AdaLN: A Vision Transformer for Multidomain Learning and Predisaster Building Information Extraction from Images

Yunhui Guo[1]; Chaofeng Wang, A.M.ASCE[2]; Stella X. Yu[3]; Frank McKenna[4]; and Kincho H. Law, F.ASCE[5]

Abstract: Satellite and street view images are widely used in various disciplines as a source of information for understanding the built environment. In natural hazard engineering, high-quality building inventory data sets are crucial for the simulation of hazard impacts and for supporting decision-making. Screening the building stocks to gather the information for simulation and to detect potential structural defects that are vulnerable to natural hazards is a time-consuming and labor-intensive task. This paper presents an automated method for extracting building information through the use of satellite and street view images. The method is built upon a novel transformer-based deep neural network we developed. Specifically, a multidomain learning approach is employed to develop a single compact model for multiple image-based deep learning information extraction tasks using multiple data sources (e.g., satellite and street view images). Our multidomain Vision Transformer is designed as a unified architecture that can be effectively deployed for multiple classification tasks. The effectiveness of the proposed approach is demonstrated in a case study in which we use pretrained models to collect regional-scale building information that is related to natural hazard risks. **DOI: 10.1061/(ASCE)CP.1943-5487.0001034.** *© 2022 American Society of Civil Engineers.*

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

Introduction

The aftermath of a natural hazard, such as an earthquake or a tropical cyclone, may significantly impact a city, causing damages to buildings and infrastructures, casualties, and economic losses. According to the National Oceanic and Atmospheric Administration (NOAA), weather-related disasters, mainly due to tropical cyclones (also known as hurricanes), have caused the most deaths and destruction among all types of natural disasters, resulting in more than $22 billion loss in 2020 in the United Sates alone (NCEI 2021). Furthermore, over the past couple decades, earthquakes and tsunamis have led to nearly 750,000 lives lost (Wallemacq and House 2018).

Buildings are the most vulnerable asset of the built environment affected by a natural hazard. Postdisaster assessments are usually performed to evaluate the building damage level and reduce the risk to responders (Torok et al. 2014; Zhu et al. 2011; Wang et al. 2020b; Zhou et al. 2016b). More importantly, for natural disaster preparedness programs, predisaster screening of the vulnerabilities of buildings is often the first task (Wang et al. 2021a; ATC 1988). Predisaster screening relies on knowledge of existing building stock. For example, a three-dimensional (3D) map of downtown Victoria, British Columbia, Canada, was created to understand the predisaster building representations (Kucharczyk and Hugenholtz 2019). Such screening processes are usually based on visual cues that indicate potential deficiencies in the structures (ATC 1988). However, the manual and labor-intensive screening process is costly and time consuming for government agencies and, in addition, poses a significant economic burden on the residents (Wang et al. 2021a). An automated, cost-effective approach that can facilitate building vulnerability investigation would be desirable.

The availability of images collected via remote sensing has been a popular source for extraction of building information (Zhang 1999; Hamaguchi and Hikosaka 2018; Ivanovsky et al. 2019; Yu et al. 2020; Wang et al. 2021a) and for natural hazard-related analyses (Joyce et al. 2009; Patino and Duque 2013; Poursanidis and Chrysoulakis 2017; de Beurs et al. 2019). Meanwhile, because of their availability and the rich visual information, another type of imagery data, street view images, have attracted significant attention from researchers and have emerged as a much-sought-after resource. Potential uses of street view images have been demonstrated for a wide variety of applications, ranging from predicting housing prices (Bency et al. 2017; Law et al. 2018) to evaluating the safety of neighborhoods (Naik et al. 2014; Liu et al. 2017). Street view images have also been shown useful for rapid screening of certain structural vulnerabilities, such as soft story of buildings (Yu et al. 2020).

The past several years have witnessed significant progress in deep learning (DL) technologies, which have been embraced by the computer vision community. DL techniques, e.g., deep convolutional neural networks (CNNs), have been applied in various fields of civil and natural hazard engineering and achieved impressive performance (Gao and Mosalam 2018; Wang et al. 2018, 2020a; Cha et al. 2018; Guo et al. 2020; Czerniawski and Leite 2020). However, directly applying CNNs to satellite or street view images faces a number of issues. First, each image typically has multiple

[1]Postdoctoral Scholar, International Computer Science Institute, Univ. of California, Berkeley, Berkeley, CA 94704.

[2]Assistant Professor, M.E. Rinker, Sr. School of Construction Management, College of Design, Construction and Planning, Univ. of Florida, Gainesvill, FL 32603 (corresponding author). ORCID: https://orcid.org/0000-0001-8534-9276. Email: chaofeng.wang@ufl.edu

[3]Director, International Computer Science Institute, Univ. of California, Berkeley, Berkeley, CA 94704. ORCID: https://orcid.org/0000-0002-3507-5761

[4]Research Engineer, Dept. of Civil and Environmental Engineering, Univ. of California, Berkeley, Berkeley, CA 94720.

[5]Professor, Dept. of Civil and Environmental Engineering, Stanford Univ., Stanford, CA 94305.

Note. This manuscript was submitted on November 11, 2021; approved on March 26, 2022; published online on July 4, 2022. Discussion period open until December 4, 2022; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, © ASCE, ISSN 0887-3801.

objects that introduce noises into the training process. Second, the object of interest, such as the roof of a building, may only occupy a small portion of the whole image. CNNs designed for image classification have no built-in capability to accurately locate a specific object of interest in the image (Wang et al. 2017). Third, while training CNNs from scratch needs a large number of labeled images, the cost of labeling and annotating satellite or street view images is prohibitive and limits the number of labeled data (Wang et al. 2020c).

Attention networks (Zheng et al. 2017; Wang et al. 2017; Woo et al. 2018) have been proposed to improve conventional CNNs with additional attention modules to refine the feature map of the CNNs by focusing on the part of the image that is critical for the specific task of interest. Recently, the Vision Transformer (Dosovitskiy et al. 2020) was proposed as an image classification architecture that relies only on attention modules and has no convolutional layers. With the attention modules, the Vision Transformer is particularly suitable to address the limitations of CNNs for street view or satellite images, i.e., to extract specific objects of interest in the images. However, the model size of the Vision Transformer is significantly larger than the conventional CNNs, requiring additional cost for training and storing the model.

In this work, we leverage a pretrained Vision Transformer as the backbone network for extracting relevant information from satellite and street view images for natural hazard analysis. As a demonstration, we tackle two tasks related to building information extraction: classification of roof types based on satellite images and identification of soft story from street view images. Instead of training one model for each task, we exploit the idea of multidomain learning (Rebuffi et al. 2017, 2018) to train a *single* Vision Transformer for both tasks with images from two disparate (satellite and street view) sources as shown in Fig. 1. Specifically, we employ the pretrained Vision Transformer on the ImageNet data set (Deng et al. 2009) as the backbone network. For each task, we adjust the layer-normalization layers (Ba et al. 2016) in the pretrained network. Our method is thus termed *adaptive layer normalization* (AdaLN). Compared to the baseline method, the proposed approach can maintain the same model accuracy but only needs 50% of the parameters. Because of the scalability of the proposed architecture, a number of tasks that involve satellite and street view images can be effectively represented and executed with a single compact model.

The motivation of using DL methods for examining the natural hazard risks of buildings from images is to save time and reduce human errors in large-scale or high-frequency tasks. However, traditional CNN-based DL methods encounter difficulties dealing with multiple aforementioned issues. To this end, we propose a transformer-based architecture that is different from traditional CNN-based methods. The contributions of this work are threefold: (1) this study, to the best of our knowledge, represents the first effort toward developing a Vision Transformer–based framework for automatic soft-story structure classification and roof-type classification; (2) we propose a multidomain Vision Transformer architecture that can deal with multiple tasks involving satellite or street view images with a compact architecture; and (3) the effectiveness of the AdaLN method is demonstrated on a real-world-use case study for natural hazard analysis.

Related Work

Vision-Based Predisaster Risks Examination

Vision-based examination has been widely adopted in building risk analysis. For example, roof type is crucial information for evaluating wind effects on structures. Fragility analyses show that different roof types could result in significant variations in the probability of damage state exceedance (FEMA 2018). In a recent study, vision-based examination of roofs is successfully implemented for hurricane risk analysis of buildings (Wang et al. 2021a). In another instance that dates back to three decades ago, the Applied Technology Council (ATC) published *FEMA 154* (ATC 1988), a guidebook for assessing the seismic performance of structures by employing a scoring scheme based on visual examinations. The scoring scheme provides the earthquake engineering community with a cost-effective way to analyze the seismic resistance of a vast stock of buildings, without accessing the buildings but instead mainly based on the visual cues manifested at building exteriors. The scoring system has been used broadly for evaluating structures in countries and regions prone to destructive earthquakes (Karbassi and Nollet 2007; Wallace and Miller 2008; Srikanth et al. 2010; Saatcioglu et al. 2013; Perrone et al. 2015; Ploeger et al. 2016; Ningthoujam and Nanda 2018).

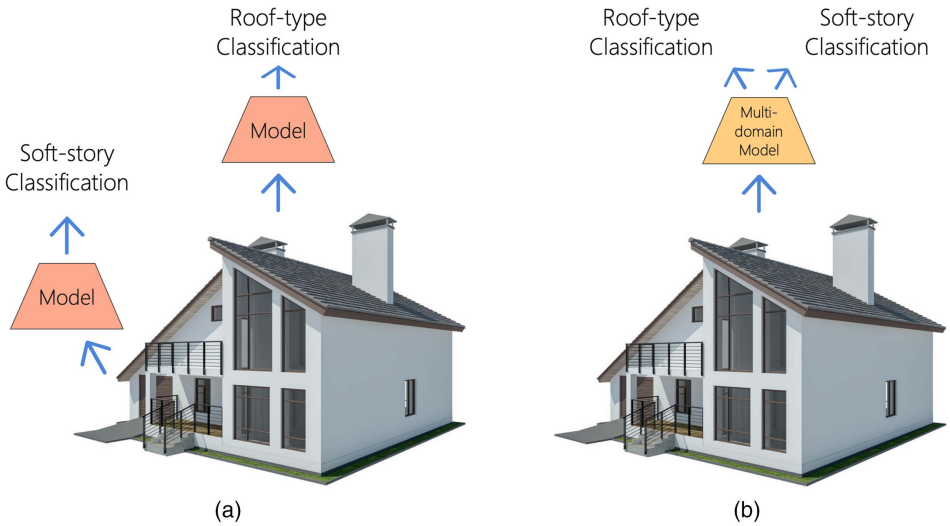


Fig. 1. Different approaches for information extraction from building images: (a) the baseline approach needs to train separate models for extracting different information from the building, which incurs large costs for training and storing the models; and (b) our proposed approach trains a single multidomain model for all the tasks with comparable accuracy to the baseline approach.

J. Comput. Civ. Eng., 2022, 36(5): 04022024

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

The intuition behind vision-based risk examination methods is that, to a considerable degree, the performance of a built structure depends on its construction type and material, geometric irregularities, maintenance conditions, etc., and such information can be visually inferred from the exterior of the structure. Although visual examination of building risks has been broadly adopted, such examination can be expensive and error-prone (due to reasons such as fatigue or neglect) as it can be labor-intensive when examining a vast number of buildings. Decisions can also be subjective, potentially leading to diverging interpretations and erroneous results. To this end, an alternative procedure for automatically evaluating the risks of buildings based on visual cues in sensed imagery data is desired.

Satellite and street level imagery can capture rich environmental objects such as buildings, vegetation, and vehicles. The use of such images has drawn considerably increasing attention from researchers in various disciplines. Satellite imagery has been an important source for building information acquisition, including the footprint extraction, roof information inference, and damage state assessment (Jin and Davis 2005; Lari and Ebadi 2007; Brunner et al. 2010; Hang and Cai 2020). Compared with satellite images, street view images can provide detailed building information that is observable from the facades. Researchers have proved the wide applicability of street view images in a variety of different studies (Naik et al. 2014; Gebru et al. 2017; Law et al. 2018; Kang et al. 2018). In the natural hazard engineering domain, building risk analyses can benefit from satellite and street level imagery. In this study, we present a transformer-based approach to extracting building information from such imagery.

Transformer

The transformer (Vaswani et al. 2017) is a sequence transduction model that is first proposed for natural language processing (NLP). Different from conventional or recurrent neural networks (Lipton et al. 2015), the transformer only relies on attention mechanisms to extract dependency relations between input and output. The transformer has achieved state-of-the-art results on several NLP tasks (Vaswani et al. 2017). Besides NLP, transformer-based models have been utilized in recommendation systems (Sun et al. 2019). CNN has been the dominant architecture in computer vision since 2012. However, CNNs are known to have limitations in long-range relation modeling, caused by the intrinsic convolution operation locality (Linsley et al. 2018). Hence, CNNs generally perform less exceptionally for visual structures that show large variations in terms of size, texture, and shape (Geirhos et al. 2018), which are common in various imagery domains, including satellite and street view images. A self-attention mechanism is found to be able to overcome this limitation. Because the transformer is an architecture that entirely and solely relies on the attention mechanism, it has emerged as a powerful alternative architecture for CNN in computer vision. There have been abundant studies on improving the Vision Transformer from different aspects such as utilizing distillation (Touvron et al. 2020), pyramid architecture (Wang et al. 2021b), and combining with convolutions (Wu et al. 2021). The recently proposed Vision Transformer (Dosovitskiy et al. 2020) has questioned the need for convolutional layers for image classification. Due to the model size of the transformer architecture, there is also a line of work trying to compress or quantize the model (Bhandare et al. 2019).

Multidomain Learning

The model size of CNNs incurs heavy costs for training and storing the model. To this end, multidomain learning is proposed as a way to create a single model to classify images from multiple visual domains in order to save costs (Rebuffi et al. 2017, 2018). Bilen and Vedaldi (2017) shows that a single neural network can be trained for multiple image domains by only fine-tuning the instance normalization layer (Ulyanov et al. 2016). Rebuffi et al. (2017, 2018) proposed universal parametric families containing a small number of task-specific parameters for multidomain learning. Rosenfeld and Tsotsos (2017) proposed deep adaptation networks (DAN) that have the constraint that the filter for the new domain is a linear combination of the existing ones trained on old domains. In Guo et al. (2019), the authors proposed a multidomain learning architecture based on depthwise separable convolution (Chollet 2017). The proposed architecture can reduce the number of parameters by 50% compared with the state-of-the-art approach. Multidomain learning can promote the application of DL-based vision models because it reduces engineers' effort to train new models for new visual domains.

Preliminary

Multidomain Learning

Multidomain learning aims at creating a compact model for multiple tasks. Consider a set of N tasks $\{T_1, T_2, ..., T_N\}$, each task T_i consists of a triplet $\{X_i, Y_i, P_i\}$. $X_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ is the image space, where C_i is the number of image channels, H_i is the height of the image, and W_i is the width of the image. $Y_i \in \{1, 2, ..., L_i\}$ is the label space, and L_i is the maximum label index. $P_i(x, y)$ is the joint probability distribution of image $x \in X_i$ and label $y \in Y_i$.

Given a deep neural network $f_i(x)$: $\mathbb{R}^{C_i \times H_i \times W_i} \rightarrow \{1, 2, ..., L_i\}$ and the cross-entropy loss function c, the expected risk of $f_i(x)$ can be computed as follows:

$$R_i = \mathbb{E}[c(y, f_i(x)] = \int c(y, f_i(x)) dP_i(x, y) \tag{1}$$

The standard way for tackling all the T tasks is to train one model for each task T_i. However, with the standard approach the total model size scales linearly with the number of tasks. The goal of multidomain learning is to design neural network architecture such that the total model size scales sublinearly with respect to the number of tasks. The general strategy of multidomain learning is to design task-agnostic parameters that can be shared across all the tasks and add task-specific parameters for each particular task. The goals of multidomain learning can be summarized as: (1) maintain the model accuracy across all the tasks; (2) maximize the number of task-agnostic parameters; and (3) minimize the number of task-specific parameters.

Transformer

The transformer (Vaswani et al. 2017) is a special kind of attention neural network that does not rely on convolutional layers. The transformer is first used in NPL, in which each word in the sentence is converted into a word embedding for encoding the meaning of the word (Mikolov et al. 2013). The input of the transformer is an embedding matrix that represents the word embbedings of the whole sentence. A transformer typically consists of an encoder and a decoder. Both the encoder and the decoder are constructed by stacking several self-attention layers.

The key operation in the self-attention layer is the scaled dot-product attention, which is shown in Fig. 2. For each word in the input sequence, an embedding vector of dimension d_e is generated. The embedding vectors of a sequence of length N can be stacked as an embedding matrix of dimension $d_e \times N$. The embedding matrix is then multiplied with different weight matrices to obtain the value

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

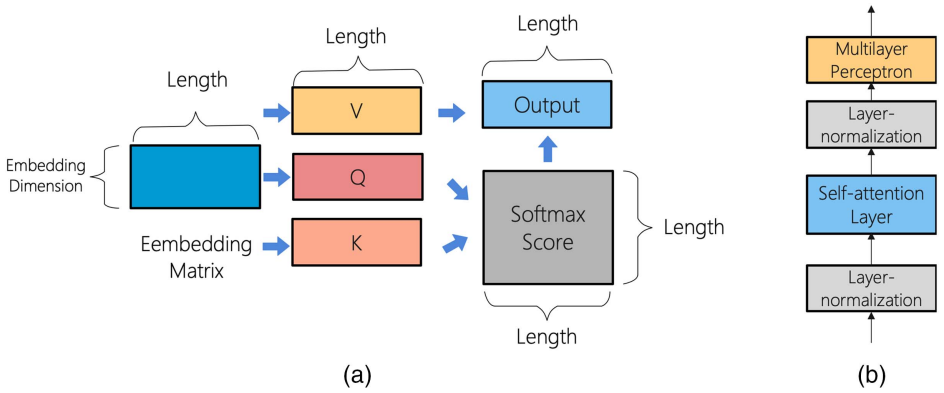


Fig. 2. (a) Operations in a self-attention layer. The embedding matrix is linearly transformed to obtain value (**V**), query (**Q**), and key (**K**) matrices. The query and key matrices are multiplied and routed through a softmax function to obtain the softmax score. (b) The architecture of the attention module.

(**V**) matrix, query (**Q**) matrix, and key (**K**) matrix. The output of the self-attention layer is computed as

$$\text{self-attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{2}$$

where $\text{softmax}(\cdot)$ is the softmax function, which is defined as

$$\text{softmax}(Z)_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} \quad \text{for } i = 1, ..., K \quad \text{and} \quad Z = (Z_1, ..., Z_K) \tag{3}$$

The $\text{softmax}((QK^T)/(\sqrt{D_k}))$ is defined as the softmax score that measures the dependencies of the words in the sequence. Practically, the embedding matrix is linearly projected h times with different linear projections to obtain the query matrix, key matrix, and value matrix, which is called multihead attention (Vaswani et al. 2017). Each of these projected versions of queries, keys, and values runs in parallel and the results are concatenated to obtain the final result.

Proposed Method

Our goal is to extract various aspects of 3D buildings with a single compact Vision Transformer as shown in Fig. 1. We introduce a multidomain Vision Transformer called AdaLN for satellite and street view image information extraction in natural hazard risk analysis. First, we introduce the network architecture of the Vision Transformer, which is used as the backbone network. Second, we introduce the details of the proposed approach for learning with multiple domains with a single Vision Transformer. Last, we analyze the model size of the proposed approach compared with several baselines.

Network Architecture

The Vision Transformer (Dosovitskiy et al. 2020) is a recently proposed architecture for image classification based on the language transformer. For the experiments, we use the base Vision Transformer architecture with 16×16 input patch size (Dosovitskiy et al. 2020). The input image $\mathbf{x}$ is reshaped to be a list of two-dimensional patches of size 16×16. A fully connected layer is applied on the image patches to obtain patch embeddings $E_{\mathbf{x}}$. There are a total of 12 attention modules. Each attention module A_i consists of two layer-normalization layers, one self-attention layer, and a multilayer perceptron that applies a nonlinear transformation on the output of the layer-normalization layer. The architecture of the attention module is shown in Fig. 2. The network ends with a softmax layer that normalizes the output of the attention modules using the $\text{softmax}(\cdot)$ function in Eq. (3) for classification. The loss function is the following cross-entropy loss

$$\ell(\mathbf{y}, \tilde{\mathbf{y}}) = -\mathbf{y} \log(\tilde{\mathbf{y}}) \tag{4}$$

where $\mathbf{y}$ = one-hot representation of the ground truth label; and $\tilde{\mathbf{y}}$ = output of the softmax layer. It is worth noting that there are no convolutional layers in the Vision Transformer. In essence, the attention can be interpreted as a vector that indicates the correlation of one pixel with other pixels. To make a prediction of one pixel, the attention module takes the sum of the values of other pixels weighted by the attention as the approximation. With the self-attention layers, the network focuses on the part of the image that is most critical for the task.

Learning Multiple Tasks

Although the Vision Transformer can obtain higher accuracy than the conventional CNN, it has many more parameters that need to be trained. In particular, compared with ResNet-50 (He et al. 2016), the base Vision Transformer contains four times more parameters. For the considered two tasks, training one separate model for each task would incur heavy costs for storing the models. This might not be a big issue for a small research project. But it is particularly undesirable for large research projects and industry projects, where there could be multiple tasks that need to be handled in a timely manner, which means that the number of parameters grows linearly with respect to the number of tasks. Instead, we propose a multidomain Vision Transformer that trains one model for all the tasks in order to save the cost for training and storing the models.

The proposed approach is based on the intuition that only the layer-normalization layers trained on the source task, i.e., ImageNet, need to be adjusted for each target task. Given a minibatch of vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$, each $\mathbf{x}_i$ is a vector of size of K that is the output of an intermediate layer in the network. For simplicity, we only consider fully connected layer here. Define $I = \{1, 2, ..., M\}$, for each $i \in I$, we can compute the mean μ_i and variance σ^2 of $\mathbf{x}_i$ as follows:

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_{i,k}$$

$$\sigma_i^2 = \frac{1}{K}\sum_{k=1}^{K} (\mathbf{x}_{i,k} - \mu_i)^2 \tag{5}$$

The sample $\mathbf{x}_i$ is then normalized as follows:

$$\tilde{\mathbf{x}}_{i,k} = \frac{\mathbf{x}_{i,k} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \tag{6}$$

where ϵ = small number used for numerical stability. The normalized feature is then transformed as follows:

$$\mathbf{o}_i = \gamma \tilde{\mathbf{x}}_i + \beta \tag{7}$$

where γ and β = learnable layer-normalization parameters. The layer-normalization layer can be used to stabilize the hidden state dynamics during training neural networks (Ba et al. 2016). Intuitively, the layer-normalization parameters capture the statistics of the source task, i.e., the ImageNet data set. When the model is repurposed for another task, it is necessary to adjust the layer-normalization parameters. In the paper, we find that it is possible to only adjust the layer-normalization parameters while keeping the rest of the layers fixed for the target task. Essentially, by adjusting the layer-normalization parameters, the self-attention layers trained on the source task can be reused on the target task regardless of the domain shift between natural images and satellite images. While the base Vision Transformer has around 86 million parameters, there are only around 40,000 layer-normalization parameters.

Model Efficiency Analysis

In this section, we compare the model size of the proposed approach with several baselines to demonstrate that the proposed approach incurs the least amount of parameters. In particular, we consider three baselines. The first baseline is called *Independent Network*, which trains one model for each task. Independent Network incurs the largest model size. The second baseline is called *Single Mask*, which applies a single trainable mask on the multihead attention while keeping the backbone network fixed. The trainable mask is a matrix of floating-point numbers that is multiplied elementwise with the softmax score matrix in Fig. 2. Intuitively, the trainable mask can adjust the attention maps for each new task. The third baseline is called *Multiple Masks*, which applies different masks to different heads of the multihead attention to consider the distinct roles played by each head.

In Table 1, we give the total number of parameters incurred by different approaches. Clearly, the proposed approach AdaLN leads to the least amount of parameters, which is critical for saving the cost of model training and storage. Compared with the standard approach, i.e., Independent Network, AdaLN reduces the number of parameters by 99.95% while maintaining accuracy, as we demonstrate in the section "Experiments."

Experiments

Data Sets

Satellite Images of Building Roofs

There are three major types of roof shapes that are widely used in many regions worldwide: flat, gable, and hip, as shown in Fig. 3. The shape of the roof influences the performance of buildings under wind pressure during hurricanes and tornadoes. It is not difficult for a trained engineer to differentiate these types by looking at the satellite image. We propose to employ DL to scale the classification task. For training, we collected 6,000 satellite images of building roofs, 2,000 for each type. For testing, we use a total of 124 images. All images were downloaded using Google Map API at the same zoom level and were cropped to the size of 256 × 256 pixels, so that the target building was located in the center of the image. All buildings (training and testing) were randomly selected in the United States.

Street View Images of Building Facades

For multistory buildings, irregular vertical geometries could cause one story's stiffness to be much less than other stories, which makes the building vulnerable to strong ground motion during earthquakes. Such buildings are termed as soft-story buildings. Fig. 4 shows an example of a soft-story building and its possible failure mode. Given the probability of severe damage or collapse in the event of an earthquake, identification of such structures is critical. This can be done by classifying the street view images of buildings. We collected 556 street view images of soft-story buildings and 736 non-soft-story buildings from Google Street View for training. For testing, we use a total of 395 images.

Baseline

We first compare the performance of the Vision Transformer with ResNet50 on the two tasks. For multidomain learning, we consider the following baselines in the experiments:

1. Independent Network: The simplest baseline we consider is Independent Network. We fine-tune the pretrained Vision Transformer for each task, which leads to two separate models. This method results in the largest model size because there is no sharing of parameters between different tasks.
2. Single Mask: As mentioned in the section "Proposed Method," in Single Mask the attention maps of the pretrained Vision Transformer are adjusted with one trainable matrix.
3. Multiple Masks: As mentioned in the section "Proposed Method," in Multiple Masks the multihead attention maps of the pretrained Vision Transformer are adjusted with different trainable matrices.

Experimental Setups

All networks are implemented using Pytorch (Paszke et al. 2019) and trained on four NVIDIA GTX 1080 GPUs. We use the pretrained Vision Transformer on ImageNet provided by the original

Table 1. Comparison of model size of different multidomain learning approaches. The proposed AdaLN has the least amount of total parameters. For the first task, each approach leverages a base Vision Transformer that has 86 million parameters. For the second task, different approaches incur a drastically different number of additional parameters

Approach	Description	Total number of parameters for two tasks
Independent Network	Train one network for each task	86 million+86 million
Single Mask	Adjust attention with a single mask	86 million + 470,000
Multiple Masks	Adjust attention with multiple masks	86 million+5 million
AdaLN	Fine-tune layer-normalization parameters	86 million + 40,000

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

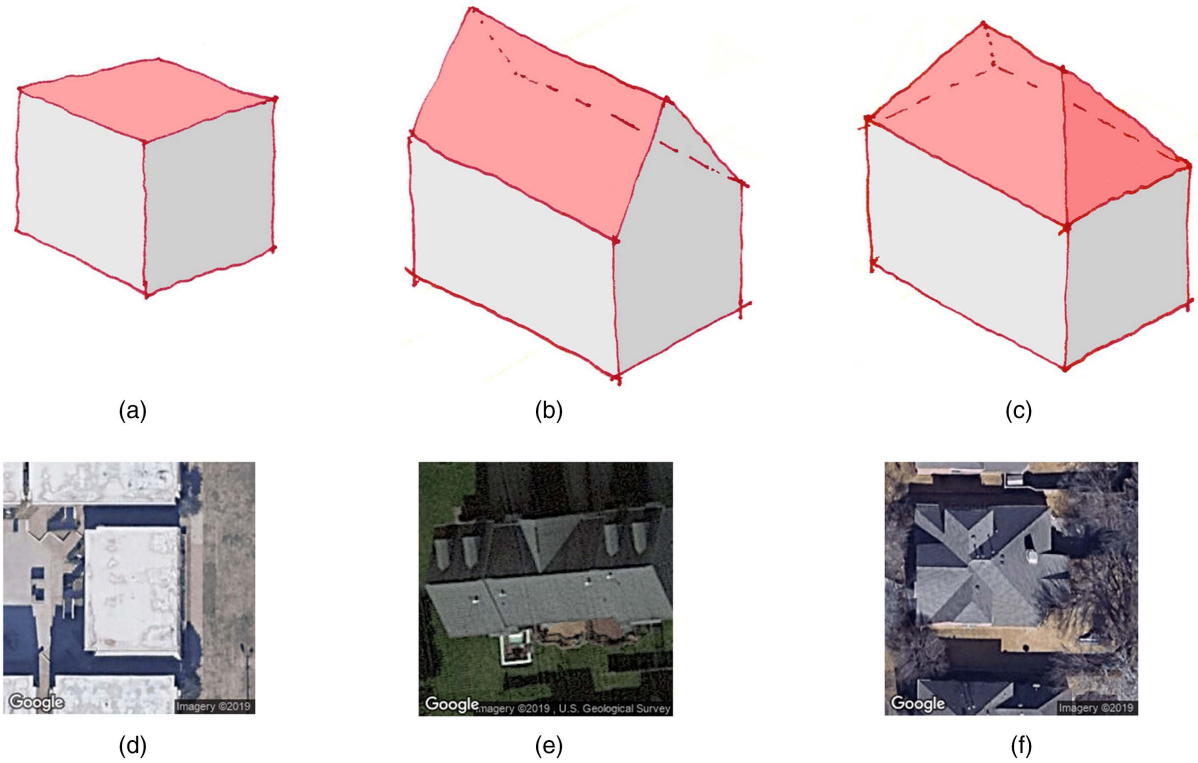


Fig. 3. Roof shape: prototypes and satellite images. There are three different roof types: flat, gabled, and hipped: (a) flat prototype; (b) gabled prototype; (c) hipped prototype; (d) flat; (e) gabled; and (f) hipped. [Images (d, e, and f) map data © 2019 Google.]

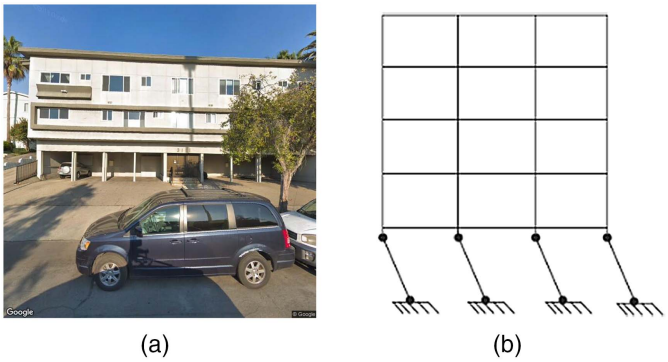


Fig. 4. A street view of a soft-story building and its failure mode: (a) an example of a soft-story building (map data © 2019 Google); and (b) soft-story failure mechanism.

authors (Dosovitskiy et al. 2020). For training the models on the target tasks, we use stochastic gradient descent with momentum (Bottou 2012) as the optimizer. We set the momentum rate as 0.9, the initial learning rate as 0.01, and use a batch size of 64. We train the network with a total of 50 epochs and use stepwise learning rate decay with a decay rate of 0.9 in each step. We use standard data augmentation techniques (Shorten and Khoshgoftaar 2019) such as RandomResizedCrop, RandomHorizontalFlip, and ColorJitter in the training process. The data augmentation techniques can improve the generalization ability of the trained network in order to perform well on images never encountered during training.

For roof-type classification, we use test accuracy as the evaluation metric, which is defined as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{8}$$

For soft-story classification, we use F1 score, precision, and recall (Goutte and Gaussier 2005) as the metric. The precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

where TP = true positive; FP = false positive; and FN = false negative. The F1 score is the harmonic mean of precision and recall. A larger value indicates better performance across all the metrics.

Results and Discussion

Quantitative Results

Vision Transformer Versus ResNet50: In Table 2 we give the comparison of ResNet50 and Vision Transformer for roof-type classification and soft-story classification. From the results we can observe that for roof-type classification, the Vision Transformer achieves 4% higher accuracy compared with ResNet50. For soft-story classification, the Vision Transformer achieves a higher F1 score compared with ResNet50. The results reveal that with the attention modules, the Vision Transformer is more effective than ResNet50 for satellite or street view images. The attention modules can locate the object of interest in the images, as we show in the section "Experiments" by illustrating the attention maps of the Vision Transformer.

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

Table 2. Vision Transformer achieves much better performance compared with ResNet50 for roof-type classification and soft-story classification

	Roof-type classification	Soft-story classification		
Network	Test accuracy (%)	F1	Precision	Recall
ResNet50	91.93	0.85	0.82	0.89
Vision Transformer	95.16	0.89	0.86	0.92

Table 3. The proposed AdaLN achieves comparable performance with Independent Network while requiring significantly fewer parameters for roof-type classification and soft-story classification. Compared with alternative multidomain learning approaches, the proposed AdaLN achieves better performance while being more resource efficient

		Roof-type classification	Soft-story classification		
Method	No. of parameters	Test accuracy (%)	F1	Precision	Recall
Independent Network	2,150×	95.16	0.89	0.86	0.92
Single Mask	11.75×	90.32	0.80	0.77	0.83
Multiple Masks	125×	91.93	0.81	0.80	0.83
AdaLN	1×	93.54	0.89	0.87	0.91

Multidomain Learning with Vision Transformer: In Table 3, we give the results of different multidomain learning approaches. We also show the comparison of the number of parameters with AdaLN as the baseline. From the results we can observe that Single Mask, Multiple Masks, and AdaLN can greatly reduce the number of parameters compared with Independent Network, which demonstrates the benefits of multidomain learning. Specially, the performance of AdaLN approaches Independent Network for soft-story classification with a significant reduction of number of parameters. By comparing Single Mask, Multiple Masks, and AdaLN, we can see that AdaLN achieves better performance while leading to a much smaller model size. This fact shows that for multidomain learning it is critical to identify the task-agnostic parameters and task-specific parameters. By only fine-tuning the layer-normalization parameters as in AdaLN, we effectively reuse the model pretrained on ImageNet while capturing the difference between natural images and satellite (or street view) images.

Qualitative Results

Attention Visualization: We show the examples of the attention maps of Independent Network and AdaLN in Figs. 5 and 6. We can see from the figures that the attention module can extract the part of the image that is critical for the task. For the roof-type classification, the attention map ignores the surroundings of the building and only focus on the roof of the building. For soft-story

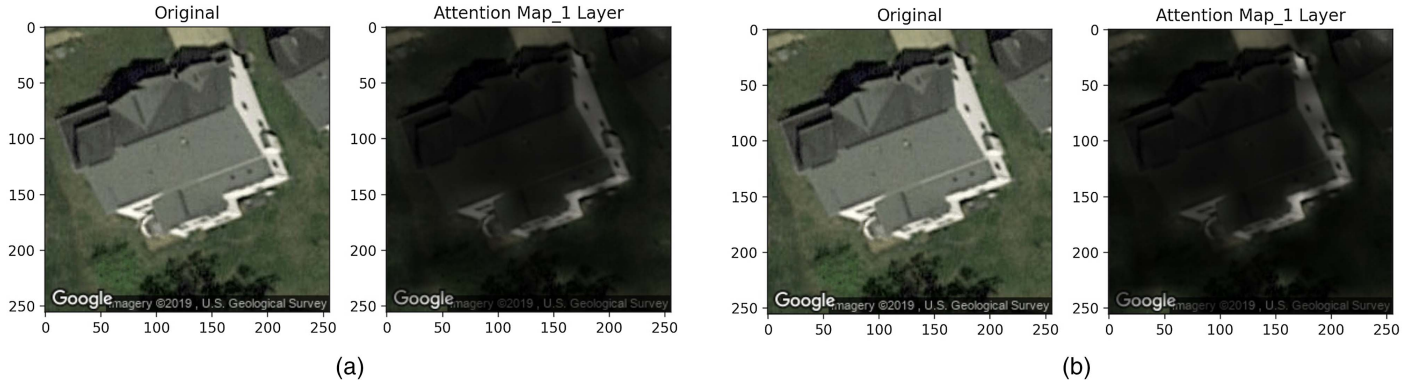


Fig. 5. Examples of the attention maps for roof-type classification: (a) the attention map obtained by Independent Network; and (b) the attention map obtained by *AdaLN*. (Map data © 2019 Google.)

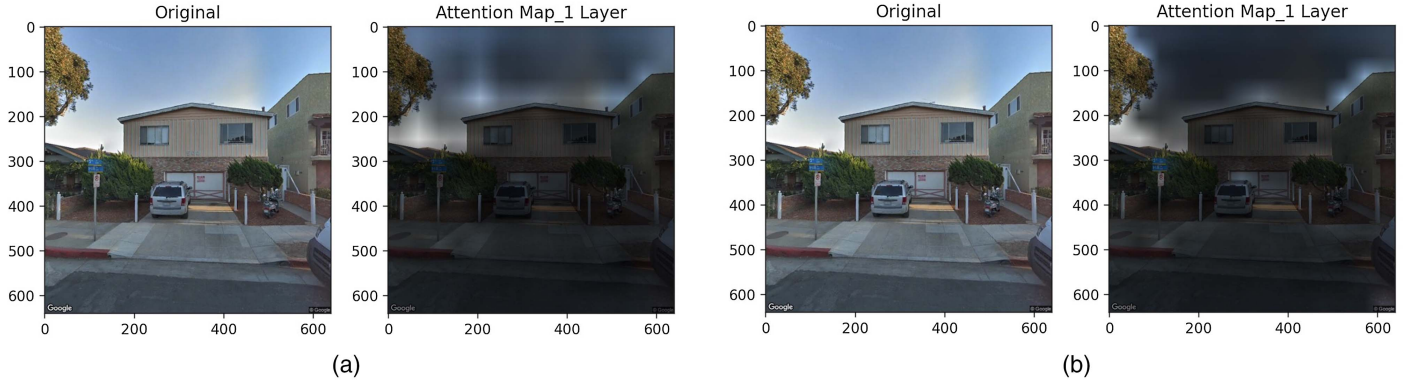


Fig. 6. Examples of the attention maps for soft-story classification: (a) Independent Network; and (b) AdaLN. (Map data © 2019 Google.)

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

classification, the attention map ignores the sky or the road and only focuses on the front of the building. There are typically multiple objects in the satellite or street view images, and with the attention modules, the network can focus on the important part of the image while ignoring background noises. By using much fewer parameters, AdaLN obtains similar attention maps compared with Independent Network. The visualizations again confirm the effectiveness of the proposed AdaLN for multidomain learning with the Vision Transformer.

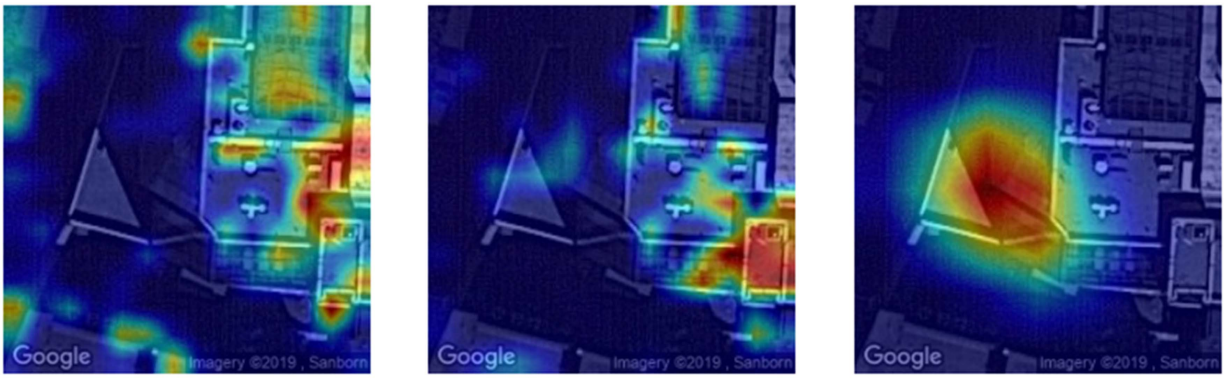
Class Activation Map Visualization: We further compare the class activation map (CAM) (Zhou et al. 2016a) generated by ResNet50, Independent Network, and the proposed AdaLN. CAM is a powerful method in computer vision for understanding which part of the image is responsible for the prediction of the category. CAM can be used for different network architectures, including both CNNs and Vision Transformers. Our purpose of using CAM is to understand why Vision Transformers can obtain higher accuracy than ResNet50.

We take the images for which ResNet50 makes wrong predictions while Vision Transformers make correct predictions. We use CAM to localize the part of the image that is responsible for the prediction. In Figs. 7 and 8 we show the CAMs generated by different approaches. It can be observed that both Independent Network and AdaLN can accurately locate the areas of the image that are critical for the prediction task. This explains the effectiveness of Vision Transformer–based approaches over ResNet50. It is worth noting that the proposed AdaLN can produce a similar CAM to Independent Network with much fewer parameters.

From Figs. 5–8, we can observe that the attention-based models can accurately focus on the object of interest, which is desirable for the tasks. The surprising fact is that without any supervision, the model can still find relevant information for classification, which indicates the effectiveness of data-driven learning.

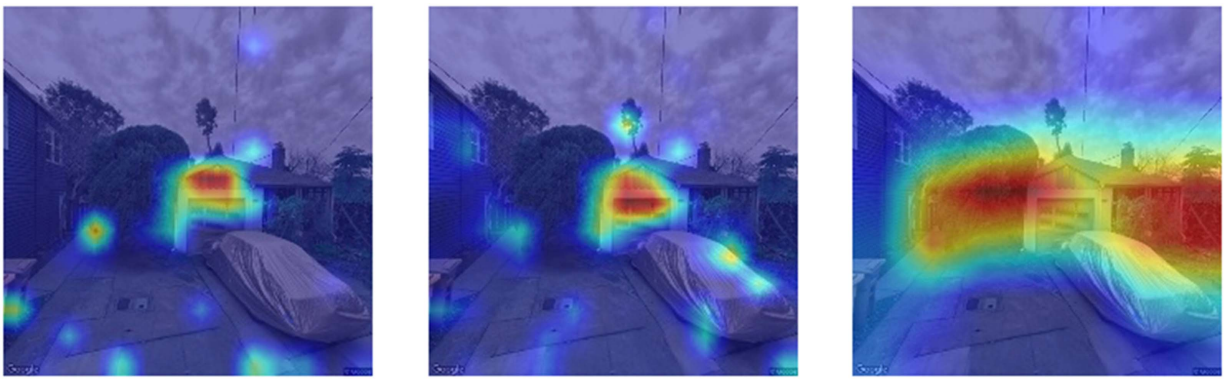
Application

We used the trained model for inferring building information in two populated neighborhoods. The first is Northeast MacFarlane in Tampa in the hurricane nation: Florida, United States. A total of 3,089 buildings were found in this neighborhood. Satellite images of each building were downloaded from Google Maps. Using our model, the roof type of each building was classified based on the satellite images. The result is injected into the building footprints for visualization in Fig. 9. The second is the Buckman neighborhood in Portland, Oregon, United States. Located in the Pacific Northwest seismic zone, Portland is one of the most populated major cities in this region. We collected street view images for 1,639 buildings in this neighborhood from Google Maps. Predictions were executed



(a) (b) (c)

Fig. 7. CAMs generated by different approaches on a sampled satellite image: (a) Independent Network; (b) AdaLN; and (c) ResNet50. (Map data © 2019 Google.)



(a) (b) (c)

Fig. 8. CAMs generated by different approaches on a sampled street view image: (a) Independent Network; (b) AdaLN; and (c) ResNet50. (Map data © 2019 Google.)

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

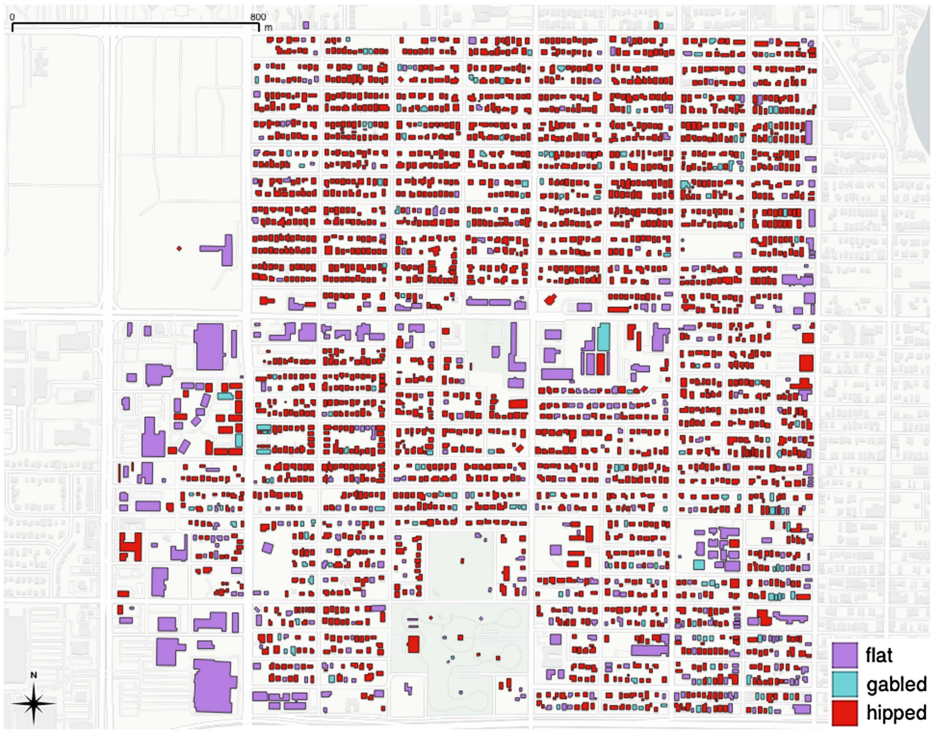


Fig. 9. Roof shape classification from satellite images for Northeast MacFarlane in Tampa, Florida, United States. (© OpenStreetMap contributors.)

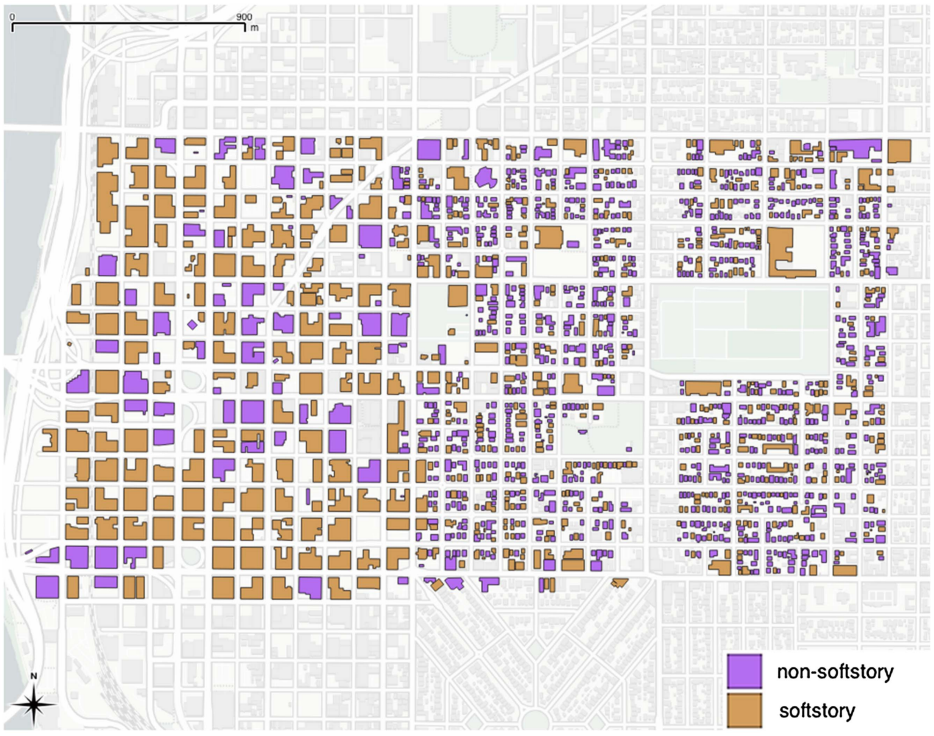


Fig. 10. Soft-story building classification from street view images for Buckman in Portland, Oregon, United States. (© OpenStreetMap contributors.)

for each image and the result is merged with building footprints as shown in Fig. 10. Like Northeast MacFarlane and Buckman, there are many populated neighborhoods located in regions of high natural hazard risks. The approach demonstrated in this study provides a tool for rapid predisaster examination of buildings in a large region at very low cost.

Conclusion

DL is prevalent in remote sensing for extracting information from sensed imagery. However, using this technique to obtain detailed building attributes for hazard risk screening has not been sufficiently explored. The first objective of this study is to explore the possibility

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

of automatic collection of building information from satellite and street view images for assisting decision-making in natural hazard preparedness. The second objective is to develop and validate a unified transformer architecture for different classification tasks of images from multiple domains. The transformer is a new deep neural network architecture. Different from the convolutional operations in CNNs, transformers are based solely on the attention mechanism.

One problem in image and DL-based information acquiring is the multidomain issue. Traditionally, it requires multiple models to be trained for different tasks. To tackle this issue, we developed a new transformer architecture, AdaLN, that can be used to extract information from images in different domains. An advantage of the unified transformer is that we can train on different tasks by only adjusting the layer-normalization parameters while all the other layers remain fixed. This saves significantly on the computation needed for training. We demonstrated that our unified transformer can be used to extract building information from both satellite and street view images. For a demonstration, this unified transformer is tested for two tasks:

- Classifying the building roof information (roof type) based on satellite images; and
- Classifying the building facade information (soft-story feature) based on street view images.

The motivation of this case study is to facilitate large-scale predisaster building information collection, where a single unified model can be used for images from multiple domains, such as satellite and street view.

This work appears to be the first study to use Vision Transformers for this purpose. Compared with the baselines, our model achieved similar performance with extremely reduced parameters for training and storage.

Data Availability Statement

Testing data, trained models, and the codes that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

A part of this study is based on work supported by the National Science Foundation under Grant No. 1612843. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

ATC (Applied Technology Council). 1988. *Rapid visual screening of buildings for potential seismic hazards: A handbook*. FEMA 154. Washington, DC: FEMA.

Ba, J. L., J. R. Kiros, and G. E. Hinton. 2016. "Layer normalization." Preprint, submitted July 21, 2016. http://arxiv.org/abs/1607.06450.

Bency, A. J., S. Rallapalli, R. K. Ganti, M. Srivatsa, and B. Manjunath. 2017. "Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery." In *Proc., IEEE Winter Conf. on Applications of Computer Vision*. New York: IEEE.

Bhandare, A., V. Sripathi, D. Karkada, V. Menon, S. Choi, K. Datta, and V. Saletore. 2019. "Efficient 8-bit quantization of transformer neural machine language translation model." Preprint, submitted June 7, 2019. http://arxiv.org/abs/1906.00532.

Bilen, H., and A. Vedaldi. 2017. "Universal representations: The missing link between faces, text, planktons, and cat breeds." Preprint, submitted January 28, 2017. https://arxiv.org/abs/1701.07275.

Bottou, L. 2012. "Stochastic gradient descent tricks." In *Neural networks: Tricks of the trade*, 421–436. New York: Springer.

Brunner, D., G. Lemoine, and L. Bruzzone. 2010. "Earthquake damage assessment of buildings using VHR optical and SAR imagery." *IEEE Trans. Geosci. Remote Sens.* 48 (5): 2403–2420. https://doi.org/10.1109/TGRS.2009.2038274.

Cha, Y.-J., W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk. 2018. "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types." *Comput.-Aided Civ. Infrastruct. Eng.* 33 (9): 731–747. https://doi.org/10.1111/mice.12334.

Chollet, F. 2017. "Xception: Deep learning with depthwise separable convolutions." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 1251–1258. New York: IEEE.

Czerniawski, T., and F. Leite. 2020. "Automated segmentation of RGB-D images into a comprehensive set of building components using deep learning." *Adv. Eng. Inf.* 45 (Aug): 101131. https://doi.org/10.1016/j.aei.2020.101131.

de Beurs, K. M., N. S. McThompson, B. C. Owsley, and G. M. Henebry. 2019. "Hurricane damage detection on four major Caribbean islands." *Remote Sens. Environ.* 229 (Aug): 1–13. https://doi.org/10.1016/j.rse.2019.04.028.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. "Imagenet: A large-scale hierarchical image database." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 248–255. New York: IEEE.

Dosovitskiy, A., et al. 2020. "An image is worth 16×16 words: Transformers for image recognition at scale." Preprint, submitted October 22, 2020. http://arxiv.org/abs/2010.11929.

FEMA. 2018. *Hazus—MH 2.1 hurricane model technical manual*. Washington, DC: FEMA.

Gao, Y., and K. M. Mosalam. 2018. "Deep transfer learning for image-based structural damage recognition." *Comput.-Aided Civ. Infrastruct. Eng.* 33 (9): 748–768. https://doi.org/10.1111/mice.12363.

Gebru, T., J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. 2017. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." *Proc. Natl. Acad. Sci. U.S.A.* 114 (50): 13108–13113. https://doi.org/10.1073/pnas.1700035114.

Geirhos, R., P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. 2018. "Imagenet-trained CNNs are biased towards texture; Increasing shape bias improves accuracy and robustness." Preprint, submitted November 29, 2018. http://arxiv.org/abs/1811.12231.

Goutte, C., and E. Gaussier. 2005. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." In *Proc., European Conf. on Information Retrieval*, 345–359. New York: Springer.

Guo, J., Q. Wang, Y. Li, and P. Liu. 2020. "Façade defects classification from imbalanced dataset using meta learning-based convolutional neural network." *Comput.-Aided Civ. Infrastruct. Eng.* 35 (12): 1403–1418. https://doi.org/10.1111/mice.12578.

Guo, Y., Y. Li, L. Wang, and T. Rosing. 2019. "Depthwise convolution is all you need for learning multiple visual domains." In Vol. 33 of *Proc., AAAI Conf. on Artificial Intelligence*, 8368–8375. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Hamaguchi, R., and S. Hikosaka. 2018. "Building detection from satellite imagery using ensemble of size-specific detectors." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 187–191. New York: IEEE.

Hang, L., and G. Cai. 2020. "CNN based detection of building roofs from high resolution satellite images." *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-3/W10: 187–192. https://doi.org/10.5194/isprs-archives-XLII-3-W10-187-2020.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep residual learning for image recognition." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*. New York: IEEE.

Ivanovsky, L., V. Khryashchev, V. Pavlov, and A. Ostrovskaya. 2019. "Building detection on aerial images using U-NET neural networks."

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

In *Proc., 24th Conf. of Open Innovations Association (FRUCT)*, 116–122. New York: IEEE.

Jin, X., and C. H. Davis. 2005. "Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information." *EURASIP J. Adv. Signal Process.* 2005 (14): 1–11. https://doi.org/10.1155/ASP.2005.2196.

Joyce, K. E., S. E. Belliss, S. V. Samsonov, S. J. McNeill, and P. J. Glassey. 2009. "A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters." *Prog. Phys. Geogr.* 33 (2): 183–207. https://doi.org/10.1177/0309133309339563.

Kang, J., M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu. 2018. "Building instance classification using street view images." *ISPRS J. Photogramm. Remote Sens.* 145 (Nov): 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.

Karbassi, A., and M. Nollet. 2007. "The adaptation of the FEMA 154 methodology for the rapid visual screening of existing buildings in accordance with NBCC-2005." In *Proc., 9th Canadian Conf. on Earthquake Engineering*, 27–29. Vancouver, BC, Canada: Canadian Association for Earthquake Engineering.

Kucharczyk, M., and C. H. Hugenholtz. 2019. "Pre-disaster mapping with drones: An urban case study in Victoria, British Columbia, Canada." *Nat. Hazards Earth Syst. Sci.* 19 (9): 2039–2051. https://doi.org/10.5194/nhess-19-2039-2019.

Lari, Z., and H. Ebadi. 2007. "Automated building extraction from high-resolution satellite imagery using spectral and structural information based on artificial neural networks." In *Proc., ISPRS Hannover Workshop*. Hannover, Germany: Leibniz Univ. Hannover.

Law, S., B. Paige, and C. Russell. 2018. "Take a look around: Using street view and satellite images to estimate house prices." Preprint, submitted July 18, 2018. http://arxiv.org/abs/:1807.07155.

Linsley, D., J. Kim, V. Veerabadran, C. Windolf, and T. Serre. 2018. "Learning long-range spatial dependencies with horizontal gated recurrent units." In *Proc., 32nd Int. Conf. on Neural Information Processing Systems*, 152–164. Red Hook, NY: Curran Associates.

Lipton, Z. C., et al. 2015. "A critical review of recurrent neural networks for sequence learning." Preprint, submitted May 29, 2015. http://arxiv.org/abs/1506.00019.

Liu, X., Q. Chen, L. Zhu, Y. Xu, and L. Lin. 2017. "Place-centric visual urban perception with deep multi-instance regression." In *Proc., 25th ACM Int. Conf. on Multimedia*. New York: Association for Computing Machinery.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient estimation of word representations in vector space." Preprint, submitted January 16, 2013. http://arxiv.org/abs/1301.3781.

Naik, N., J. Philipoom, R. Raskar, and C. Hidalgo. 2014. "Streetscore–Predicting the perceived safety of one million streetscapes." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. New York: IEEE.

NCEI (National Centers for Environmental Information). 2021. *US billion-dollar weather and climate disasters*. Washington, DC: National Oceanic and Atmospheric Administration.

Ningthoujam, M., and R. P. Nanda. 2018. "Rapid visual screening procedure of existing building based on statistical analysis." *Int. J. Disaster Risk Reduct.* 28 (Jun): 720–730. https://doi.org/10.1016/j.ijdrr.2018.01.033.

Paszke, A., et al. 2019. "Pytorch: An imperative style, high-performance deep learning library."http://arxiv.org/abs/1912.01703.

Patino, J. E., and J. C. Duque. 2013. "A review of regional science applications of satellite remote sensing in urban settings." *Comput. Environ. Urban Syst.* 37 (Jan): 1–17. https://doi.org/10.1016/j.compenvurbsys.2012.06.003.

Perrone, D., M. A. Aiello, M. Pecce, and F. Rossi. 2015. "Rapid visual screening for seismic evaluation of RC hospital buildings." In Vol. 3 of *Structures*, 57–70. Amsterdam, Netherlands: Elsevier.

Ploeger, S., M. Sawada, A. Elsabbagh, M. Saatcioglu, M. Nastev, and E. Rosetti. 2016. "Urban RAT: New tool for virtual and site-specific mobile rapid data collection for seismic risk assessment." *J. Comput. Civ. Eng.* 30 (2): 04015006. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000472.

Poursanidis, D., and N. Chrysoulakis. 2017. "Remote sensing, natural hazards and the contribution of ESA sentinels missions." *Remote Sens. Appl. Soc. Environ.* 6 (Apr): 25–38. https://doi.org/10.1016/j.rsase.2017.02.001.

Rebuffi, S.-A., H. Bilen, and A. Vedaldi. 2017. "Learning multiple visual domains with residual adapters." In *Advances in neural information processing systems*, 506–516. Red Hook, NY: Curran Associates.

Rebuffi, S.-A., H. Bilen, and A. Vedaldi. 2018. "Efficient parametrization of multi-domain deep neural networks." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 8119–8127. New York: IEEE.

Rosenfeld, A., and J. K. Tsotsos. 2017. "Incremental learning through deep adaptation." Preprint, submitted May 11, 2017. http://arxiv.org/abs/1705.04228.

Saatcioglu, M., M. Shooshtari, and S. Foo. 2013. "Seismic screening of buildings based on the 2010 National Building Code of Canada." *Can. J. Civ. Eng.* 40 (5): 483–498. https://doi.org/10.1139/cjce-2012-0055.

Shorten, C., and T. M. Khoshgoftaar. 2019. "A survey on image data augmentation for deep learning." *J. Big Data* 6 (1): 1–48. https://doi.org/10.1186/s40537-019-0197-0.

Srikanth, T., R. P. Kumar, A. P. Singh, B. K. Rastogi, and S. Kumar. 2010. "Earthquake vulnerability assessment of existing buildings in Gandhidham and Adipur cities Kachchh, Gujarat (India)." *Eur. J. Sci. Res.* 41 (3): 336–353.

Sun, F., J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. 2019. "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer." In *Proc., 28th ACM Int. Conf. on Information and Knowledge Management*, 1441–1450. New York: Association for Computing Machinery.

Torok, M. M., M. Golparvar-Fard, and K. B. Kochersberger. 2014. "Image-based automated 3D crack detection for post-disaster building assessment." *J. Comput. Civ. Eng.* 28 (5): A4014004. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000334.

Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. 2020. "Training data-efficient image transformers & distillation through attention."Preprint, submitted December 23, 2020. http://arxiv.org/abs/2012.12877.

Ulyanov, D., A. Vedaldi, and V. Lempitsky. 2016. "Instance normalization: The missing ingredient for fast stylization." Preprint, submitted July 27, 2016. http://arxiv.org/abs/1607.08022.

Vaswani, A., et al. 2017. "Attention is all you need." In *Advances in neural information processing systems*, 5998–6008. Red Hook, NY: Curran Associates.

Wallace, N. M., and T. H. Miller. 2008. "Seismic screening of public facilities in Oregon's western counties." *Pract. Period. Struct. Des. Constr.* 13 (4): 189–197. https://doi.org/10.1061/(ASCE)1084-0680(2008)13:4(189).

Wallemacq, P., and R. House. 2018. *Economic losses, poverty and disasters: 1998-2017*. Technical report. Brussels, Belgium: United Nations Office for Disaster Risk Reduction.

Wang, C., Q. Yu, K. H. Law, F. McKenna, X. Y. Stella, E. Taciroglu, A. Zsarnóczay, W. Elhaddad, and B. Cetiner. 2021a. "Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management." *Autom. Constr.* 122 (Feb): 103474. https://doi.org/10.1016/j.autcon.2020.103474.

Wang, F., M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. 2017. "Residual attention network for image classification." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 3156–3164. New York: IEEE.

Wang, N., Q. Zhao, S. Li, X. Zhao, and P. Zhao. 2018. "Damage classification for masonry historic structures using convolutional neural networks based on still images." *Comput.-Aided Civ. Infrastruct. Eng.* 33 (12): 1073–1089. https://doi.org/10.1111/mice.12411.

Wang, N., X. Zhao, Z. Zou, P. Zhao, and F. Qi. 2020a. "Autonomous damage segmentation and measurement of glazed tiles in historic buildings via deep learning." *Comput.-Aided Civ. Infrastruct. Eng.* 35 (3): 277–291. https://doi.org/10.1111/mice.12488.

Wang, W., E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. 2021b. "Pyramid vision transformer: A versatile backbone for

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.

dense prediction without convolutions." Preprint, submitted February 24, 2021. http://arxiv.org/abs/2102.12122.

Wang, X., C. Wittich, T. Hutchinson, Y. Bock, D. Goldberg, E. Lo, and F. Kuester. 2020b. "Methodology and validation of UAV-based video analysis approach for tracking earthquake-induced building displacements." *J. Comput. Civ. Eng.* 34 (6): 04020045. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000928.

Wang, Y., Q. Yao, J. T. Kwok, and L. M. Ni. 2020c. "Generalizing from a few examples: A survey on few-shot learning." *ACM Comput. Surv.* 53 (3): 1–34. https://doi.org/10.1145/3386252.

Woo, S., J. Park, J.-Y. Lee, and I. S. Kweon. 2018. "CBAM: Convolutional block attention module." In *Proc., European Conf. on Computer Vision (ECCV)*, 3–19. Cham, Switzerland: Springer.

Wu, H., B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. 2021. "CVT: Introducing convolutions to vision transformers." Preprint, submitted March 29, 2021. http://arxiv.org/abs/2103.15808.

Yu, Q., C. Wang, F. McKenna, S. X. Yu, E. Taciroglu, B. Cetiner, and K. H. Law. 2020. "Rapid visual screening of soft-story buildings from street view images using deep learning classification." *Earthquake Eng. Eng. Vibr.* 19 (4): 827–838. https://doi.org/10.1007/s11803-020-0598-2.

Zhang, Y. 1999. "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering." *ISPRS J. Photogramm. Remote Sens.* 54 (1): 50–60. https://doi.org/10.1016/S0924-2716(98)00027-6.

Zheng, H., J. Fu, T. Mei, and J. Luo. 2017. "Learning multi-attention convolutional neural network for fine-grained image recognition." In *Proc., IEEE Int. Conf. on Computer Vision*, 5209–5217. New York: IEEE.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016a. "Learning deep features for discriminative localization." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*. New York: IEEE.

Zhou, Z., J. Gong, and M. Guo. 2016b. "Image-based 3D reconstruction for posthurricane residential building damage assessment." *J. Comput. Civ. Eng.* 30 (2): 04015015. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000480.

Zhu, Z., S. German, and I. Brilakis. 2011. "Visual retrieval of concrete crack properties for automated post-earthquake structural safety evaluation." *Autom. Constr.* 20 (7): 874–883. https://doi.org/10.1016/j.autcon.2011.03.004.

Downloaded from ascelibrary.org by Chaofeng Wang on 07/04/22. Copyright ASCE. For personal use only; all rights reserved.