Computer Vision for Transit Travel Time Prediction: An End-to-End Framework Using Roadside Urban Imagery

Awad Abdelhalim[1*] and Jinhua Zhao[1]

[1]Department of Urban Studies and Planning, Massachusetts Institute of Technology, 77 Mass Ave, Cambridge, 02139, MA, USA.

*Corresponding author(s). E-mail(s): awadt@mit.edu;
Contributing authors: jinhua@mit.edu;

Abstract

Accurate travel time estimation is paramount for providing transit users with reliable schedules and dependable real-time information. This work is the first to utilize roadside urban imagery to aid transit agencies and practitioners in improving travel time prediction. We propose and evaluate an end-to-end framework integrating traditional transit data sources with a roadside camera for automated image data acquisition, labeling, and model training to predict transit travel times across a segment of interest. First, we show how the General Transit Feed Specification (GTFS) real-time data can be utilized as an efficient activation mechanism for a roadside camera unit monitoring a segment of interest. Second, Automated Vehicle Location (AVL) data is utilized to generate ground truth labels for the acquired images based on the observed transit travel time percentiles across the camera-monitored segment during the time of image acquisition. Finally, the generated labeled image dataset is used to train and thoroughly evaluate a Vision Transformer (ViT) model to predict a discrete transit travel time range (band). The results of this exploratory study illustrate that the ViT model is able to learn image features and contents that best help it deduce the expected travel time range with an average validation accuracy ranging between 80%-85%. We assess the interpretability of the ViT model's predictions and showcase how this discrete travel time band prediction can subsequently improve continuous transit travel time estimation. The workflow and results presented in this study provide an end-to-end, scalable, automated, and highly efficient approach for integrating traditional transit data sources and roadside imagery to improve the estimation of transit travel duration. This work also demonstrates the added value of incorporating real-time information from computer-vision sources, which are becoming increasingly accessible and can have major implications for improving transit operations and passenger real-time information.

Keywords: Travel time prediction, computer vision, vision transformers.

1 Introduction

Accurate and reliable travel time prediction plays a critical role in all aspects of transportation planning, this is even more crucial when it comes to public transit (Lin et al, 2005). Efficient operations, service delivery, riders' experience, and the general perception of public transit are greatly shaped by the reliability and on-time performance of the system. Travel and arrival time prediction, in general, and particularly for public transit applications are fields of research that have been well studied over the years. The recent advancements in machine learning and artificial intelligence technologies have significantly improved the state of practice. This is in part due to the wealth of data available for and from day-to-day public transit operations, including high granularity data for passenger count and movements (APC), automated fare collection (AFC), automated vehicle locations (AVL), and the information from the General Transit Feed Specification (GTFS). While these data sources offer an abundance of scheduling, usage, and performance measures that can be used for tasks like travel time prediction for transit, transit vehicles still - for the most part - share road infrastructure with other roadway users. This necessitates combining transit-specific data with more generalized data sources that can offer information about the overall traffic state and roadway infrastructure, which are often challenging to acquire.

The recent advancements in the fields of deep learning, machine perception, and computer vision have made image data extremely useful. Tasks including image classification, detection, and tracking of objects within images can be accomplished with ease, with an ever-growing multitude of frameworks and architectures to choose from. The biggest drawback when opting to utilize computer vision architectures is that they remain extremely data-hungry, requiring copious amounts of data that needs to be acquired and, more often than not, manually labeled to train vision models. And while this is a challenge facing the broader computer vision community, it is further exacerbated in domain-specific applications like transportation systems (Dilek and Dener, 2023). In the specific context of public transit, the need to acquire external technology and talent for these tasks of data acquisition and integration is often cost-prohibitive, hindering the ability to benefit from fusing the abundant traditional and new-found data sources (Ge et al, 2021). The authors of this paper see immense value in incorporating the domain knowledge of relevant data sources to create a streamlined framework for image data acquisition, labeling, and vision model training combining roadside imagery with transit data sources. This study presents our exploratory framework for this concept and experimental results from a pilot study conducted in a segment of Massachusetts Avenue in Cambridge, MA, USA.

In this study, we propose and evaluate TranViT, an end-to-end framework for efficiently integrating real-time GTFS, AVL, roadside urban imagery, and a Vision Transformer (ViT) architecture for predicting transit travel time through a camera-monitored segment. The main contribution of this work is presenting a blueprint to transit practitioners and researchers for the efficient utilization of traditional transit data sources to acquire, label, and train data for computer vision tasks. We demonstrate the resulting highly accurate and interpretable predictions and discuss their implications for the state of the practice. The following sections of this paper are as follows: (a) a survey of related literature in transit travel time prediction and applications of deep learning and computer vision, (b) a detailed breakdown of the proposed TranViT framework, (c) results and analyses of the case study, and, (d) discussion and conclusions.

2 Related Work

2.1 General Traffic and Transit Travel Time Prediction

There is a rich body of literature on travel time prediction, a topic that has been well-studied for years. In the context of general traffic, this often falls into larger-scale traffic state estimation, which along with travel time and speed includes predicting traffic flow and density (Wang and Papageorgiou, 2005; Yang, 2005; Work et al, 2008; Yildirimoglu and Geroliminis, 2013). Those works among others have mostly relied on probe vehicle GPS and loop detector data, which are often limited in size and temporal coverage, noisy, and require the use of complementary techniques to overcome the data quality issues, including particle and Kalman filtering, and machine learning (Wang et al, 2008; Chen et al, 2011; Jenelius and Koutsopoulos, 2013) in combination with underlying traffic state models and, in some cases, simulation modeling.

For transit, accurate travel time prediction is paramount for providing transit users with reliable schedules and dependable real-time information about their transit vehicles' arrival times, and for the efficient implementation of operation strategies such as transit signal priority (Zeng et al, 2014; Abdelhalim and Abbas, 2018). The methods used in general traffic state estimation and travel prediction, particularly from probe vehicle analyses, may not always translate directly into the context of transit. Albeit sharing the same roadway infrastructure; the vehicle dynamics and operations of transit (continuous start-stopping), among other factors like passenger interactions with operators, influence the movement of transit within traffic streams in a way that is not necessarily reflective of the overall traffic. The need to monitor this level of complexity in size (fleets as opposed to individual vehicles) and operation, however, has resulted in an abundance of transit data sources being available to practitioners. Of the aforementioned data sources, AVL data has particularly been at the center of numerous research efforts. An early study by Cathey and Dailey (2003) offered a generalized framework to utilize AVL data for making transit arrival time predictions. The growing adoption of AVL systems by transit agencies allowed researchers to develop real-time applications (Jeong and Rilett, 2005; Shalaby and Farhan, 2004). While these studies have demonstrated the ability to obtain accurate travel time predictions from AVL data, the lack of information regarding other influencing factors like overall congestion state and the vehicles' own dwell time were limiting factors. Works by Yu et al (2010, 2011) have demonstrated that incorporating additional external data sources (e.g. weather data) and using stop-level travel time that captures this variation in dwell time more than the route-level analysis results in improved predictions.

The past decade has witnessed substantial strides in data availability (higher frequency AVL, GTFS-RT, APCs, etc) and real-time assessment and predictive methodologies (Park et al, 2020; Elliott and Lumley, 2020; Aemmer et al, 2022; Samal et al, 2017). While state estimation methods based on Kalman filters, and machine learning models based on boosting and ensemble remained popular (Zhang and Haghani, 2015; Gaikwad and Varma, 2019), deep learning architectures based on recurrent neural networks (RNNs) have also proven high competency in the travel time prediction task. Studies by Zhou et al (2019) and Pang et al (2018) concluded that RNN-based models significantly outperform other state-of-the-art methods. The proposed Long Short-Term Memory (LSTM) RNN achieved a minimum of 10% improvement in the mean absolute error compared to other traditional models. Han et al (2020) proposed a method based on position calibration and an LSTM model that accurately predicts transit arrival time, with an error below two minutes for the 8^{th} downstream stop.

2.2 Computer Vision, Vision Transformers, and Related Works in Transportation Applications

Computer vision (CV) is a field of artificial intelligence focused on deriving useful information from images and image-based data (e.g. videos). This includes tasks like image classification, object recognition, and multi-object recognition and tracking. Albeit being proposed since the early 1990s (LeCun et al, 1998), the rapid evolution of Convolutional Neural Networks (CNNs) in the past decade helped establish their place as the undisputed backbone for innumerable architectures that conduct the various aspects of computer vision tasks. These CV-based methods have been widely adopted for a variety of applications in the transportation field, including traffic speed, turn count, and density estimation (Buch et al, 2011; Abdelhalim et al, 2021; Gokasar and Timurogullari, 2021), safety (Tageldin et al, 2014; Sayed et al, 2013; Abdelhalim, 2021), and autonomous driving (Janai et al, 2020). The use of CV in transit applications remains extremely rare, with few applications in railway intrusion detection (Wang and Yu, 2021), and a recent study by Sipetas et al (2020) who utilized a CV-based video processing component as a part of a larger framework to estimate left-behind subway passengers.

The Vision Transformer (ViT) architecture introduced by Dosovitskiy et al (2020) was inspired by the immense success of transformer architectures in the field of natural language processing (NLP) (Vaswani et al, 2017). Converse to CNNs, which have been the de-facto method for most computer vision tasks, the vision transformer architecture doesn't introduce the implicit bias of the convolutional and pooling layers present in CNNs, allowing the trained model to better extract global information from images. Although this comes at the cost of requiring more training data, the ViT-based models have been shown to outperform their CNN counterparts at their introduction in 2020. While there has been a tug of war between ViT and CNN-based models since then, ViT-based models have been the standout performer in tasks that are more involved than simple everyday image classification, including but not limited to holography (Cuenat and Couturier, 2021), classification of COVID-19 CT scan images (Gao et al, 2021), and identifying distracted driving (Li et al, 2022). Those studies provide an indication that the global-attention aspect of the ViT architecture allows it to make more accurate predictions using information that could be lost within convolution and pooling layers of a CNN. While CNNs are still useful for transportation-related classification tasks where information is locally concentrated in an image such as traffic sign classification (Zheng and Jiang, 2022), studies that evaluated vision models for transportation applications where the information needed to make accurate predictions are expected to be sparse across the image show the considerable benefits of using ViT. This includes the work of Liang et al (2022) demonstrating the ability of ViT in detecting driver distraction, and a study by Abdelraouf et al (2022) who demonstrated the ability of a ViT-based architecture to detect rain and roadway surface conditions with an impressive F-1 score of up to 98%.

Computer vision methods and models have proven to add tremendous value to different fields of science and practice. There remains, however, a significant gap between the abundance of available transit data sources and the integration of these data sources with methods and frameworks to extract complementary information through computer vision. Such integration can provide a significant boost to the state of the practice. The authors believe that this gap is, by and large, due to the demanding process of computer vision models' data acquisition, labeling, and training. There is an immense need for developing generalizable and transferable frameworks that can seamlessly integrate these existing transit and urban data sources, and streamline the

data acquisition, model training, and inference processes. We propose TranViT as a pioneering example for this much-needed systems integration task.

3 Methodology

3.1 Site of Study

The site of this exploratory study was near Central Square in Cambridge, MA, USA, at the intersection of Massachusetts Avenue and Sidney Street. We selected this site due to the availability of a public IP (Internet Protocol) camera that streams a live view of this area, which is served by multiple Massachusetts Bay Transportation Authority (MBTA) bus routes; namely Route 1, Route 70, and Route 64 with, respectively, 10, 15, and 45-minute weekday headways. The Google Earth view of the site is illustrated in Figure 1. The north direction is indicated by the arrow at the top-right of the figure. The top-left shows the position of the public camera that was used to acquire images for this study, with the yellow lines showing the field of view of the camera. The bus stops at the bottom-right quarter of the image serve either direction of the MBTA Route 1 bus operating between Nubian and Harvard Square. The inbound (to Boston) and outbound (to Harvard Sq) directions of the route are respectively illustrated by the red and blue arrows. Figure 2 shows the Google Maps view from the camera angle, alongside the true camera perspective of the site.

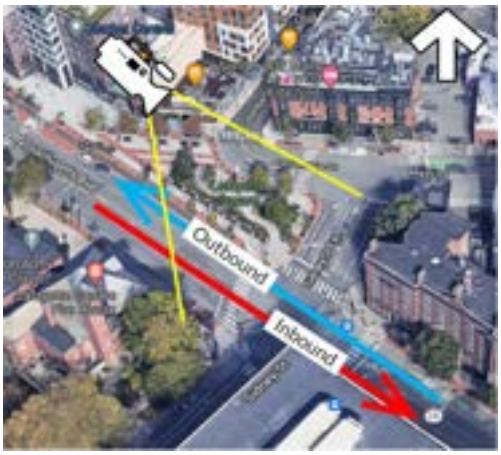


Fig. 1: Site of study near Central Square in Cambridge, MA, USA.

3.2 Proposed Framework, Data Collection and Pre-processing, and ViT Model Training

3.2.1 Data Sources and Training Data Acquisition

Figure 3 illustrates the data sources and the modules of our proposed framework. The attributes of data acquired from each source are described in Table 1. At the core of our proposed TranViT framework is the General Transit Feed Specification real-time component (GTFS-RT). The GTFS-RT is available for the MBTA vehicles through onboard GPS equipment which updates and publishes vehicle location, heading, and occupancy data in real time at high-frequency intervals (up to every 1 second) (Massachusetts Bay Transportation Authority, 2022). We utilize this high-frequency of

(a) Google Maps view. (b) Camera view.

Fig. 2: Camera perspective for the site of study.

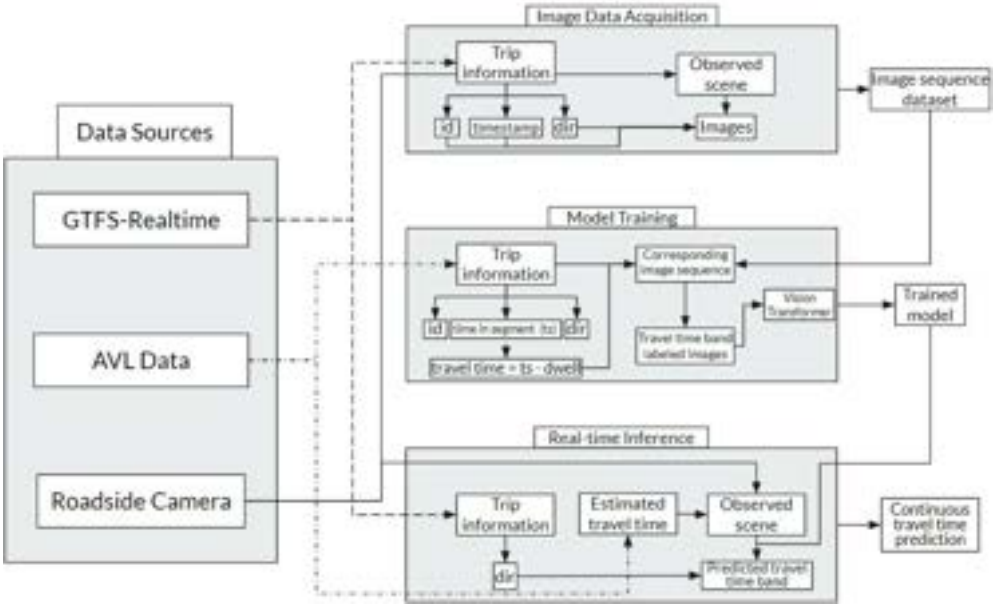


Fig. 3: TranViT data sources, modules, and workflow.

the GTFS-RT to trigger image acquisition *only* as transit vehicles approach the area of the study, which accomplishes the following:

1. Enables linking the acquired image sequences and the travel times associated with the trip ID that activated the image acquisition.
2. Optimizes the image acquisition process, ensuring data is only collected as needed.

Image acquisition from camera livestream at the site is activated once a transit vehicle is approaching (within 500m). A total of six images are acquired for each activation, with a 15-second wait time between the images to allow for traffic movement across the intersection. If more than one transit vehicle approaches and activates the camera

at the same time, a single acquisition stream is initiated and separate images will be labeled by the trip ID and direction of all transit vehicles (with no maximum limit) that cross the area during the given timeframe. The outputs of the data acquisition process are a database containing the trip IDs, directions, and the timestamp of the transit vehicles' approach to the site of study, and an image dataset with six images associated with each trip ID. The data acquisition process for this pilot study is summarized by the following:

- Three separate rounds of data acquisition (to account for the seasonal variations in daylight and roadside greenery).
  - Feb 23$^{\text{rd}}$ - March 4$^{\text{th}}$, 2022
  - March 28$^{\text{th}}$ - April 6$^{\text{th}}$, 2022
  - May 9$^{\text{th}}$ - May 16$^{\text{th}}$, 2022
- Data was collected for each day between 6 AM - 9 PM.
- Transit trip data is only recorded if the image acquisition process is successful. Livestream buffering and connection issues can result in failed acquisitions.
- A total of 2,992 MBTA Route 1 bus trips were recorded out of 3,013 trips in AVL records for the same time period (99.3%).
- A total of 17,905 images were acquired and associated with these trips.

Table 1: Data Sources and Attributes Used

Source	Attribute	Type	Description
GTFS-RT	id	Integer	Unique trip identifier.
	dir	Binary	Trip Direction (0 = outbound, 1 = inbound).
	timestamp	Integer	Unix timestamp during image acquisition.
AVL	ts	Float	Total travel time across the segment (sec).
	dwell	Float	Stop dwell time for a given trip (sec).
Camera	site image	1280 x 720 x 3 Array	Image of the observed scene.

3.2.2 Generating Travel Time Bands Based on Effective Travel Time Percentiles

The acquired trip-image dataset requires pre-processing to make it suitable for the ViT image classification task. First, the overall travel time of transit vehicles is acquired from the MBTA's AVL database. The AVL database also keeps a record of transit vehicles' stop events, including the associated dwell time with a stop event (if any). We associate the trip ID for each transit vehicle in our dataset with the stop events at either of the two stops in our site of study (in the inbound and outbound direction) and define the effective travel time as the time spent by the vehicles on the 1-km segment (which includes camera monitored area in addition to the 500m activation buffer in either direction), minus any dwell time associated with that trip at its respective stop on the site.

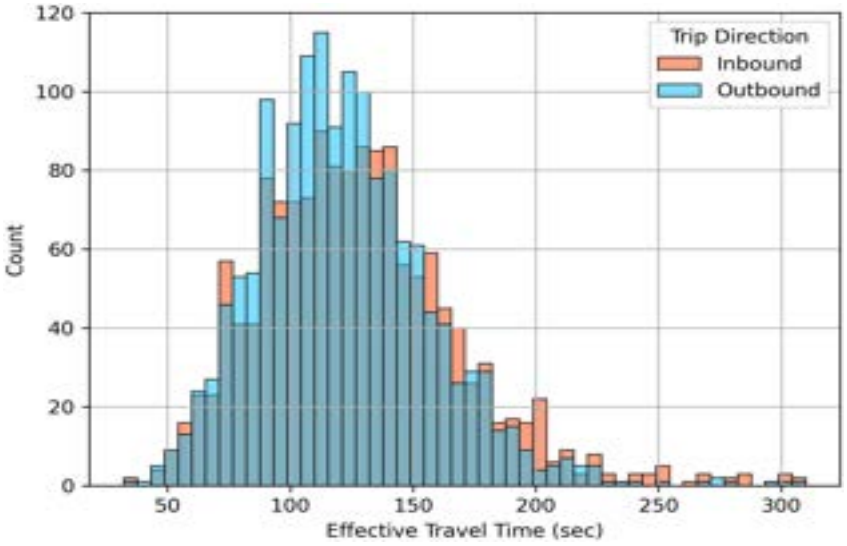
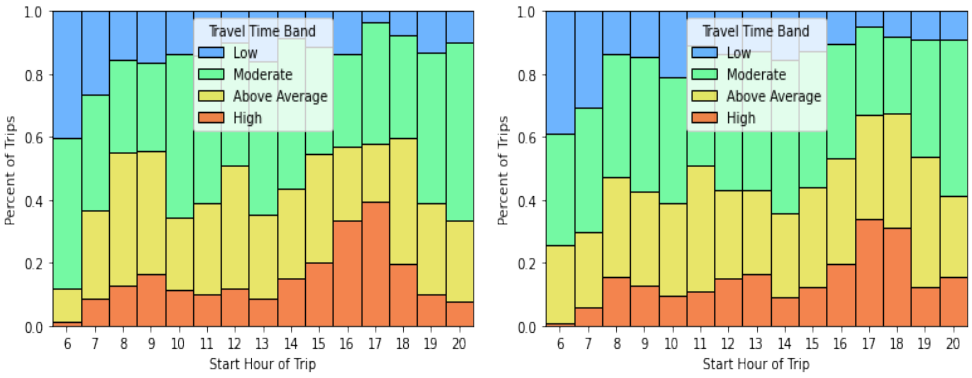


Fig. 4: Observed effective transit travel time for training dataset.

We use this effective travel time across the segment as a ground truth label, with the assumption that after accounting for stop events (if any) a transit vehicle's travel time across a segment is representative of the overall traffic. This assumption, however, does not take into account the impact of signal control due to the unavailability of signal time data. The signal, however, was observed to operate a 180-second cycle length, with an 85-15 split (25 seconds green for the side street) which considerably favors the street on which Route 1 transit vehicles run. The observed effective travel time for the dataset is illustrated in Figure 4. Figure 5 shows the distribution of these travel time bands per trip direction and hour of the day for the acquired image sequence dataset. It is worth noting that the dataset contains more outbound (1,536) than inbound (1,456) trips.



(a) Inbound. (b) Outbound.

Fig. 5: Travel time band distribution for acquired image dataset.

The overall average effective travel time observed was 124 seconds, with a minimum of 35 seconds and a maximum of 310 seconds. The descriptive statistics are shown in Table 2. To normalize the labels for the training dataset, we utilize a four-class labeling approach to discretize observed effective travel time percentile and label them as follows:

- Effective Travel Time $\leq 10\% \rightarrow$ Low.
- $10\% <$ Effective Travel Time $\leq 50\% \rightarrow$ Moderate.
- $50\% <$ Effective Travel Time $< 90\% \rightarrow$ Above Average.
- Effective Travel Time $\geq 90\% \rightarrow$ High.

Table 2: Descriptive Statistics for Effective Travel Time

Statistic	Trip Direction		
	Overall	**Inbound**	**Outbound**
μ	124	127	121
σ	38	41	35
10%	79	79	79
50%	121	124	118
90%	160	166	156
Min	35	35	35
Max	310	309	310
Count	2,992	1,456	1,536

3.2.3 ViT Model Training and Fine-tuning

The ViT model utilized as a part of our framework is the one proposed by Dosovitskiy et al (2020). The gist of the model is that it simplifies the pixel-wise attention calculation that would take place in a transformer's encoder module by splitting the image into N fixed-size patches (P). Those fixed-size patches are linearly inserted into the transformer encoder alongside their positional embeddings, which simply tell the encoder where each patch belongs in an image. Meaning that an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is reshaped into $\mathbf{x}_{\mathrm{p}} \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where H, W, C are, respectively, the height (720 pixels), width (1280 pixels), and the number of color channels (3) in the source image. A constant latent vector (D) is used across all layers, and learnable positional embedding parameters ($\mathbf{E}_{lin}$, $\mathbf{E}_{pos}$) are utilized to extract an embedding token ($\mathbf{z}_o$). Flattened embeddings pass through a series of encoders (L) each with a multi-head self-attention (MSA) layer, and a feed-forward Multi-Layer Perceptron (MLP), both preceded by normalization layers (LN). This process is mathematically described by the following:

$$\mathbf{z}_o = [\mathbf{x}_{class};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \ldots\ ;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}_l^{'} = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \qquad l = 1 \ldots L \tag{2}$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}_l^{'})) + \mathbf{z}_l^{'}, \qquad l = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

Where $\mathbf{y}$ is the final image representation out of the transformer encoder, then passed into a classification MLP outputting the most likely class for the image, which in our case is the travel time band prediction based on the effective travel time percentile. We start with the base ViT model pre-trained on the ImageNet-21K dataset and open-sourced by Google Research (Dosovitskiy et al, 2020). To accommodate for the data-hungry nature of ViT, prior to re-training on our data we increase the size of the dataset by creating an image augmentation pipeline; randomly cropping, tilting, and slightly adjusting the brightness and contrast of images acquired in our dataset as shown in Table 3, while maintaining the original images' class label based on effective travel time rank. Every image is passed through the pipeline six times, and the probability of an augmentation action is the probability that it is performed on the image during an iteration in a random magnitude within action bounds. This results in a final dataset with 78,385 images.

Table 3: Image Augmentation Pipeline Parameters

Action	**Augmentation Bounds**		**Probability**
	Lower	**Upper**	
Crop	560 × 560	1280 × 720	0.33
Rotate	-30 °	+30 °	0.33
Brightness	-20%	+20%	0.50
Contrast	-20%	+20%	0.50

We initially train multiple instances of the ViT model while performing a grid search to optimize the model's hyper-parameters to maximize the overall F-1 Score. The bounds and optimal values of this fine-tuning process are shown in Table 4. After finding optimal hyper-parameters, the vision transformer was trained using a 5-fold cross-validation, with an 80-20 stratified training-testing split for each fold. We further evaluate the model in terms of precision, recall, and accuracy.

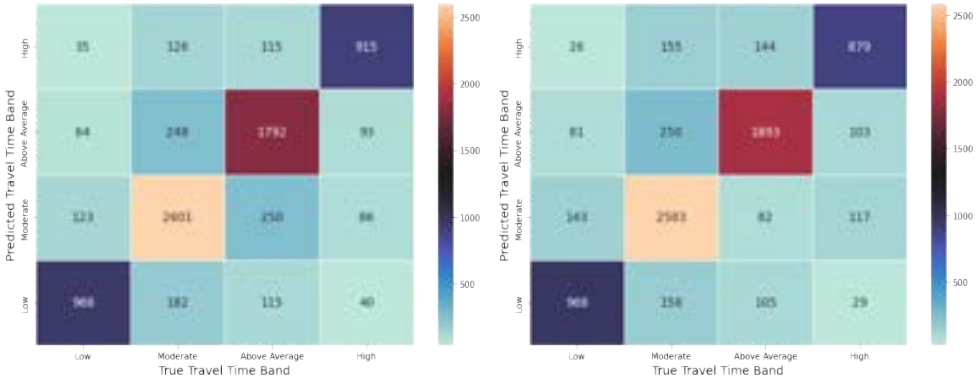
Table 4: ViT Parameter Fine-Tuning

Parameter	**Parameter Limits**		**Optimal**
	Lower	**Upper**	
Hidden Layers	2	12	12
Attention Heads	2	12	12
Batch Size	8	256	32
Learning Rate	$1e^{-6}$	$1e^{-2}$	$2e^{-5}$
Dropout Probability	0	0.25	0.10

4 Results and Discussion

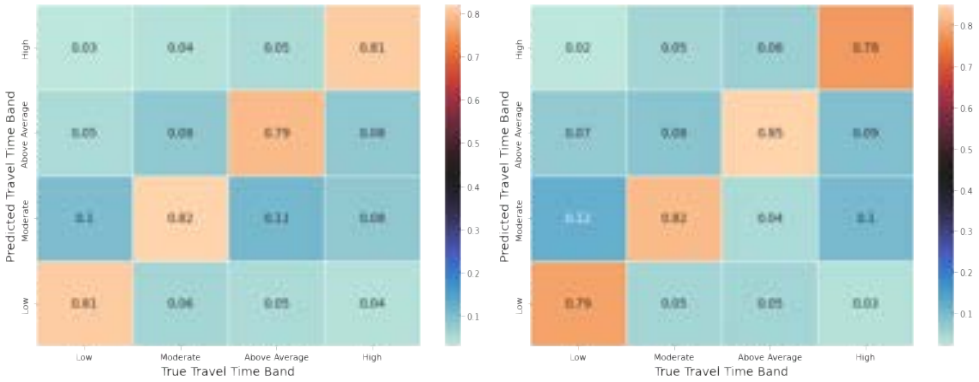
4.1 ViT Performance on Effective Travel Time Band Prediction

After training on 80% of the images in the k-fold augmented dataset, the averaged 5-fold performance of the optimized ViT model on the test sets for either direction is shown in the confusion matrices in Figure 6 below, illustrating the true and predicted travel time band for test images. Figure 7 illustrates the results normalized to the total number of true travel time band images in the test set (sum over columns = 1, diagonal cells represent accuracy for each travel time band).



(a) Inbound. (b) Outbound.

Fig. 6: Confusion matrices for test subsets.



(a) Inbound. (b) Outbound.

Fig. 7: Normalized confusion matrices for test subsets.

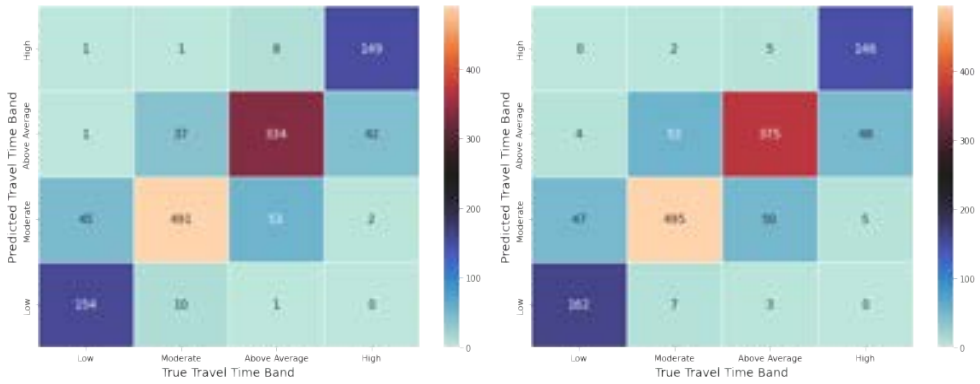
The results indicate that the model is able to differentiate the discrete expected travel time ranges based on the observed effective transit travel time used for labeling the dataset. While a higher number of misclassifications happen between the intermediate states (moderate and above-average conditions), the misclassification rate between the extreme conditions (low and high) is very low. This result is of utmost importance,

indicating that when using image data in complementing the travel time prediction process, the misclassifications will rarely lead to an extreme-end under or overestimation. Detailed results on the ViT model's performance are provided in Table 5, showing the 5-folds average for each metric ± the range of variation for each metric across the 5-folds.

Table 5: 5-Fold Cross-Validation Test Sets Performance Metrics

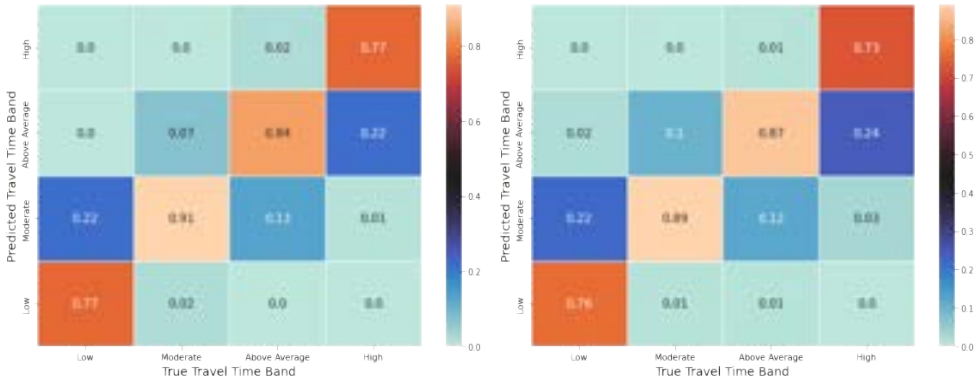
Class	Inbound			Outbound		
	Prc (±)	**Recall** (±)	**F-1** (±)	**Prc** (±)	**Recall** (±)	**F-1** (±)
Low	0.74 (0.05)	0.81 (0.02)	0.78 (0.01)	0.76 (0.02)	0.80 (0.02)	0.79 (0.03)
Normal	0.84 (0.01)	0.83 (0.02)	0.83 (0.01)	0.82 (0.01)	0.82 (0.01)	0.82 (0.01)
Above Average	0.81 (0.01)	0.78 (0.01)	0.80 (0.00)	0.81 (0.01)	0.79 (0.02)	0.79 (0.01)
High	0.77 (0.02)	0.81 (0.02)	0.79 (0.01)	0.74 (0.02)	0.79 (0.03)	0.75 (0.03)
Accuracy		**0.81** (0.01)			**0.80** (0.01)	

Given that the ViT training and inference are based on images that provide a constrained view of the area of study, variation in the number of vehicles and other factors between images taken from the same sequence can occur (6 images at 15-second intervals, as detailed in the Methodology). To account for this variation, we run inference on image sequences for a given trip ID instead of single images. Images from the test set of the best-performing fold were grouped by their trip ID before making travel time band predictions based on the average of all predictions for the image sequence. The results are illustrated in Figures 8 and 9. Classification metrics are detailed in Table 6, where support is the number of actual occurrences of the class in the test subset.



(a) Inbound. (b) Outbound.

Fig. 8: Confusion matrices using averaged image sequence score.



(a) Inbound. (b) Outbound.

Fig. 9: Normalized confusion matrices using averaged image sequence score.

Table 6: Performance Metrics Using Averaged Image Sequence Score

Class	Inbound				Outbound			
	Prc	**Recall**	**F-1**	**Support**	**Prc**	**Recall**	**F-1**	**Support**
Low	0.93	0.77	0.84	201	0.94	0.76	0.84	213
Moderate	0.83	0.91	0.87	539	0.83	0.89	0.86	557
Above Average	0.81	0.84	0.82	396	0.78	0.87	0.82	433
High	0.94	0.77	0.85	193	0.95	0.73	0.83	199
Accuracy	**0.85**				**0.84**			

The results in Table 6 demonstrate significant improvements in model performance when using image sequences, particularly in nearly eliminating all misclassifications between non-consecutive travel time ranges. While the accuracy (the number of travel time band labels that were correctly classified divided by the total occurrences of the class label in the training dataset) slightly drops for the under-represented classes (low and high travel time bands), the precision for those classes increases significantly. Given that precision is an indicator of the quality of the prediction (the number of true positives divided by the total number of class predictions made by the model) a precision of over 93% for those extreme classes is of utmost importance when those labels are to be used to enhance travel time predictions. The precision, recall, F-1 score, and accuracy all increase for the moderate and above-average travel time band labels.

Of the 2,992 trip-image sequences in the dataset acquired for this study, 2,731 transit trip IDs had two or more images in the test set of the best-performing training fold. The effect of the number of images in a sequence is illustrated in Figure 10. The line plots show the average classification accuracy for each travel time band, while the envelopes illustrate the 95^{th} confidence interval for a given number of images in sequence. While accuracy and confidence increase as the number of images used for prediction increases, particularly for under-represented classes (the low and high travel time bands), a decline is observed after the number of images in a sequence exceeds six. This is attributable to the original number of images for each trip-image-sequence

acquisition being six images, hence exceeding that number indicates the presence of augmented images in a test sequence that could make it more challenging to classify.

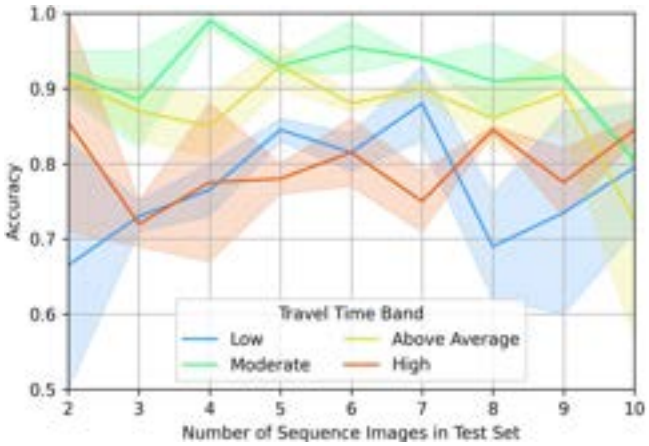


Fig. 10: Accuracy versus image count in a test sequence.

Figure 11 illustrates the variation in travel time band prediction accuracy for both directions during different times of the day at the trip-images-sequence level (i.e. the percent of images accurately labeled within for unique trip-images-sequence). Consistent with previously discussed results, the model's performance for the inbound direction outperforms the outbound. The model was found to learn and make better predictions during the AM and PM peak hours for both directions compared to the off-peak hours. This may be attributed to the apparent variation in traffic volumes observed during peak hours which is a good indicator of the expected travel time across the monitored segment during these times. In addition to that, the higher frequency of public transit vehicles during those times resulted in a higher volume of ground-truth image acquisitions, which improved the training process of images acquired during those times. Higher travel times can occur during off-peak hours without the presence of high traffic due to traffic control, driving behavior, or vehicles impeding access to bus stops. Such factors can not be identified from the image data used in this study.

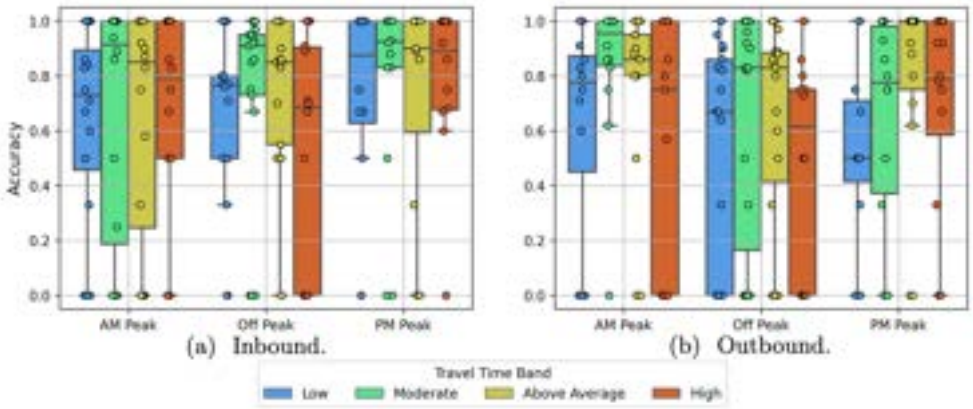
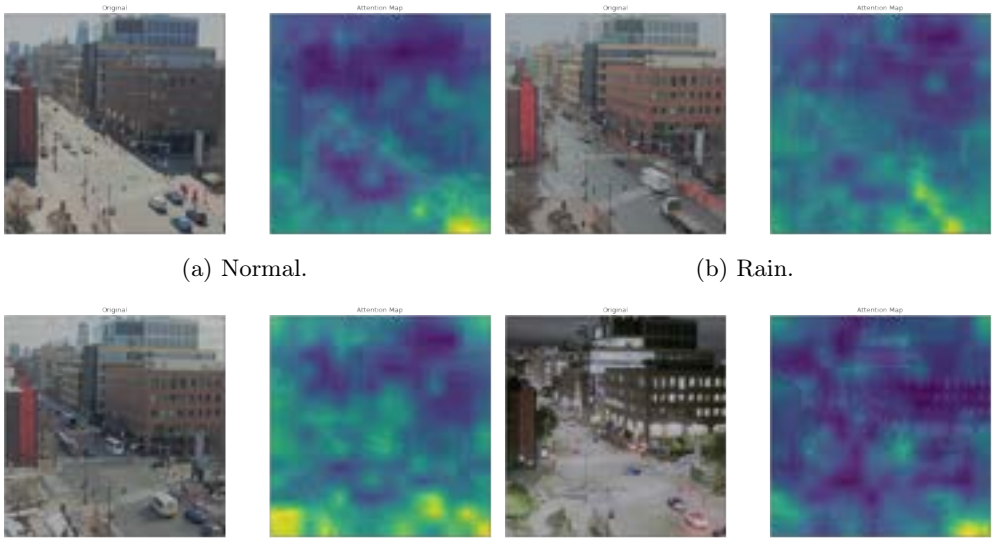


Fig. 11: Variation in classification accuracy by direction and time of day.

Next, we assess the interpretability of the results obtained by the ViT. This is accomplished by mapping the averaged attention scores (values between 0 to 1) of the model's 12 attention heads. Attention illustrates the parts of the image from which the model learns to gain the most information in making the predictions for travel time bands. Higher scores are illustrated by brighter pixels. Figures 12 and 13 illustrate the attention maps for different scenarios for the inbound and outbound directions.



(a) Normal. (b) Rain.

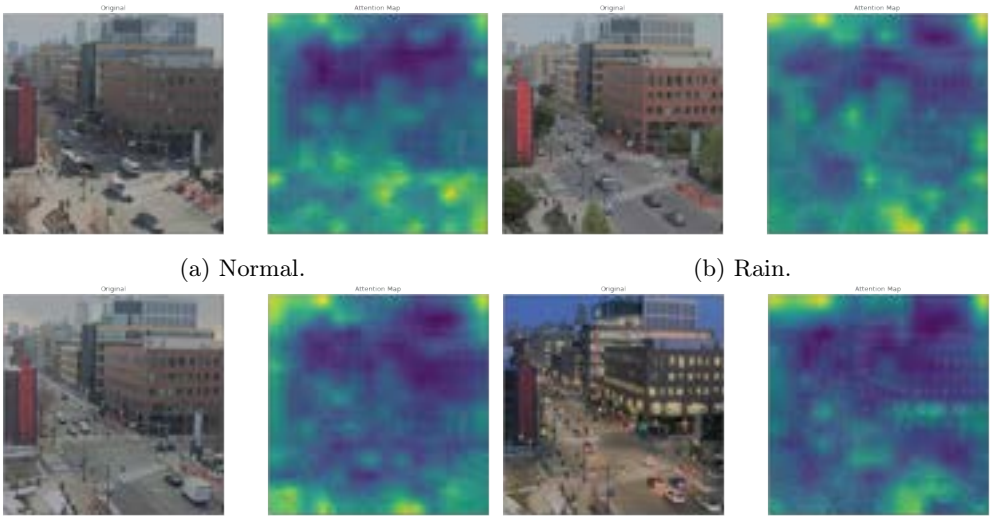
(c) Snow. (d) Night.

Fig. 12: Inbound direction attention maps across different conditions.

We observe the ViT model learning to make predictions in a very logical and sensible manner, where a lot of attention is directed to the pixels containing information about vehicles traveling in the direction of interest. For the inbound direction illustrated in Figure 12, we see the majority of attention being directed to the right side of the image, mostly to the vehicle in the queue in the inbound direction (when present), but also towards the downstream. The ViT learns to look in the opposing direction with lower attention while learning to ignore the buildings that provide little to no information (although attention to buildings' indoor lights was observed, the model might use it as an indicator for the time of day). Other sensible actions illustrated by the ViT include paying attention to the snow when there is a snowfall, which impacts the travel speeds across the segment. In Figure 12d, it can be seen that at night the transformer learns to include information from the inbound direction traffic light which becomes visible from the camera's vantage point in the evening.

We observed the ViT model exhibiting similar behavior in the outbound direction in Figure 13. The camera view for the outbound direction, however, is a lot more limited compared to that of the inbound, as can be seen with the building on the left of the image blocking the upstream, and a very limited view of the downstream. This explains the better predictive performance for the model in the inbound direction previously presented in the evaluation tables and confusion matrices. Aside from learning to look at vehicles and their locations, the model was observed to pay more attention to the sky and buildings' lights during the night. This behavior seems to be an attempt to compensate for the lack of information induced by the restricted camera view for the

outbound direction by making some form of time-of-day inference. The extent to which image resolution and/or combinations of extreme weather (i.e. heavy night precipitation, morning mist, etc.) affect the prediction accuracy was not explored in this study.



(a) Normal. (b) Rain.

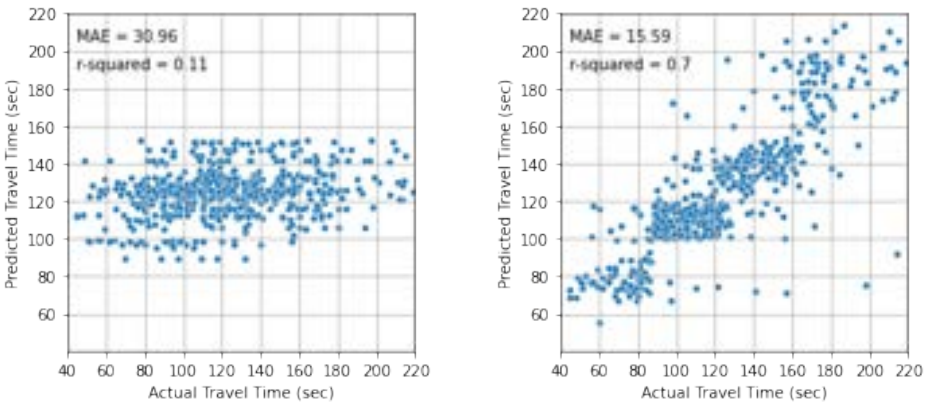
(c) Snow. (d) Night.

Fig. 13: Outbound direction attention maps across different conditions.

4.2 Implications for Travel Time Prediction

We conclude this assessment with a brief proof-of-concept looking into what this use of real-time, computer-vision-based travel time class predictions could mean to the broader task of travel and arrival time estimation. We fit a linear regression model to predict the effective segment travel time of the 2,731 vehicles in the image sequence test set based on the recorded information for these vehicles which was obtained from the GTFS real-time component. The linear regression model (Ordinary Least Squares, OLS) makes baseline travel time predictions based on the time of day, the direction of travel, and the occupancy of the transit vehicle as it approaches the segment. Occupancy is a continuous variable showing the percentage of passengers estimated from on-bus automatic passenger counters to total seating capacity, ranging from 0 - 150%, with $\mu = 34.6$ and $\sigma = 26.67$. Hours of the day were encoded as binary variables. We then run the same model, with the addition of a predicted travel time band label (OLS+) obtained from the inference of the images associated with the transit vehicle's approach to the segment. The predicted travel time bands are denoted by the TTB variables, with the moderate band (TTB_Mod) as the baseline. The results of actual versus predicted travel times are illustrated in Figure 14.

Table 7 shows the coefficient estimates for the different models. Hours of the day that were not found significant for any model were dropped from the table for brevity. The model-agnostic addition of the travel time band labels predicted from images works to create bounds that enhance the continuous travel time estimation. This demonstratory evaluation shows significant improvement achieved by the linear regression model, both in terms of the predictions' mean absolute error and r-squared. As the data and image sequence acquisition for this study was initiated when a transit vehicle is within 500 meters of the camera, we utilized state predictions from the previous transit



(a) AVL-based linear regression. (b) With image-based band prediction.

Fig. 14: Linear regression fit for predicting transit travel time through a segment.

vehicle traveling in the same direction and obtained comparable results, indicating that the traffic state does not change drastically between consecutive vehicles (10-minute scheduled headway for MBTA Route 1). Future work will investigate the extent to which this look-ahead horizon can provide reliable predictions, and assess the extent of performance improvements to more sophisticated state-of-the-practice models.

Table 7: Ordinary Least Squares Regression Model Estimates

Variable	Type	Inbound		Outbound	
		OLS	**OLS+**	**OLS**	**OLS+**
Occupancy	**Continous**	0.23**	0.04	0.02	0.03
Hour_6	**Binary**	-25.11**	0.16	-15.63**	1.93
Hour_7	**Binary**	-12.92**	2.80	-15.79**	2.84
Hour_9	**Binary**	8.53*	4.91*	11.54**	5.76**
Hour_12	**Binary**	6.12	0.29	11.10**	7.15**
Hour_13	**Binary**	1.49	2.50	2.49	5.40*
Hour_14	**Binary**	-2.47	3.77	7.42**	7.91**
Hour_16	**Binary**	21.79**	11.69**	7.04**	3.69
Hour_17	**Binary**	24.85**	6.46*	22.47**	8.03**
Hour_18	**Binary**	25.10**	11.57**	37.19**	15.88**
Hour_19	**Binary**	3.11	4.57	18.18**	9.64**
Hour_20	**Binary**	6.14	10.16**	11.82**	6.05**
Hour_21	**Binary**	-8.35	1.24	9.87**	8.71**
TTB_Low	**Binary**	-	-33.87**	-	-30.68**
TTB_Aav	**Binary**	-	36.73**	-	29.80**
TTB_High	**Binary**	-	83.32**	-	70.81**
Intercept		*116.55*	*104.63*	*112.58*	*99.87*
R-Square		*0.111*	*0.693*	*0.128*	*0.680*

Coefficients denoted with "**" are significant at $p < 0.05$ level; "*" are significant at $p < 0.10$.

The figures in the Appendix of this paper briefly illustrate the broader implications of improving travel time prediction for a specific camera-monitored segment and how it relates to overall transit trip duration estimation. This is specifically demonstrated in terms of travel time variability relative to the median travel time at the trip and monitored segment level, both obtained from AVL data. It can be observed that the relative travel time at the segment explains 11% of the variability in the overall trip duration. While this area of study was pre-determined due to the availability of video data, this approach can be used as a screening mechanism to identify ideal sites for image sensor installation, such that improving travel time prediction for select feasible sites could substantially improve the overall prediction of trip duration. Our analysis identified another segment of similar length in Massachusetts Avenue where the relative segment travel time explains up to 45% of the total trip duration variability.

5 Conclusions

This study provided a detailed implementation and assessment for TranViT, an integrated framework utilizing a combination of traditional transit data sources with roadside computer vision to improve real-time transit travel time prediction. An exploratory assessment of our proposed framework was conducted for a segment of Massachusetts Avenue in Cambridge, MA, USA, with the results providing evidence for the potency of this framework. First, we introduce a workflow for automated roadside image data acquisition and labeling utilizing traditional transit data sources. Second, we train and thoroughly evaluate the ViT component of the framework which was able to successfully learn image features and contents that best help it deduce the expected travel time range across the segment of interest, with a validation accuracy ranging between 80%-85%, and a precision of up to 95%. Finally, we demonstrate how this prediction of the travel time range can subsequently be utilized to improve continuous travel time prediction. This study demonstrates the added value of creating end-to-end, scalable, automated, and highly efficient approaches integrating traditional transit data sources and roadside imagery to extract real-time information, providing a blueprint for transit agencies and practitioners to address a fundamental gap impeding the broader utilization of computer vision in transit operations. Such integration can be extended to numerous use cases in transit as an agency's data and assets permit (e.g. dwell time estimation, platform crowding, left-behind passengers, etc.) to improve operational and passenger real-time information.

Unlike existing studies utilizing computer vision for transportation applications, our framework does not require having certain data in place. On the contrary, the major contribution of this work is creating a generalized workflow for acquiring and labeling the data for the computer vision tasks, which could be extended to other use cases (e.g. focusing on areas of images containing bus stops for anticipating the dwell time). Utilizing the GTFS real-time component to initiate transit data and corresponding image sequence acquisitions results in an extremely computationally efficient workflow, running on a single CPU and with no more than 128 MBs of RAM. Based on our observations from this study, we expect a more demanding data acquisition and training task for models to be deployed in areas where the scenery changes noticeably between seasons (like our case in Cambridge, MA). The models we trained in between rounds of data acquisition did not perform well without retraining. The subsequent model re-training, however, is faster. This is attributed to the gigantic number of parameters (86 Million in the base model) that a ViT employs, which requires substantial training to ensure generalizability. The output travel time range labels (based on the percentiles used to create these ranks) can complement any existing travel time prediction algorithm by adding a complementary real-time attribute describing

the currently observed state of traffic at an area of interest, supplementing what models traditionally learn from historic AVL observations. We concluded this study with a proof of concept demonstrating the potential impact of this additional vision-based input for improving transit travel time predictions. In practice, we anticipate that a data-driven deployment of roadside cameras only in locations where high transit delays are observed would be sufficient in providing reliable travel time predictions without the need to monitor large segments of the transit network.

One of the limitations of this study is the need to generate augmented images to satisfy the data-hungry nature of vision-transformed training. Future works will look into acquiring a larger quantity of image data, alongside integrating additional data sources that provide information on signal timing (if the work is conducted in an intersection setting) or adapting the model in ways that account for the impact of traffic control variations on segment travel times. Integrating methods that account for label uncertainty could also help improve the performance of the predictions (Northcutt et al, 2021). The observations noted in the discussion of Figures 12 and 13 indicate some best practices for the camera installation location, where a wider field of view is expected to provide better image data for model training. Another potential improvement is utilizing the observed speed and/or acceleration profiles of the transit vehicles as opposed to their travel time. This could either be accomplished by logging the transit vehicles' location coordinates from GTFS-RT during the data acquisition phase (Huang et al, 2023), or by utilizing computer-vision-based trajectory tracking for all vehicles, which would come at the expense of higher computational load but will allow for the estimation of the true overall traffic speed and state more accurately, leading to a better quality labeling for the training data.

6 Acknowledgements

The authors would like to thank The Massachusetts Bay Transportation Authority (MBTA) for providing data access which enabled this study, and the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing high-performance computing resources that have contributed to the research results reported in this paper.

7 Author Contribution Statement

The authors confirm their contribution to the paper as follows: study conception and design: A.A., J.Z.; data collection: A.A.; analysis and interpretation of results: A.A.; draft manuscript preparation: A.A., J.Z. Both authors reviewed the results and approved the final version of the manuscript. The authors do not have any conflicts of interest to declare.

8 Data Availability

The transit AVL data used in this study has not been made publicly available by the MBTA. Code for image acquisition, model training, and analysis to replicate this study is available from the corresponding author upon reasonable request.

References

Abdelhalim A, Abbas M (2018) Impact assessment of a cooperative bus-holding transit signal priority strategy. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp 1908–1913

Abdelhalim A, Abbas M, Kotha BB, et al (2021) A framework for real-time traffic trajectory tracking, speed estimation, and driver behavior calibration at urban intersections using virtual traffic lanes. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) pp 2863–2868

Abdelhalim AT (2021) A real-time computer vision based framework for urban traffic safety assessment and driver behavior modeling using virtual traffic lanes. PhD thesis, Virginia Tech

Abdelraouf A, Abdel-Aty M, Wu Y (2022) Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways. IEEE Transactions on Intelligent Transportation Systems 23(10):18546–18556

Aemmer Z, Ranjbari A, MacKenzie D (2022) Measurement and classification of transit delays using GTFS-RT data. Public Transport 14:263–285

Buch N, Velastin SA, Orwell J (2011) A review of computer vision techniques for the analysis of urban traffic. IEEE Transactions on Intelligent Transportation Systems 12(3):920–939

Cathey F, Dailey DJ (2003) A prescription for transit arrival/departure prediction using automatic vehicle location data. Transportation Research Part C: Emerging Technologies 11(3-4):241–264

Chen H, Rakha HA, Sadek S (2011) Real-time freeway traffic state prediction: A particle filter approach. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE, pp 626–631

Cuenat S, Couturier R (2021) Convolutional neural network (cnn) vs visual transformer (vit) for digital holography. arXiv preprint arXiv:210809147

Dilek E, Dener M (2023) Computer vision applications in intelligent transportation systems: a survey. Sensors 23(6):2938

Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929

Elliott T, Lumley T (2020) Modelling the travel time of transit vehicles in real-time through a gtfs-based road network using GPS vehicle locations. Australian & New Zealand Journal of Statistics 62(2):153–167

Gaikwad N, Varma S (2019) Performance analysis of bus arrival time prediction using machine learning based ensemble technique. In: Proceedings 2019: Conference on Technologies for Future Cities (CTFC)

Gao X, Qian Y, Gao A (2021) COVID-ViT: Classification of COVID-19 from CT chest images based on bision transformer models. arXiv preprint arXiv:210701682

Ge L, Sarhani M, Voß S, et al (2021) Review of transit data sources: potentials, challenges and complementarity. Sustainability 13(20):11450

Gokasar I, Timurogullari A (2021) Real-time prediction of traffic density with deep learning using computer vision and traffic event information. 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) pp 1–5

Han Q, Liu K, Zeng L, et al (2020) A bus arrival time prediction method based on position calibration and LSTM. IEEE Access 8:42372–42383

Huang Y, Abdelhalim A, Stewart A, et al (2023) Reconstructing transit vehicle trajectory using high-resolution GPS data. arXiv preprint arXiv:230515545

Janai J, Güney F, Behl A, et al (2020) Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision 12(1–3):1–308

Jenelius E, Koutsopoulos HN (2013) Travel time estimation for urban road networks using low frequency probe vehicle data. Transportation Research Part B: Methodological 53:64–81

Jeong R, Rilett LR (2005) Prediction model of bus arrival time for real-time applications. Transportation Research Record 1927(1):195–204

LeCun Y, Bottou L, Bengio Y, et al (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Li Y, Wang L, Mi W, et al (2022) Distracted driving detection by combining vit and cnn. In: 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, pp 908–913

Liang J, Zhu H, Zhang E, et al (2022) Stargazer: A transformer-based driver action detection system for intelligent transportation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3160–3167

Lin HE, Zito R, Taylor M, et al (2005) A review of travel-time prediction in transport and logistics. Proceedings of the Eastern Asia Society for transportation studies 5:1433–1448

Massachusetts Bay Transportation Authority (2022) GTFS Documentation. URL https://github.com/mbta/gtfs-documentation/

Northcutt C, Jiang L, Chuang I (2021) Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research 70:1373–1411

Pang J, Huang J, Du Y, et al (2018) Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network. IEEE Transactions on Intelligent Transportation Systems 20(9):3283–3293

Park Y, Mount J, Liu L, et al (2020) Assessing public transit performance using real-time data: spatiotemporal patterns of bus operation delays in columbus, ohio, usa. International Journal of Geographical Information Science 34(2):367–392

Samal C, Sun F, Dubey A (2017) Speedpro: A predictive multi-model approach for urban traffic speed estimation. In: 2017 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, pp 1–6

Sayed T, Zaki MH, Autey J (2013) Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis. Safety science 59:163–172

Shalaby A, Farhan A (2004) Prediction model of bus arrival and departure times using AVL and APC data. Journal of Public Transportation 7(1):3

Sipetas C, Keklikoglou A, Gonzales EJ (2020) Estimation of left behind subway passengers through archived data and video image processing. Transportation Research Part C: Emerging Technologies 118:102727

Tageldin A, Sayed T, Zaki MH, et al (2014) A safety evaluation of an adaptive traffic signal control system using computer vision. Advances in transportation studies

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30

Wang Y, Papageorgiou M (2005) Real-time freeway traffic state estimation based on extended kalman filter: a general approach. Transportation Research Part B: Methodological 39(2):141–167

Wang Y, Yu P (2021) A fast intrusion detection method for high-speed railway clearance based on low-cost embedded GPUs. Sensors 21(21):7279

Wang Y, Papageorgiou M, Messmer A (2008) Real-time freeway traffic state estimation based on extended kalman filter: Adaptive capabilities and real data testing. Transportation Research Part A: Policy and Practice 42(10):1340–1358

Work DB, Tossavainen OP, Blandin S, et al (2008) An ensemble kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In: 2008 47th IEEE Conference on Decision and Control, IEEE, pp 5062–5068

Yang JS (2005) Travel time prediction using the GPS test vehicle and kalman filtering techniques. In: Proceedings of the 2005, American Control Conference, 2005., IEEE, pp 2128–2133

Yildirimoglu M, Geroliminis N (2013) Experienced travel time prediction for congested freeways. Transportation Research Part B: Methodological 53:45–63

Yu B, Yang ZZ, Chen K, et al (2010) Hybrid model for prediction of bus arrival times at next station. Journal of Advanced Transportation 44(3):193–204

Yu B, Lam WH, Tam ML (2011) Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C: Emerging Technologies 19(6):1157–1170

Zeng X, Zhang Y, Balke KN, et al (2014) A real-time transit signal priority control model considering stochastic bus arrival time. IEEE Transactions on Intelligent Transportation Systems 15(4):1657–1666

Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies 58:308–324

Zheng Y, Jiang W (2022) Evaluation of vision transformers for traffic sign classification. Wireless Communications and Mobile Computing 2022

Zhou X, Dong P, Xing J, et al (2019) Learning dynamic factors to improve the accuracy of bus arrival time prediction via a recurrent neural network. Future Internet 11(12):247

Appendix: Relationship between segment travel time and total trip duration.

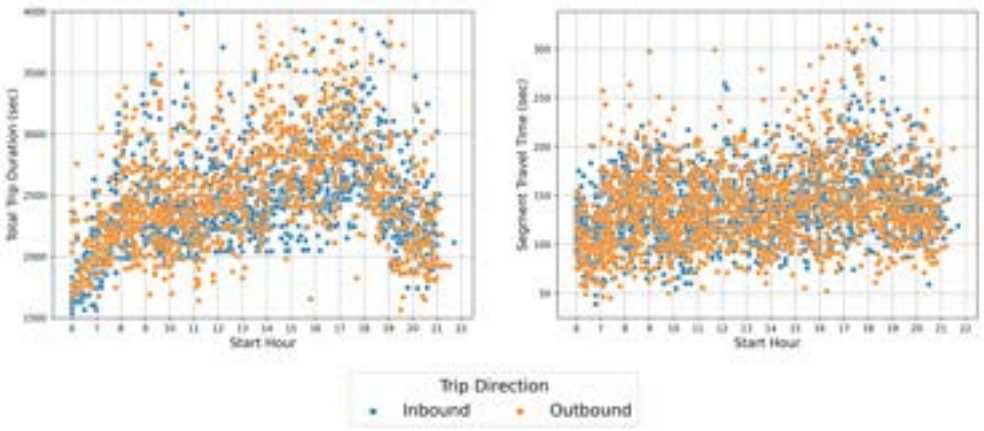


Fig. 1: Temporal distribution of the total trip duration on MBTA Route 1 and travel time across the segment of the study.

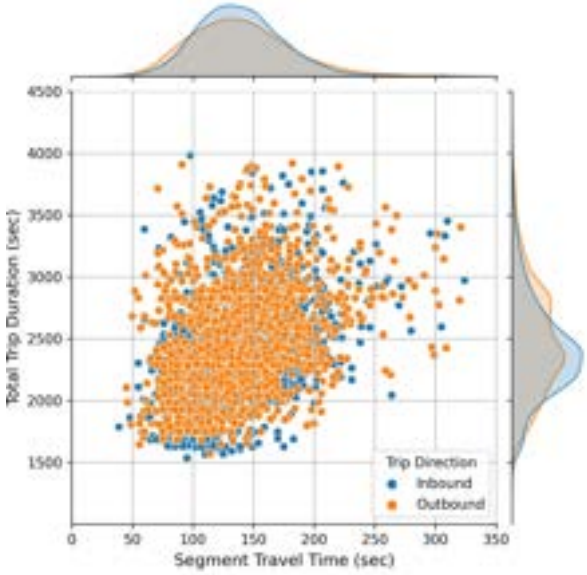
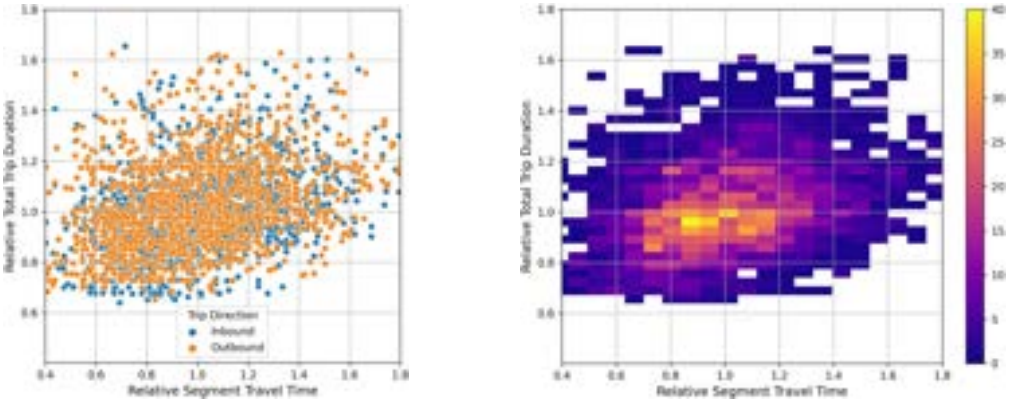
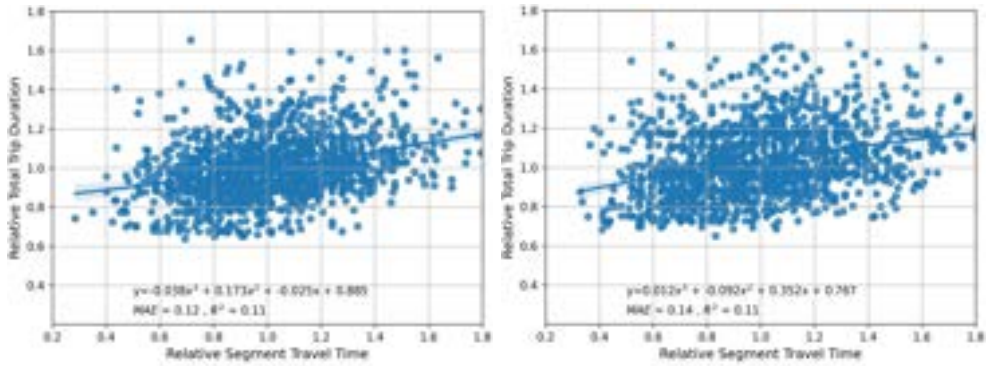


Fig. 2: Relationship between the total trip duration and travel time across the segment of the study.



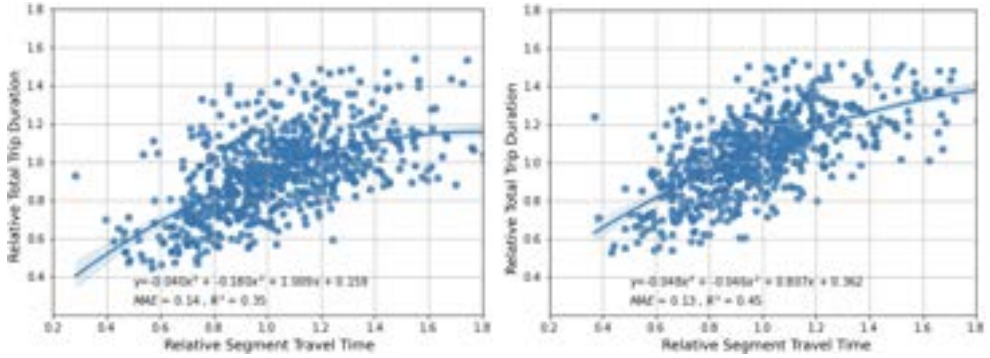
(a) Observed (b) Frequency Count

Fig. 3: Distribution and frequency of observed travel times relative to the median.



(a) Inbound (b) Outbound

Fig. 4: Regression of relative trip duration on relative travel time (segment of study).



(a) Inbound (b) Outbound

Fig. 5: Regression of relative trip duration on relative segment travel time (ideal segment candidate).