CS231N Final Report
Snowy Road Classification with a Data-Centric Approach

Joanne Zhou
Stanford University
joannezhou@stanford.edu

Li Tian
Stanford University
lii@stanford.edu

Brian Hill
Stanford University
bwhill@stanford.edu

Guangyuan Pan
Linyi University
g5pan@uwaterloo.ca *

Abstract

We explored the state of art computer vision neural network models to classify winter road conditions by snow coverage level (barely, partially and fully). We implemented the following improvements to previous works: (1) split the train and test data by camera stations to prevent test data leakage (2) apply more recently developed model architectures and (3) employ subclass labeling to improve the learning capacity of the model.

We experimented with pre-trained and non pre-trained versions of ResNet, DenseNet and Vision Transformer. After testing different numbers of layers fine-tuned, and performing a random hyperparameter search, we reached a best test accuracy of 0.8518 with a DenseNet-121 model, pre-trained on ImageNet. Our model achieved similar performance with the previous work but without data leakage between the train and test data. To further improve test performance, we added daytime/nighttime labels to the images, and trained a six-class classifier before evaluating its performance on the original task. It was found that this data-centric approach did not result in significantly higher accuracy, but provided a more balanced classification performance across different snow coverage levels.

1. Introduction

After a winter trip to Alaska, we were inspired to develop and apply computer vision architectures to recognize winter road conditions to improve driving safety. The difficulty of the classification problem is multi-fold. The ideal model should not only recognize snow coverage in a variety of road context, but also differentiate the levels of coverage, during both daytime and nighttime. Traditional image recognition technology currently cannot achieve the fast real-time high-accuracy performance necessary for road recognition in intelligent and safe driving. Deep learning models have emerged as promising tools to achieve this performance, yet most of the architectures currently applied to this problem are outdated by modern standards.

We explored a variety of CNN models (ResNet, DenseNet, Vision Transformer) with fine-tuned parameters, to find the best performing model for detection of different levels of snow-coverage (barely covered, partly covered and fully covered) in images captured by roadside cameras. We conduct a literature review on previously published work tackling similar road condition classification problems, as well as the various existing computer vision architectures for image classification. We then introduce the Road Weather Information System (RWIS) dataset used in this project, obtained from a research group from the University of Waterloo, and explain the data processing pipeline. By exploring different deep learning models, including ResNet, DenseNet, and Vision Transformer (ViT), and performing hyperparameter tuning, we obtained a test accuracy of 0.8518 on the three-class classification problem. We further implemented a novel data-centric AI method to append subclass labeling of images that can help our model to learn diverse classes. This method improved the test accuracy to 0.853, and has a better recall value for heavier snow coverage.

2. Related Work

Road condition detection with deep learning techniques has been studied from a variety of angles. These selected works focus on classifying weather-related road surface condition using CNN architectures, and using images collected from either in-vehicle or road-side cameras.

*This co-author is not enrolled in CS231N, they provided the dataset we used (no additional guidance).

Lu et. al(2010) [3] developed a neural network model that takes road temperature, radiation, air temperature, air humidity, season, location, and time as independent variables, and predicted road surface conditions (dry road or wet road) with 90% accuracy.

Pan et. al (2017) [7] proposed the idea of making use of a pre-trained convolutional neural network (CNN) with an addition of extra layers of neurons for model fine-tuning and localization. The authors developed a road surface condition image dataset with labels representing different levels of snow coverage. The resulting model achieved a three-class classification accuracy of 87.3%.

Khan et. al (2021) [4] used an ResNet-18 architecture for weather detection: clear, light snow, and heavy snow; as well as three surface conditions: dry, snowy, wet/slushy. With image data from roadside webcams by Department of Transportations (DOTs) in the U.S., they achieved an overall detection accuracy of 97% on weather and 99% surface condition detection respectively.

Our main takeaway from the literature review is that the more novel methods exhibit improved performance on the task, a trend we hope to continue. Further, the input to the CNN models in [7] used only the image, whereas the model in [3] used many additional numeric variables, demonstrating the power of computer vision and image data as an approach to the road condition classification task. That said, it is hard to directly compare the papers since different road classification datasets are used. It could be interesting in future work to explore the inter-operability of the models: do they work when the dataset is slightly different? It is worth noting that while we hoped to obtain a dataset with road condition pictures from the viewpoint of a car, we instead have pictures from the viewpoint of a weather observation station, a related task.

Further, there has been limited past work on the problem of subclass learning. Back in 2001, Hoffmann et. al [1] explored subclass learning using pre-neural modeling techniques and observed improvements in model performance with additional labels of the subclasses. In addition, the presentation of ImageNet discusses the hierarchical nature of the classes in the dataset (e.g. different types of dog), and how humans take advantage of this hierarchy to develop compact object classification models [8].

3. Dataset

The Road Weather Information System (RWIS) data used in this paper was collected from two highway sections in South-Western Ontario, Canada near Mount Forest. This area experiences an annual average of 59 days of snowfall of at least 0.2cm. The data set includes images of the highway surface taken from 60 separate weather stations, captured during the winter of 2014 at different time across a day. We obtained the data with permission from Prof. Pan from Linyi University, who collected and used this dataset in his and his co-authors work *Winter Road Surface Condition Recognition Using a Pretrained Deep Convolutional Network* (2018) [7]. Although pictures were taken from RWIS stations, models trained using this dataset can be applied to classify images from roadside CCTV cameras, which have lower installation costs.

For pre-processing, we cropped the date and time bar on the top of the image, and resized them to 256x256 pixel size before converting them to RGB channel tensors. The target variable, road surface snow coverage condition is categorized into three levels: barely covered, partly covered and fully covered. The three-class definition is in accordance with Transportation Association of Canada's route reporting terminology, which is used to convey road surface condition to the general public. The distribution of classes is presented in Table 1.

Snow Coverage	Image count	Percentage
Barely	9,506	44.1%
Partly	8,930	41.5%
Fully	3,104	14.4%
Total	21,540	100%

Table 1. RWIS dataset by snow coverage categories.

The data set is split into train, validation, and test set by the ratio of 70%-15%-15%. In previous works, Pan et. al separated all images at random to form their splits. However, we believe this technique would result in data leakage, since similar images from the same camera station have likely already been seen during training. Thus, we decided to split the data set by station, such that the validation and test sets contain a similar distribution of different classes, but from an unseen camera perspective. This will make the classification task more challenging, but more realistic and generalizable to newly installed camera stations that have no labeled data available.

For reference, we have provided some example images after pre-processing as an appendix at the bottom of this report.

4. Methods

In this section, we explain the different approaches we ran experiments on for the snow coverage classification task. Different approaches were compared using prediction accuracy on the test set. We thank the AI for Climate Change Lab, part of the Stanford Machine Learning Group, for providing starter code framework for us to build our models off of [2].

4.1. 3-Class Classification

As the first part of our project, we aim to achieve an improved performance compared to the previous published work that uses the same dataset [7], by applying more recent and higher-performing CNN architectures.

4.1.1 Models and Finetuning

The two questions we are interested in are

- Would pre-trained model yield better snowy road classification accuracy? What is a good number of fine-tuning parameters?
- Would more complicated and deeper models structure benefit our specific task?

We experimented with ResNet(18 and 34), DenseNet(121 and 161) and VisionTransformer(16 and 32). For each model architecture, we loaded the parameters pre-trained on ImageNet [8], and we experimented both partially fine-tuned and fully fine-tuned versions for performance comparison. The non pre-trained models were also included as comparisons. We experimented different sizes of each model to see if the model complexity improves prediction accuracy. And finally we performed random search for finding the most suitable learning rate specific to each architecture. The models and pre-trained parameters are obtained from the TorchVision library [6].

The ResNet architecture is comprised of convolutional, pooling and fully connected layers, with added residual connections. We believe ResNet to be a powerful architecture for our task as explored previously by Khan et. al (2021) [4]. The added identity mapping in the residual block ensures that the model takes the representation from earlier layers as an input in following steps, and enables training of deeper models.

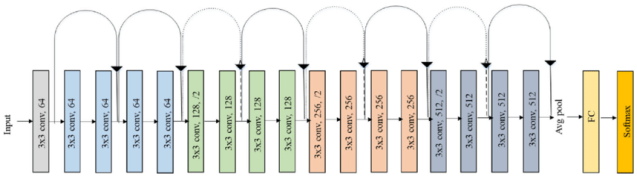


Figure 1. ResNet18 architecture

The DenseNet architecture utilizes dense blocks to connect all layers directly with each other. For each layer, the feature-maps of all preceding layers as well as its own feature-maps are used as inputs. Though DenseNet is relatively computationally inefficient, we believe DenseNet can be effective for snowy road classification by strengthening feature propagation and encouraging feature reuse with small dataset. We implemented pre-trained DenseNet 121 and DenseNet 161 with various levels of fine-tuning for comparison.

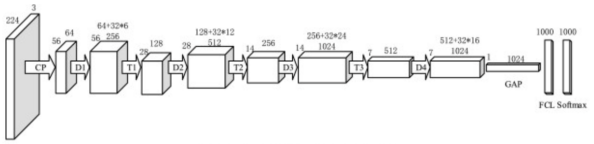


Figure 2. DenseNet121 architecture

Vision Transformer emerges as the state of art sequence processing architecture. It incorporates multihead attention blocks which allow for efficient training of long sequence and parallel computing. Given that the RWIS dataset is relatively small compared to other general image classfication problems, we are curious to see if the vision transformer applied directly to sequences of image patches can enhance learning. We implemented pre-trained ViT-16 and ViT-32 for comparison.

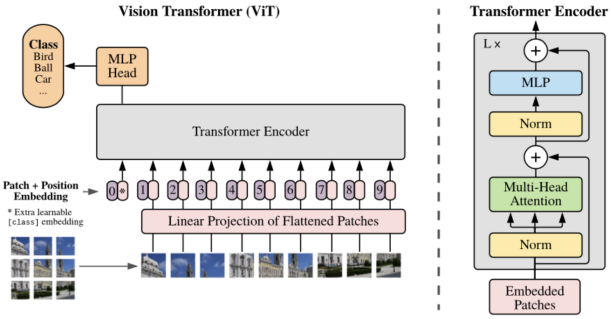


Figure 3. Vision Transformer architecture

4.2. Subclass Classification

The second part of our project was inspired by inspecting the training data more carefully and finding that daytime images look completely different from the nighttime images, even from the same station. We hypothesize that it could be challenging for our model to recognize snowy road sections in both daytime and nighttime conditions. All of the nighttime images are much more similar to each other than fully-snowy-nighttime is to fully-snowy-daytime. We would like to investigate if the model would perform better if we separated the data into 6 classes instead of 3 (with each snow coverage having daytime and nighttime components) and then map the six-class classification results to three classes for evaluation.

4.2.1 Day/Night Classifier

The subclass classifier experiment requires a daytime/nighttime label for each image. To obtain this label, we

first manually labeled a small data set and trained a simple convolutional neural network to produce labels for all images. We define a nighttime image as having no observable sunlight, regardless of the time that the picture was taken. The manually labeled data set consists of 121 daytime and 134 nighttime images from randomly chosen stations, with 20% of this data used for evaluation. The classifier consists of two CNN blocks and a linear layer to predict the probability of being nighttime. The network architecture is illustrated in Figure 4.

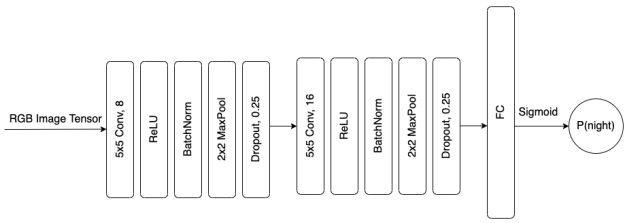


Figure 4. Day/Night classifier architecture, own work.

Using an Adam optimizer with a learning rate of 0.001, the day/night classifier achieved 0.98 accuracy on the train data and 0.96 accuracy on the test data. By concatenating the snow coverage label and the predicted day/night label, the subclass distribution is as follows:

Snow coverage - day/night subclass	Image count	Percentage
Barely-Day	5,772	26.8%
Barely-Night	3,734	17.3%
Partly-Day	4,970	23.1%
Partly-Night	3,960	18.4%
Fully-Day	1,056	4.9%
Fully-Night	2,048	9.5%

Table 2. RWIS dataset by subclass categories.

It is interesting to note that fully covered images are nearly twice as likely to be labeled as night than day, whereas barely covered images are more likely to be day than night, indicating that perhaps the three-class model may try to somehow use night as a signal for fully covered.

4.2.2 Subclass Classification Pipeline

In order to conduct the subclass experiment, we modified our classification pipeline to handle the day/night data with six classes. During training, the models were trained to perform a 6-class classification task, and the cross-entropy loss was computed using six classes instead of three. For evaluation, we aggregate the predicted subclass scores to map the six-class predictions to the original 3-class target, then compute accuracy metrics accordingly, so that we can obtain a fair comparison with the previous models.

There are two potential choices of how to aggregate the predicted scores to produce 3-class predictions. We evaluated every model using both approaches.

- The first recalculation approach was to take each of the subclass pairs, e.g. barely-day and barely-night, and sum their scores to obtain the score for the three-class task.
- The second recalculation approach was to instead take the maximum score of e.g. barely-day and barely-night for each of the three classes to get scores for barely snow covered.

The same hyperparameter combinations were tested for the subclass models as the original models. Before experimenting the subclass approach on the RWIS data, we first tested it on the CIFAR-10 dataset [5]. For the CIFAR-10 experiment, a ResNet-18 with learning rate of 0.003 was used for both models.

5. Experiments and Analysis

In this section we present the evaluation and visualization results of the models we experimented.

5.1. 3-Class Classification Results

Table 3 displays the test set accuracy of the models prior to splitting into subclasses. After hyperparameter tuning, the learning rates used for ResNet, DenseNet, ViT-L/16, and the rest of the vision transformers are 0.0001, 0.0001, 0.00001, 0.00005 respectively.

Test Set Results		
Model	**Fine-tune**	**Acc**
ResNet-18	1	0.7760
ResNet-18	3	0.8190
ResNet-18	all	0.8385
ResNet-18	not pre-trained	0.7370
ResNet-34	all	0.8335
DenseNet-121	1	0.7605
DenseNet-121	**all**	**0.8518**
DenseNet-121	not pre-trained	0.7556
DenseNet-161	all	0.8082
ViT-L/16	all	0.8270
ViT-L/32	all	0.8165
ViT-B/16	all	0.8499
ViT-B/32	all	0.8143

Table 3. Number of layers fine-tuned and test set accuracy for each 3-class classifier model.

Comparing different number of layers fine-tuned, it was found that fine-tuning more layers results in better accuracy, likely because the RWIS data is insufficient and different from the pre-training ImageNet data. Despite the discrepancy between the pre-training dataset and the RWIS data, the pre-trained models still achieved significantly better accuracies than the non pre-trained ones. The best performing model obtained has a DenseNet-121 structure, pre-trained and with all layers and parameters fine-tuned.

The test accuracy obtained is 0.8518, which we consider comparable with the previous work [7] (pre-trained CNN with test accuracy = 0.87), taking into account that we partitioned the test data by camera stations instead of randomly. In fact, we experimented the same model on a train and test set that are randomly split regardless of stations, and obtained a test accuracy higher than 0.94, demonstrating that the random split indeed overestimates model performance on unseen camera stations, and that the model architecture we tested outperform the ones from the previous work on this dataset.

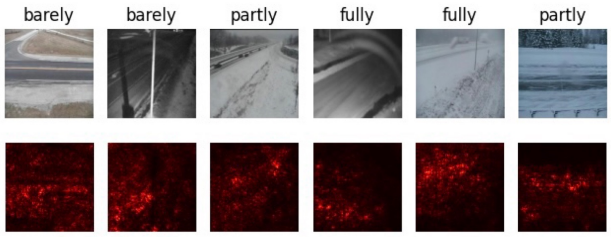


Figure 5. Saliency map samples of the best-performing DenseNet-121 model.

We visualized some outputs from the best performing model. Figure 5 illustrates two saliency map samples per class. The saliency map was constructed by performing a gradient ascend of the correct class score (logits), with respect to each pixel element, and taking the maximum gradient absolute value for each pixel across three channels. The red highlight in the saliency maps indicates that the pixels in this region contribute more to the correct classification result. It was found that the model was able to locate the road surface in the image and ignore irrelevant information such as street lights. For barely and partly covered roads, the saliency maps show a visible line corresponding to the road surface. For fully covered roads, the saliency map tends to form a cluster, and this behavior aligns with our expectation. Figure 6 displays some misclassified samples. We found that model sometimes underestimates snow coverage when there is a streak of dark line and overestimates snow coverage when there is a streak of while line. Table 4 shows the confusion matrix of the test set prediction. Barely covered road has the best prediction accuracy, and misclassification mostly happens between two adjacent levels.

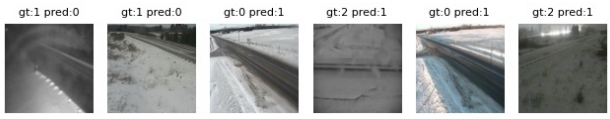


Figure 6. Misclassification samples (ground truth vs. predicted). 0: barely, 1: partly, 2: fully.

	barely	partly	fully
barely	0.9009	0.0941	0.0050
partly	0.1059	0.8315	0.0626
fully	0.0348	0.1905	0.7747

Table 4. Confusion matrix of the DenseNet-121 model.

During training, given that we fine-tuned enough layers and selected a proper learning rate, we always observed high train accuracy (¿ 0.98). This indicates that the neural network is able to describe most of the pattern in the train data we provided, which explains why increasing model complexity (e.g. from ResNet-18 to ResNet-34) doesn't guarantee a better performance. Although we observed small increase in test accuracy by exploring different models, we concluded that it is difficult to achieve a substantial improvement by further changing model architecture, due to the insufficient data, the way we split the train/test data, and the ambiguity and subjectivity in labeling snow coverage. We believe that the key to further improve model accuracy is to focus on data instead of the model choice. The subclass experiment is an example of such data-centric approach.

5.2. Subclass Classification Results

5.2.1 CIFAR-10 Experiment

We first tested out our subclass hypothesis on the CIFAR-10 dataset. To do this, we proposed an artificial classification task between animal and non-animal (Ship-Truck-Car-Airplane). Would the model perform better on this task if we used the ten subclasses or only the aggregated two classes?

Test Set Results	
2-classes	0.8742
10-classes Sum	0.8736
10-classes Max	0.8682

Table 5. Test set accuracy for the 10-class and 2 class metrics on CIFAR-10.

We observe that the additional information for each of the 10 classes does not actually yield an improvement in model performance over using the two classes alone. This caught us by surprise – we had expected the highly diverse

animal class to be challenging for the model to learn as one joint class separate from non-animals. However, it seems the model was able to learn this. Perhaps, despite our intuition this task was actually too easy for the model to require additional guidance on, since it was already achieving somewhat high accuracy.

As a further exploration, we ran the same experiment but with a reduced training set (5000 images instead of 45000), and observed similarly minimal differences in model accuracy.

A key takeaway from this experiment is that providing the model with more information will not always help, and that advanced conventional models already have high capacity to learn the latent classes within the data.

5.2.2 Snowy Road Classification Results

Below we present the evaluation results of the subclass approach on classifying snowy roads using different models. We also present the saliency maps and misclassified samples from the best subclass model.

Subclass Test Results	
Model	**Acc**
ResNet-18	0.8382
ResNet-34	0.8363
DenseNet-121	0.8502
DenseNet-161	**0.853**
ViT-B/16	0.7438

Table 6. Model accuracy results for subclass version with 6 classes. Note that all models here had a learning rate of 0.0001, with all layers fine-tuned.

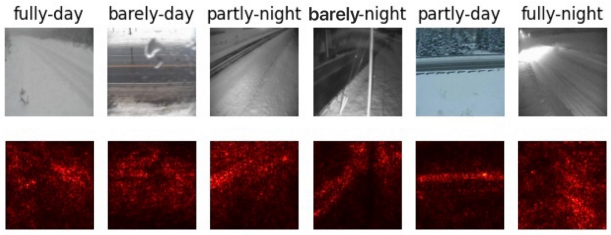


Figure 7. Saliency map samples of the best-performing DenseNet-161 subclass model.

Comparing the two aggregation approaches of mapping predictions from six to three classes, for most models, summing the scores usually outperform taking the maximum. The saliency maps demonstrate similar behavior with the non subclass models, for example focusing on the road surface and ignoring irrelevant information. Out of all of the models that we experimented, the highest accuracy on the test set was the DenseNet-161 with subclasses, with an accuracy of 0.853. However, it is in line with the best performing non sub-class model, and likely higher just by random chance. The impact of the subclass model gets more interesting when we compare the confusion matrices of the best performing non-subclass to the best performing subclass model. We observe higher accuracies for the partly and fully covered classes using the sub-class approach, and more balanced accuracies across classes. For our task, it would be most important to catch the cases where the road is partly or fully covered in snow, and the subclass model has notably higher recall for the partly and fully covered cases, as shown in table 7.

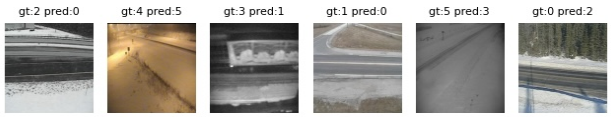


Figure 8. Misclassification samples (ground truth vs. predicted). 0: barely-day, 1: barely-night, 2: partly-day, 3: partly-night, 4: fully-day, 5: fully-night.

Recall by Model and Road Condition			
	Barely	Partly	Fully
3-Class	**0.9009**	0.8315	0.7747
Subclass	0.8728	**0.8508**	**0.8077**

Table 7. Model recall by road condition, shown for the best 3-class model (DenseNet-121) and best subclass model (DenseNet-161, with sum).

We did not anticipate our model to perform in this way, and are still developing an understanding of how it was able to achieve higher accuracy on the fully and partially covered classes. It is possible that the night and day distinction was actually helpful in learning these classes, or the model could have for some reason been pushed towards favoring more snowy road labels when it was uncertain.

5.3. Analysis Expansion

Overall, the data-centric approach of sub-class labeling was not immediately helpful in improving model performance on CIFAR nor on our task. That said, there are a few directions we would be curious to explore to better understand our subclass experiment.

First, it would be interesting to see how well the three-class model could do on the six-class tasks with a minimal amount of fine-tuning at different stages of training. If we were to observe that as the three-class performance improves, it also more easily fine-tunes to the six-class task, then this would indicate that the three-class model is actually learning the latent day/night classes as part of the

model.

Second, it could be interesting to see if subclasses are more helpful on smaller models. This hypothesis is influenced by the earlier work that suggested having subclasses was helpful in improving model performance.

Third, it would be interesting to see if instead of using manual labeling and a CNN model for night/day, if we could have instead used an autoencoder of images to create image embeddings, and grouped them to get the latent night/day (or another relevant latent property) without the need to manually label.

6. Conclusion

We compared different state-of-the-art classification models and evaluated their performance on recognizing road surface snow coverage level. Our best performing model achieves good test accuracy (0.85) when compared to previous work on this dataset (0.87 [7]), even though we split the data in a way that made for a more challenging task. When splitting data at random as Pan did we achieved accuracies of 0.94, though we do not highlight these results as the test set is certainly polluted. From model comparisons, it was found that pre-training and fine-tuning all layers in general yields better results. Using deeper models with more parameters does not necessarily improve results significantly because the less complicated models might have already reached the learning capacity, indicated by the high train accuracy.

We acknowledge the challenge in improving test accuracy by solely modifying model structure, and hence proposed a data-centric approach to further split the data into daytime/nighttime subclasses. This approach was not significantly helpful in improving the test accuracy, potentially because models have already learned the latent classes without being given explicit labels. However, the subcalss approach results in better recall rate for the partly and fully snow coverage classes. It piqued our interests and ideas for future work to see if the subclass method or other data-centric approach, such as different data augmentation techniques, could boost model performance, especially in the case of having smaller models or smaller datasets.

Overall, the task of recognizing winter road condition is an interesting and important application of computer vision techniques. Future work on this task can focus on augmenting the data with images from different camera angles, and perform input fusion with other numeric features such as weather information.

References

[1] Hoffman et. al. Using subclasses to improve classification learning. 2001. 2

[2] Irvin et. al. Artificial intelligence for climate change starter code. 2022. 2

[3] Lu Junhui and Wang Jianqiang. Road surface condition detection based on road surface temperature and solar radiation. In *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering*, volume 4, pages 4–7, 2010. 2

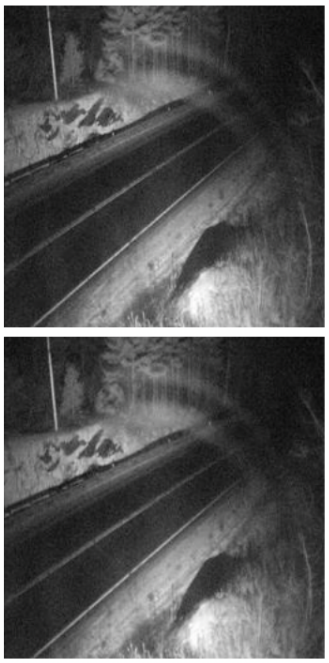
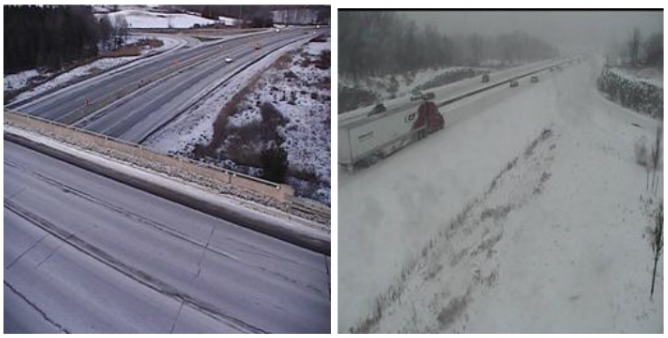
[4] Md Nasim Khan and Mohamed M. Ahmed. Weather and surface condition detection based on road-side webcams: Application of pre-trained convolutional neural network. *International Journal of Transportation Science and Technology*, 2021. 2, 3

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 4

[6] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. 3

[7] Guangyuan Pan, Liping Fu, Ruifan Yu, and Matthew Muresan. Winter road surface condition recognition using a pre-trained deep convolutional network, 12 2018. 2, 3, 5, 7

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014. 2, 3

Road Snow Coverage	Daytime	Nighttime
Barely		
Partly		
Fully		

Dataset Example Appendix: We present here select distinct images from the dataset based on their labeled snow coverage, with manual separation for daytime and nighttime. Note some images look nearly identical, and others within the same class look completely different, making for a challenging classification task. Note that the two barely nighttime images are in fact different images in the dataset, reflecting how a random split of train-validation-test may give misleadingly high test accuracies.