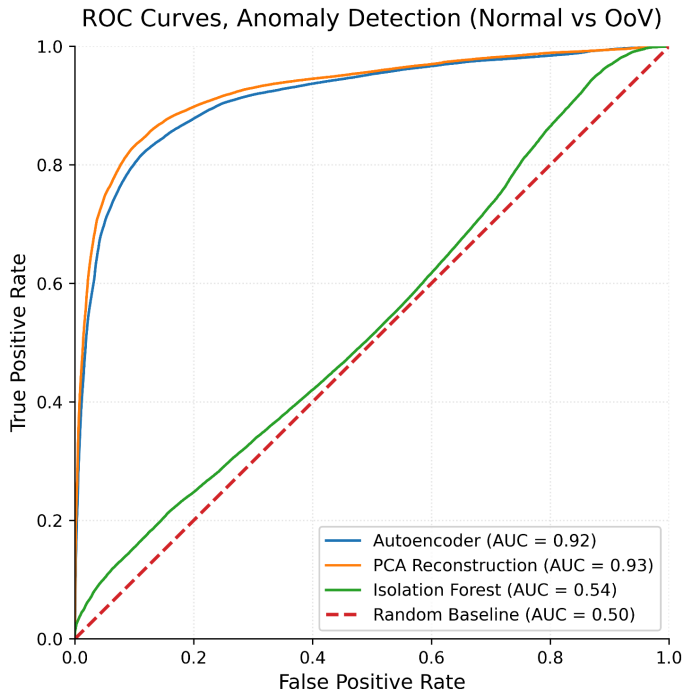

Figure 3: Receiver Operating Characteristic (ROC) curves for Anomaly Detection between Normal and Out-of-Vocabulary poses for each of the trained models (Autoencoder, PCA, Isolation Forest) on the TCG dataset. The Area Under the Curve (AUC) is displayed in the figure. The blue curve illustrates the performance of the Autoencoder (AUC: 0.92). The Principal Component Analysis (PCA) in orange (AUC: 0.93) and the Isolation Forest (ISO) (AUC: 0.54) in green are the baseline models. The red dashed line is a random classifier.

Statistical analysis of the highest performing autoencoder confirmed a statistically significant separation of the reconstruction errors between the normal and OoV frames (Mann-Whitney U = 5,938,675,636.5; $p < 0.001$) (Figure 4) as well as 90.81% accuracy, 56.03% precision, and 76.48% recall. The Mann-Whitney U test showed that OoV frames had higher reconstruction error than normal frames. The effect size between the two distributions was notably large (Cliff's Delta = 0.83).

Figure 4: Reconstruction Errors for Normal and Out-of-Vocabulary Poses. The highest performing autoencoder indicated a statistically significant separation of the reconstruction errors between the normal and OoV frames (Mann-Whitney U = 5,938,675,636.5; $p < 0.001$, Cliff's Delta = 0.831).

Individual pose skeletons were evaluated to visually comprehend where reconstruction errors manifested within both normal and OoV frames (Figure 5).

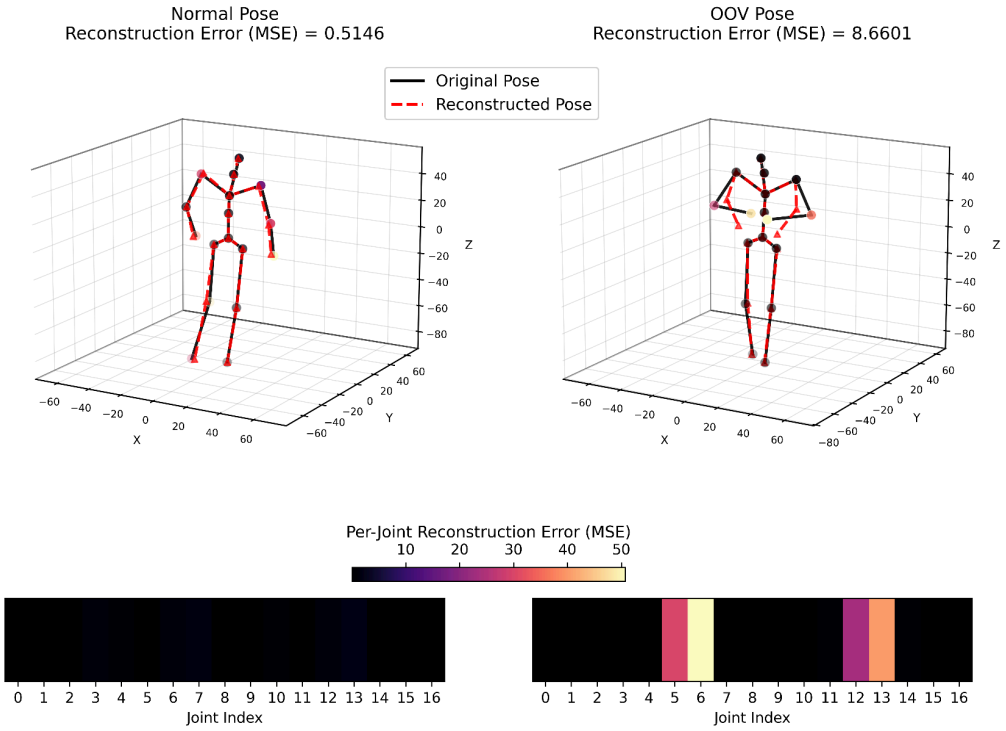

Figure 5: Heat Map of Reconstruction Errors for Normal and Out-of-Vocabulary Poses. MSE represents mean squared errors (reconstruction errors). OoV poses demonstrated higher reconstruction errors (MSE: 8.6601) compared to Normal poses (MSE: 0.5146).

Pose Skeletons Encode Semantic Information for Gesture Classification

Of the models tested, the MLP models consistently outperformed logistic regression models with higher accuracy and F1 score (Table 2).

Model	Feature Set	Accuracy (95% CI)	F1 Score (95% CI)
MLP	MeanPose	86.4% (85.9%, 86.8%)	0.800 (0.792, 0.808)
	MeanPose+STD	87.1% (86.6%, 87.5%)	0.821 (0.813, 0.829)
Logistic Regression	MeanPose	76.9% (76.4%, 77.5%)	0.630 (0.620, 0.639)
	MeanPose+STD	78.9% (78.3%, 79.5%)	0.677 (0.668, 0.687)

Table 2: Test Performance of Supervised Machine Learning Models. MLP is a Multilayer Perceptron. STD is Standard Deviation. CI is Confidence Interval. The MLP model with MeanPose+STD performed the best with an accuracy of 87.1% and an F1 Score of 0.821.

The best-performing model was the MLP with MeanPose+STD, which achieved 87.1% accuracy, 81.3% macro precision, 83.0% macro recall, and an F1 score of 0.821. All the models across both feature sets gave significant results to show that the pose structures do contain semantic information.

Autoencoder Regains Cross-Dataset Separability After Calibration

The goal of Experiment 3 was to evaluate the TCG-trained autoencoder's ability to generalize its learning across the CTP dataset. This experiment tests the autoencoder beyond the limits of its expected data, which mirrors AVs in real-life, unexpected circumstances.

After the first stage, direct transfer of the original autoencoder on the CTP dataset resulted in a substantially higher reconstruction error. The median reconstruction error for CTP "inactive" or normal poses was 1.83, compared to a median reconstruction error of 0.0036 for normal poses from the TCG dataset. This is approximately 500 times higher than the TCG baseline, thus indicating a failure in generalizing across the datasets without adaptation.

In the second stage, after normalizing the CTP dataset using a dataset-specific StandardScaler, the median reconstruction error for CTP "inactive" or normal poses reduced to 0.26, still roughly 70 times higher than the TCG baseline.

After the final stage, the autoencoder, fine-tuned on normal-only poses from the CTP dataset, regained performance. The median reconstruction error dropped to 0.0054, which is comparable to the TCG baseline. The autoencoder achieved 62.7% accuracy, 63.7% precision, and 13.4% recall. The model also regained anomaly separability between CTP "inactive" or normal and "active" or active poses used as proxy anomalies, achieving an ROC-AUC of 0.727 (95% CI: 0.722-0.733). The results indicate that while the model did not automatically generalize across the European (TCG) and Chinese (CTP) datasets, some fine-tuning using representative "normal" (inactive) frames enabled the model to adapt and gain meaningful separation again.

DISCUSSION

Our autoencoder successfully detected anomalous traffic gestures using a reconstruction error. Our results showed that the reconstruction error for OoV poses was significantly higher for normal poses, which indicated clear separability between normal and anomalous poses. As shown by the results of the Supervised Classifiers from Experiment 2, the pose features contained meaningful gesture information. In Experiment 3, despite the cross-dataset transfer initially failing due to differences in pose extraction pipelines, separation and performance were improved after normal pose calibration.

Although the autoencoder regained pose separation performance after calibration, limitations existed in the experiments. The TCG dataset contains limited information about the pose skeleton and limited information about the surrounding environment or the direction that the officer is facing. In experiment 1, we used static poses as opposed to full temporal sequences or a temporal aware model, which could pose a challenge in dynamic and animated environments. In experiment 3, the CTP and TCG datasets had label mismatches, so we chose to represent the CTP dataset's normal poses as inactive and active poses as anomalous. Because the CTP dataset does not contain explicit OoV labels, active gestures were treated as a proxy distribution shift relative to inactive frames, rather than as confirmed unsafe or invalid gestures. It is important to note, however, that this proxy is not equivalent to authentic anomalous gestures that do actually appear in the form of unsafe, ambiguous, or invalid road signals. Furthermore, the pose estimation pipelines were different, and although the results accounted for this, the domain shift between the two pipelines is still a challenge for cross-dataset generalization.

Results from Experiment 1 indicated that the trained autoencoder learned the structural representation of normal poses and that OoV poses lay outside of this learned structure. This was shown by the reconstruction error increasing with the unfamiliar or OoV gestures. Because we had a statistically significant Mann-Whitney ($p<0.001$) and a considerably large Cliff's Delta ($|\boldsymbol{\delta}| = 0.83$), we confirmed that there existed a strong separation of the underlying pose structures. Because autoencoders do not automatically fail on unseen data, the model's high reconstruction error for OoV frames indicates that the model reliably distinguished between anomalous and normal frames and that it successfully learned meaningful representations of traffic control gestures. So, if the OoV poses were instead formatted with the same learned structure as the normal poses, then the autoencoder, despite not being trained on them, would have generalized well and had a lower reconstruction error. While the results indicate that PCA slightly outperformed the autoencoder, ultimately, we chose the autoencoder as the goal of this study is not only to identify the strongest baseline on one dataset, but also to evaluate a reconstruction-based framework that could be adapted across pose-estimation pipelines. PCA provides a useful linear reconstruction baseline, but its linear structure limits its ability to model more complex pose relationships and makes it less flexible for future extensions. However, the PCA is still valuable because it shows that pose structure is strongly informative, even with a simple linear model. Autoencoders can learn nonlinear relationships among joints, can be fine-tuned on normal examples from a new domain, and can be extended to architectures that utilize temporal data to model motion across frames. Thus, the autoencoder architecture provides a more flexible system for cross-dataset calibration and future temporal modeling structures.

Experiment 1 yielded promising results, but in order to validate that the pose structures contained meaningful semantic information, we tested supervised classifiers on the known gestures. We utilized a window-based approach for training and evaluation because deciding on every frame can be brittle and unreliable for the model to perform at such speed. Rather, it is more meaningful to make decisions over an aggregation of frames. We tested two different variations of this aggregation, first with just the average pose vectors in the window (MeanPose), and second, both the average pose vectors and the standard

deviation (MeanPose+STD), to quantify how much a pose moved within a certain window. Our results show high accuracy and F1 score for both MeanPose and MeanPose+STD across both Logistic Regression and MLP models, indicating that the poses contain semantic information. This further supports that the Experiment 1 autoencoder is learning meaningful pose representations.

While we showed that the autoencoder works specifically for this pose estimation pipeline from the TCG data, Experiment 3 was conducted to evaluate how this model would perform in a real-world deployment setting. We saw that direct transfer of using the TCG autoencoder on the CTP-mapped joints failed due to a domain shift or an incompatible pose estimate pipeline with the CTP dataset. Additionally, the independent normalization of the CTP dataset partially reduced the domain shift gap, but only when we fine-tuned the autoencoder on normal poses did we achieve regained anomaly separation. This is a specifically important result because in real-world settings, it allows us to maintain one strong base autoencoder anomaly separation model but fine-tune it on specific geographies, other domains, or other pose estimation pipelines to create a system that is easier to maintain but applicable across multiple systems.

Potential future studies on this work could include using temporal sequencing models to predict human gestures rather than static pose geometry recognition processes. This would offer increasing reliability for anomalous pose detection and more accurately reflect real-world situations. To further this study, sourcing additional traffic gesture datasets and verifying the results would be valuable for dataset generalizations, such that it would encompass a varying range of poses, gestures, and environments. Finally, testing the pipelines in this study with real-time video pipelines that extract poses would expand this study's application in live AVs to bolster reliable real-time AV reactions and decisions.

Traffic control gestures occur in unpredictable environments such as pedestrian crossings, construction, and heavily populated roads, so it is imperative that AV systems recognize gestures that fall outside of expected patterns to increase safety. Our anomaly reconstruction-based system can identify meaningful information from pose data to recognize unusual or out-of-scope traffic gestures. Additionally, we show that the autoencoder can adapt across pose estimation pipelines with a small calibration step. The proposed anomaly detection pipeline in this study can act as a component of a safety guardrail in a future AV perception safety pipeline.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Ronil Synghal, who provided invaluable insights, constant support, and technical guidance for this research. I would also like to thank Julian Wiederer from Mercedes-Benz AG, Germany, who graciously provided access to the European Traffic Control Gestures (TCG) dataset for this study.

REFERENCES

Againerju. (n.d.). GitHub - againerju/tcg_recognition: Training and evaluation kit for traffic control gesture recognition on the TCG dataset. GitHub. https://github.com/againerju/tcg_recognition/tree/master

Alonso, M., & Jordan, P. (2025, April). Autonomous vehicles: Timeline and roadmap ahead. World Economic Forum. https://www.weforum.org/publications/autonomous-vehicles-timeline-and-roadmap-ahead/

Bank, D., Koenigstein, N., Giryes, R. (2023). Autoencoders. In: Rokach, L., Maimon, O., Shmueli, E. (eds) Machine Learning for Data Science Handbook. Springer, Cham. https://doi.org/10.1007/978-3-031-24628-9_16

Delić, A., Grcić, M., & Šegvić, S. (2025). Sequential keypoint density estimator: an overlooked baseline of skeleton-based video anomaly detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2506.18368

Federal Highway Administration. (2023). Manual on Uniform Traffic control devices for streets and highways (pp. 765–765) [Manual]. https://mutcd.fhwa.dot.gov/pdfs/11th_Edition/part6.pdf

Garrett, K., Ma, J., Krueger, B., Morgan, A., Schaefer, R., Leidos, Inc., University of California, Los Angeles, Synesis Partners, LLC, & Kittelson & Associates, Inc. (2024). Automated Driving Systems Operational Behavior and Traffic Regulations Information: Model Plan. https://ops.fhwa.dot.gov/publications/fhwahop21039/fhwahop21039.pdf

He, J., Zhang, C., He, X., & Dong, R. (2019). Visual Recognition of traffic police gestures with convolutional pose machine and handcrafted features. Neurocomputing, 390, 248–259. https://doi.org/10.1016/j.neucom.2019.07.103

Hirschorn, O., & Avidan, S. (2022). Normalizing Flows for Human Pose Anomaly Detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.10946

Mishra, A., Kim, J., Cha, J., Kim, D., & Kim, S. (2021). Authorized Traffic Controller hand gesture recognition for Situation-Aware autonomous driving. Sensors, 21(23), 7914. https://doi.org/10.3390/s21237914

Liu, Fei Tony, et al. “Isolation Forest.” 2008 Eighth IEEE International Conference on Data Mining, Dec. 2008, https://doi.org/10.1109/icdm.2008.17.

Maćkiewicz, Andrzej, and Waldemar Ratajczak. "Principal Components Analysis (PCA)." Computers & Geosciences, vol. 19, no. 3, Mar. 1993, pp. 303–42, https://doi.org/10.1016/0098-3004(93)90090-r.

Safety – Waymo. (n.d.). Waymo. https://waymo.com/safety

The Highway Code - Signals by authorised persons - Guidance - GOV.UK. (n.d.). https://www.gov.uk/guidance/the-highway-code/signals-by-authorised-persons

Wiederer, J., Bouazizi, A., Kressel, U., & Belagiannis, V. (2020a, July 31). Traffic control gesture recognition for autonomous vehicles. arXiv.org. https://arxiv.org/abs/2007.16072

Xiong, X., Wu, H., Min, W., Xu, J., Fu, Q., & Peng, C. (2021). Traffic police gesture recognition based on gesture skeleton extractor and multichannel dilated graph convolution network. Electronics, 10(5), 551. https://doi.org/10.3390/electronics10050551

Zc. (n.d.). GitHub - zc402/traffic-gesture-datasets: Summary of the Chinese traffic police gesture datasets published in our papers. GitHub. https://github.com/zc402/traffic-gesture-datasets/tree/main

ABOUT THE AUTHOR

Yash Sooreea is a high school senior at The College Preparatory School in Oakland, California. He has interests in engineering, computer science, and STEM-related fields. He plans to study in these areas in college.