DevianceNet: Learning to Predict Deviance from a Large-Scale Geo-Tagged Dataset

Jin-Hwi Park[*1], **Young-Jae Park**[*1], **Junoh Lee**[2] and **Hae-Gon Jeon**[1,2]

[1]AI Graduate School, GIST, South Korea
[2]School of Electrical Engineering and Computer Science, GIST, South Korea
{jinhwipark, youngjae.park, juno}@gm.gist.ac.kr, haegonj@gist.ac.kr

Abstract

Understanding how a city's physical appearance and environmental surroundings impact society traits, such as safety, is an essential issue in social artificial intelligence. To demonstrate the relationship, most existing studies utilize subjective human perceptual attributes, categorization only for a few violent crimes, and images taken from still shot images. These lead to difficulty in identifying location-specific characteristics for urban safety. In this work, to address this problem, we propose a large-scale dataset and a novel method by adopting a concept of "Deviance" which explains behaviors violating social norms, both formally (e.g. crime) and informally (e.g. civil complaints). We first collect a geo-tagged dataset consisting of incident report data for seven metropolitan cities, with corresponding sequential images around incident sites obtained from Google street view. We also design a convolutional neural network that learns spatio-temporal visual attributes of deviant streets. Experimental results show that our framework can reliably recognize real-world deviance in various cities. Furthermore, we analyze which visual attribute is important for deviance identification and severity estimation. We have released our dataset and source codes at our project page: https://deviance-project.github.io/DevianceNet/.

1 Introduction

Urban planners and policymakers have traditionally utilized social sciences, such as economics, criminology, and sociology, to inform their decisions. Identifying location specific attributes is particularly important for urban safety development. The most relevant research can be categorized into two classes: micro-level (Arietta et al. 2014; Naik et al. 2014; Porzi et al. 2015) and macro-level approaches (Suel et al. 2019; Maharana, Nguyen, and Nsoesie 2019; Alves, Ribeiro, and Rodrigues 2018).

The micro-level approaches focus on only parts of visible cues from disordered environments (e.g., broken windows and graffiti) where may affect perceived safety of places (Kelling, Wilson et al. 1982). One notable example which utilizes the perceived safety is "Deep Learning the City" proposed by Dubey (Dubey et al. 2016). With a Place Pulse 2.0 dataset consisting of pairwise image comparison data obtained through crowdsourcing, a convolutional neural network (CNN) was trained to predict perceived safety and other visual attributes of cities. However, these meth-

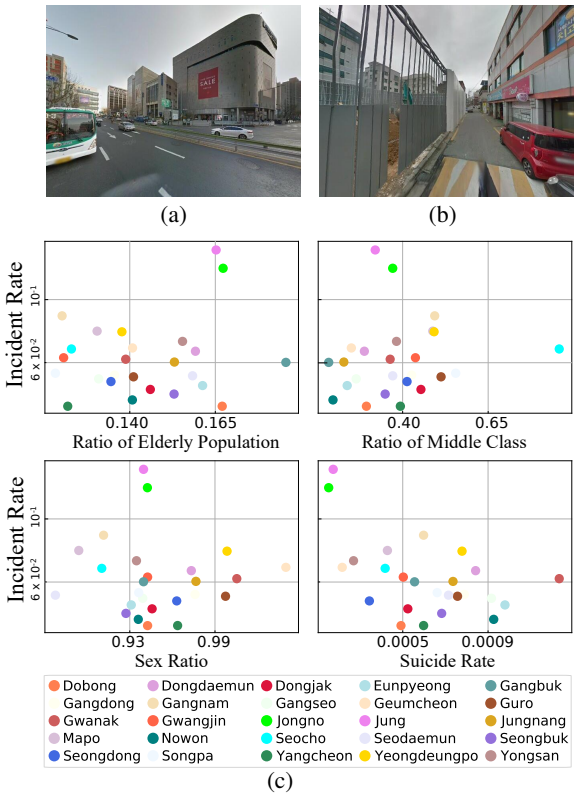


Figure 1: Which place looks more dangerous? People and existing model (Dubey et al. 2016) perceive the place (a) as safe, but actual crimes have occurred. By contrast, the place (b) seems to be dangerous due to the construction site. Surprisingly, there is no crime occurrence. Based on official crime statistics of all 25 local districts in Seoul, we report relationships between four major sociodemographic indicators and incident rates, a ratio of incident to a population of each district, in (c). The scatter plots represent no correlation exists between them.

* J.-H. Park and Y.-J. Park provided equal contributions to this work.
Corresponding author: Hae-Gon Jeon

Dataset	Input	Crime label	Image#	Viewpoint	GPS	Open	City#
PlacePulse1 (Salesses et al. 2013)	Single Image	Crowdsource Perceived Safety	4K	Street-level	X	O	4
PlacePulse2 (Dubey et al. 2016)	Single Image	Crowdsource Perceived Safety	110K	Street-level	X	O	56
Crime-Rate (Andersson, Birck, and Araujo 2017)	4-directional Images	Crowdsource Incident Report	83K	Street-level	O	X	1
Satellite (Najjar, Kaneko, and Miyanaga 2018)	Single Image	Official Incident Report	57K	Satellite-level	O	X	3
StreetNet (Fu, Chen, and Lu 2018)	Single Image	Official Incident Report	44K	Street-level	O	X	2
UK Dataset (Suel et al. 2019)	4-directional Images	Official Socio-demographics	1561K	Street-level	X	X	4
Ours	12-directional Sequential Image	Official Incident Report	760K	Street-level	O	O	7

Table 1: Dataset comparison. GPS indicates whether images and incident labels are geo-tagged together, or not.

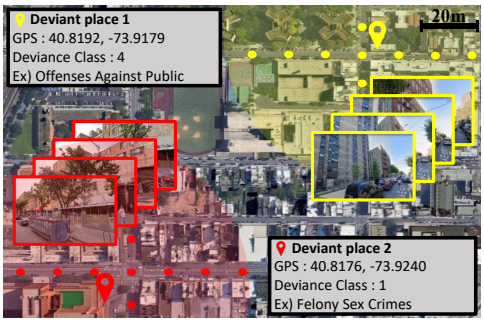


Figure 2: Dataset acquisition. Given official incident report data consisting of GPS locations and incident types, we take sequential images around an incident site and classify it into one of the deviance categories. The coverage of a deviant place is 100m.

ods rely on subjective attributes such as safe, lively, wealthy, and boring. Since disorderliness depends on the perception of survey respondents (Gau and Pratt 2010), corresponding perceived safety has no relation to actual crime occurrences (Keizer, Lindenberg, and Steg 2008). Figure 1(a) provides an example of place that is perceived safe, but violent crimes frequently occur. By contrast, even though Figure 1(b) looks like a dangerous place, the place is safe in reality. The gap between reality and human perceptions results from using single images with a specific viewpoint or 4-direction images which provide insufficient coverage of an incident sites' visual appearance.

On the other hand, the macro-level approaches leverage sociodemographic information which is mainly based on statistical data such as income, crime, and region population. (Suel et al. 2019) developed a CNN that measures the spatial distributions of education, health, unemployment, housing, living environment, and crime from raw images. However, there is no significant correlation between actual crime occurrences and sociodemographic information. As an example, we plot the relationship between sociodemographic information (i.e., elderly population, middle-income class, sex ratio, and suicide rate) of all 25 local districts in Seoul and incident rates in 2018 in Figure 1(c).

To address these challenges, we introduce a large database of geo-tagged images at a city-scale based on "Deviance" which explains deviant incidents violating social norms, both formally (e.g., crime) and informally (e.g., civil complaints). Our dataset contains objective incident report data for seven metropolitan cities with various incident types including violent crimes and civil complaints collected from government agencies, and their corresponding sequential images from Google street view. The images cover the entire street, and are not limited to individual viewpoints, whose example is illustrated in Figure 2. Since our dataset is the first deviance dataset which contains sequential images based on objective incident reports, there is no dataset that can be directly compared with ours. To highlight the novelty of our dataset, we summarize the attributes of relevant datasets in Table 1.

With our dataset, we design a CNN framework, called *DevianceNet*. DevianceNet can identify potential deviant places and their dangerousness from sequential images. Since existing video understanding models with sequential images are not suitable to handle large gaps between image frames obtained from Google street view, we use an interest point matching to find reliable correspondences between associated descriptors (DeTone, Malisiewicz, and Rabinovich 2018). In addition, since not all crimes are equivalent to one another, many works (Kwan, Ip, and Kwan 2000; Huey 2016; Ratcliffe 2015; Koss, Woodruff, and Koss 1990) impose different weights on the severity of crime types. Inspired by statistical theory (Hayhurst 1932), we propose a severity-aware loss for deviance prediction.

Using DevianceNet and the severity-aware loss, we obtain state-of-the-art results for various places. In particular, our network shows consistently promising performances for seven different cities in South Korea and the US. Additionally, we investigate whether the DevianceNet trained on only one city is transferable to other cities, indicating the extent to which visual attributes linked to measures of deviance are shared between cities. Ablation studies also indicate that each of these technical contributions leads to appreciable improvements in deviance prediction. Lastly, we conduct various analyses to understand roles of visual attributes which affect deviance occurrences.

2 The Proposed Approach

In this work, we aim to learn a CNN to predict deviance classes from visual attributes of streets. For this, we need to overcome three main challenges: (1) the huge gap between actual crime occurrences and perceived safety, (2) the representation of location-specific attributes, and (3) a consideration of the severity of crimes.

To address these challenges, we firstly construct a large-scale dataset consisting of sequential images corresponding to actual crime reports, which takes advantage of both the objectiveness of macro-level approaches and local specific

Class #	South Korea	Chicago	New York
1	Homicide	Homicide	Murder & Non-negligent manslaughter
2	Grand theft	Motor vehicle theft	Grand larceny of motor vehicle
3	Demoralization	Public indecency	Disorderly conduct
4	Public peace violation	Interference with public officer	Offenses against public administration
5	-	Non-Deviance	-

Table 2: Incident data categorization. Incident types are categorized into 4 classes based on criteria of incident reports for each city. The class 5 is for negative samples, which does not include any incident.

Class #	Seoul	Busan	Deajeon	Daegu	Incheon	NewYork	Chicago
Class 1	3369 (672)	648 (156)	1044 (240)	946 (204)	1124 (204)	1013 (245)	1044 (324)
Class 2	3486 (742)	852 (204)	1164 (228)	993 (204)	1164 (276)	977 (180)	1056 (240)
Class 3	3731 (816)	833 (180)	1421 (228)	1095 (267)	1200 (300)	988 (264)	1076 (240)
Class 4	3109 (705)	801 (168)	1130 (262)	1048 (192)	1080 (264)	984 (216)	1179 (226)
Class 5	3313 (720)	804 (144)	804 (156)	916 (195)	924 (168)	870 (205)	932 (240)

Table 3: The number of clips of our dataset. $(\cdot)$ indicates the number of clips in test set. Seoul contains 150 deviant places for each deviance class and other cities have 50 places.

representations of micro-level approaches. We then build a CNN model to predict deviance classes from input sequential images. Lastly, we propose a novel loss function to cope with the difference of severity among deviance classes because violent crimes such as murder cases are much serious than civil complaints in the real-world.

Through our dataset and DevianceNet, we focus on two tasks: deviance identification and severity estimation. A goal of the identification task is to detect whether an event of deviance appears in a sequential image. The task is actually binary classification like RSS-CNN (Dubey et al. 2016). We also perform a 5-way classification task which estimates severities of deviances.

2.1 Dataset Construction

Here, we introduce our novel large-scale dataset consisting of objective incident report data with its corresponding sequential images to fully represent the visual attributes of deviant locations. The Deviance dataset is based on official incident report data of South Korea and the US for 2018. The report data of South Korea is provided by the National Police Agency, and the US data is collected from official open data portal of Chicago[1] and New York[2]. The reports consist of many different types of incidents including violent crimes and civil complaints. Based on the reports, our dataset is collected from 5 major cities in South Korea (Seoul, Busan, Incheon, Daegu, and Daejeon) and 2 major cities in the US (Chicago and New York) in the following steps:

1. We first categorize incident types into four classes according to severities of deviances. We follow the incident classification criteria of incident reports for each city (e.g., criminal classification codes and levels of offense) in Table 2. In addition, we add a non-deviance class, which includes places where no deviance has occurred.
2. We then sort out the deviant places for each class, where deviance frequently occurred at the GPS-level. To avoid a vagueness and a class imbalance, we exclude places which have less than five occurrences. For further explanation, we report the percentage of deviant places according to incident occurrences in Seoul where 95% of places have less than five occurrences.
3. Based on the selected deviant places, we obtain surrounding Google Street View images of each deviant place. At least 10 GPS coordinates within a radius of 50m of a sorted deviant location are selected to consider entire neighborhood environments. The reason for the coverage is that we adopt a standard range for urban environments and planning with location-specific attributes (Özbil, Yeşiltepe, and Argin 2015; Gorgul et al. 2019). From the selected GPS coordinates, we collect images with 12 directions for each GPS position. As a result, each deviant place has at least 120 images in total.

We extract a total of 2,250 deviant places, consisting of

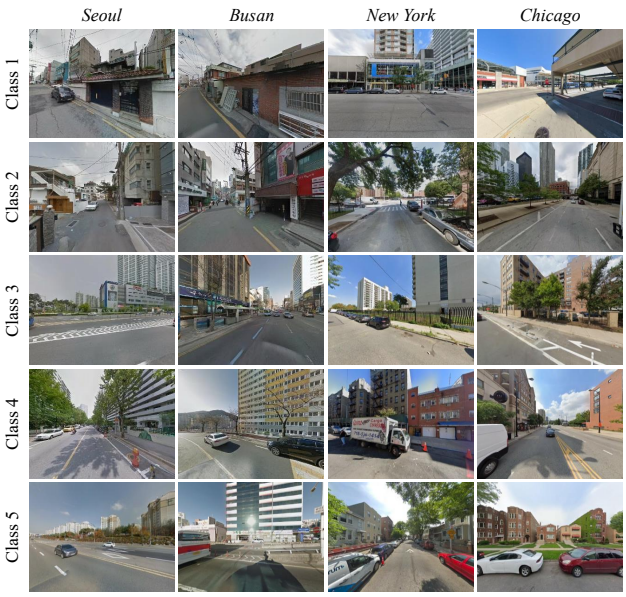


Figure 3: Examples of our dataset. Each row and column indicates a deviance class and its corresponding city, respectively. The examples have visually similar attributes, but there are differences between actual crime levels. This means that people's perception of urban appearance often fails to take account of it.

[1]Chicago Data Portal: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[2]NYC Open Data: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

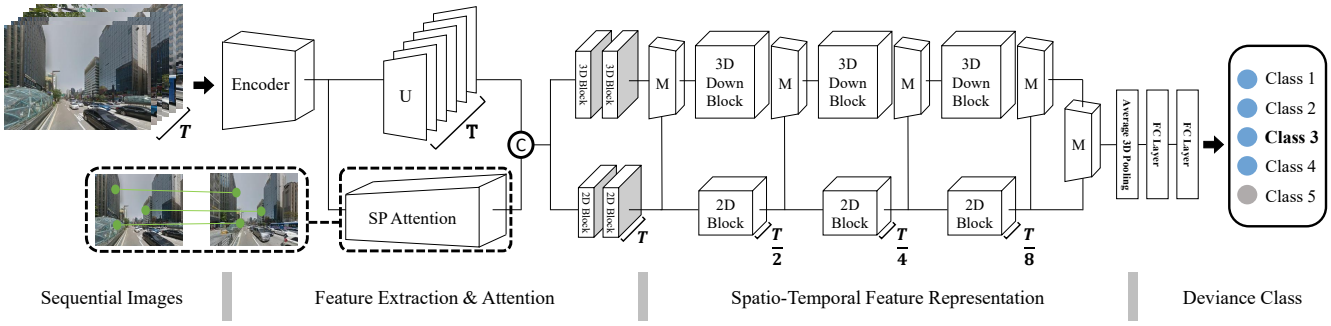


Figure 4: An overview of DevianceNet. T is the number of input images, U is an upsampling block, C is a concatenation, and SP Attention means an attention module based on an interest point matching network. The two branches with 3D block, 2D blocks learn generic representations and, specifically, 3D Down Block is a reduction block for temporal and channel dimension. M is a merging block that fuses the outputs of 3D block and 2D block.

760,952 images. The number of sequential image clips for the training and test set is 46,630 and 10,275, respectively. We note that the test set are unseen places in training set, whose details are reported in Table 3. The same number of deviant places were selected for each deviance class, and none of the deviant places overlaps. In the identification task, clips of deviant places for each deviance class are randomly selected as many as non-deviant places in the test set for data balance. The examples of our dataset are displayed in Figure 3.

2.2 Framework Architecture

We train a spatio-temporal network to learn deviance judgments, which is an extended concept of crimes, from input sequential images, and encourage it by introducing a linear combination of identification and severity estimation losses based on sociology (Hayhurst 1932; Carine and Park 2019).

Accompanying the concept of deviance and our dataset, we design a CNN which infers different types of deviance from sequential street-level images (Figure 4). Although existing works, which mainly deal with violent crimes, only focus on a small part of the whole street such as graffiti and broken windows, the overall surrounding environments are needed to be considered to find visual attributes which affect an individual's deviance. Accordingly, holistic representations for input sequential images should be learned for deviance representation.

There have been several works regarding video understanding which mainly focus on small parts of scenes where many changes occur. In contrast, we choose Holistic Appearance and Temporal Network (HATNet) (Diba et al. 2020) as a baseline of our architecture. HATNet learns a holistic representation by merging outputs of 2D and 3D convolution blocks at intermediate stages. The 2D convolutional blocks capture static cues from single frames, and the 3D convolutional blocks extract relative temporal information between frames. By fusing feature maps from these blocks, HATNet learns spatio-temporal representations. However, the 3D convolutional block requires many learnable parameters, which causes unstable training and overfitting problems. To address this issue, we decompose each 3D convolution block into 2D and 1D convolution blocks (Tran et al. 2018). Despite the same number of parameters, it doubles the nonlinear activation between the 2D and 1D convolution in each block. This leads to learn more complex structures in the data that can be represented separately, and to make the optimization tractable.

Due to the nature of Google street view, our dataset has large viewpoint gaps between frames, compared to regular videos. We overcome the limitation by adopting an interest point matching network (Sinha et al. 2020). We bring this idea from a recent work (Sarlin et al. 2019) on scene consistency between consecutive frames for visual localization. With the matched features and descriptors from the interest point matching, DevianceNet enables to capture temporal coherency of the image sequences with the large gap.

2.3 Severity-Aware Loss with Heinrich Weight

Violent crimes and civil complaints vary in severity; therefore, it is necessary to consider the severity of each deviance class. We design an effective loss function to enhance the discriminative power of learned features from DevianceNet.

The proposed loss function is a linear combination of severity estimation loss L_S and identification loss L_I:

$$Loss = L_S + \lambda_1 L_I, \tag{1}$$

where λ_1 is a scale factor to properly balance the expectation values from the deviance severity estimation and the identification errors. In this work, we use binary cross-entropy as the identification loss L_I to determine whether deviance occurs in the scenes.

The severity estimation loss L_S is calculated to reflect different severities among deviance classes. We define relatively incidental deviance classes with less severe incidents as prior classes. For example, in the case of class 2, prior classes indicate less severe classes (i.e., class 3,4 and 5). To incorporate the severity of deviance into our loss function, we modify a cross-entropy loss as below:

$$L_S = ylog(\hat{y}) + \lambda_2 H \tag{2}$$

$$H = \sum_{i=1}^{3} h_i[y_{i,prior} log(\hat{y}_{i,prior})] \quad (3)$$

where y and $\hat{y}$ are a ground-truth and prediction, respectively. In addition, $y_{1,prior}$ and $\hat{y}_{1,prior}$ are an indicator and a prediction of a prior class which is the closest class to the target class, respectively. H is a regularization term, which is the summation of the log probabilities of prior classes with a set of scalar values h_i. λ_2 is also a balance term between the cross-entropy and H.

We determine the weight values h_i using *Heinrich's law* in (Hayhurst 1932). According to the law, for every accident that causes a major injury there are 29 accidents that cause minor injuries, and 300 accidents that cause no injuries. The law is also applicable for analyzing man-made disasters including crime (Carine and Park 2019). Based on the statistical background, we set the weight values to $h_1 = 0.909(= \frac{300}{330})$, $h_2 = 0.088(= \frac{29}{330})$ and $h_3 = 0.003(= \frac{1}{330})$ because our dataset is classified into four classes except for the non-deviance class. The regularization term H based on Heinrich's law imposes a penalty for unrelated classes.

3 Experiments and Results

DevianceNet is evaluated under four different perspectives. First of all, we demonstrate the effectiveness of DevianceNet by comparing with state-of-the-art methods and by exhibiting its generality. Second, we show its transferability by training it on one city data (Seoul) whose validation is carried out of other cities. Third, we perform an extensive ablation study to examine the effects of different components on DevianceNet performance including the proposed loss function. Lastly, dominant visual attributes are analyzed to understand what is more or less salient for deviance prediction.

We follow evaluation manners of RSS-CNN (Dubey et al. 2016) and SEHNet (Suel et al. 2019). We use quantitative measures of visual perception: deviance Severity Estimation Accuracy (SEA), Deviance Identification Accuracy (DIA), and Mean Absolute Error (MAE). The SEA is the percentage of correctly predicted deviance class. The DIA indicates the percentage of whether a given image is correctly determined as a deviant place or not. Lastly, the MAE is an error margin among deviant classes. We compute the difference of class index between GT and prediction in terms of the severity-classification. For example, the MAE is 3 if a model infers a deviance class 4 while its GT deviance class is 1.

3.1 Implementation Details

We implemented our model for 100K iterations using the publicly available PyTorch framework with a batch-size of 32 and an ADAM optimizer with a learning rate of 0.0001 (β_1 = 0.9, β_2 = 0.999), which takes about 8 hours with two NVIDIA RTX 3090 with 24GB memory. An inference time for one sequential image is about 0.1 seconds.

We use a pretrained weight of the interesting point matching (DeTone, Malisiewicz, and Rabinovich 2018) whose number of points and detection threshold are 256 and 0.0005, respectively. Both $2D$ and $(2+1)D$ convolutional blocks used are constructed with ResNet 18-layer as a backbone.

All the experiments are performed with image sequences with 16 frames. To handle optical distortions of Google street view images, we crop the center part of an original image into 480×640 and downsample it into 224×224 in both training and test phase. We set λ_1 and λ_2 in our severity-aware loss to 0.5 and 0.15, respectively. In addition, we set h_i of the loss based on weight values of Heinrich's law. Further details of ablation study for the parameter selection are provided in our supplementary material.

Method	Severity-Aware			Cross-Entropy		
	SEA	DIA	MAE	SEA	DIA	MAE
RSS-CNN (R18)	31.20	-	1.25	30.40	-	1.19
RSS-CNN (R50)	32.60	-	1.22	32.40	-	1.22
SEHNet (R18)	41.18	-	1.09	38.15	-	1.15
SEHNet (R50)	41.80	-	1.10	39.83	-	1.14
I3D	41.96	77.12	1.08	40.74	77.88	1.07
C3D	40.17	68.24	1.13	38.98	68.39	1.24
R3D	40.54	76.51	1.15	39.89	77.57	1.92
R2D+LSTM	42.96	80.82	1.04	41.37	75.65	1.05
R2D+Concat	38.77	73.11	1.19	39.02	75.90	1.16
R2D+Mean	40.55	71.79	1.10	39.52	74.18	1.12
R(2+1)D (R18)	41.39	75.44	1.07	40.71	76.97	1.08
R(2+1)D (R50)	42.50	78.89	1.02	41.90	77.17	1.03
HATNet	44.01	83.34	0.99	43.07	80.52	1.01
HATNet+SP	45.08	84.42	0.99	43.86	79.54	1.04
HATNet+(R2+1)D	46.64	87.97	0.95	44.45	81.79	1.02
DevianceNet	**48.17**	**89.77**	**0.88**	**47.17**	**82.51**	**1.01**

Table 4: Quantitative evaluation. We compare DevianceNet with existing video understanding methods and location-specific attributes inference models. We also apply each component of DevianceNet into HATNet to study how it affects overall result including severity-aware loss.

3.2 Comparisons with State-of-the-Art Methods

We compare DevianceNet with state-of-the-art methods including location-specific attribute prediction methods (Dubey et al. 2016; Suel et al. 2019) with advanced backbones (ResNet18 and ResNet50). Similar to DevianceNet, they aim to represent attributes of locations from images. RSS-CNN (Dubey et al. 2016) and SEHNet (Suel et al. 2019) use a single image and 4 direction images per place as input, respectively. In addition, because our dataset consists of sequential images, we also compare it with state-of-the-art video understanding models such as RGB-I3D (Carreira and Zisserman 2017), C3D (Tran et al. 2015), ResNet3D (Hara, Kataoka, and Satoh 2017), multi-view recognition models (i.e., LSTM, mean, and concatenation) (Facil et al. 2019), R(2+1)D (Tran et al. 2018) and HATNet (Diba et al. 2020). In this experiment, we train DevianceNet and other methods from scratch, and the results are reported in Table 4.

Our DevianceNet provides the best performance on all measures. Interestingly, the video understanding methods show slightly better performance than the location-specific

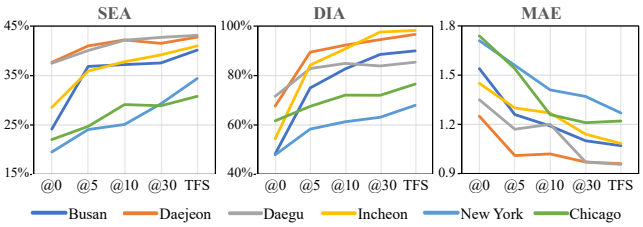


Figure 5: Transferability in different cities. Performance of DevianceNet trained on Seoul in South Korea and applied to other South Korea cities (i.e., Busan, Incheon, Daegu, and Daejeon) and US cities (i.e., Chicago and New York). @A means A% of the dataset for each city, and TFS indicates "train from scratch".

Figure 6: Results on Place Pulse 2.0 dataset. We test our DevianceNet on the images captured in Hongkong and Warsaw, and confirm the generality of DevianceNet.

attributes. However, there are huge gaps between DevianceNet and the video understanding methods, even with interest point matching for handling large gaps between sequential images. The holistic representation of DevianceNet extracted from entire image sequences makes it possible to accurately classify and detect deviant places.

We also compare the severity-aware loss with a cross-entropy loss. Table 4 shows the severity-aware loss generally outperforms the cross-entropy loss in DevianceNet. The regularization term H enforces the minimization of MAE because the predictions to adjacent classes are considered.

3.3 Transferability

Following (Suel et al. 2019), we evaluate how well DevianceNet trained on only one city data (Seoul, South Korea) predicts deviance for other city clips. We also fine-tune the pre-trained weight using small sets of deviant places for each city (5%, 10%, 30%, and all training data). As shown in Figure 5, the performance improvement plateaus when 5% data for each city are used for fine-tuning. We note that the prediction results for South Korea cities are relatively better than US cities because each country shares similar visual attributes which have an impact on the visual perception-based deviance prediction. Through this evaluation, it is noticeable that street images can potentially serve as low cost surveillance tools in data-poor geographies.

In addition, we conduct an experiment on Place Pulse 2.0 dataset (Dubey et al. 2016). We modify the dataset by augmenting single images into sequential frames based on

Figure 7: Application to developing countries. Since developing countries do not have publicly available incident record data in usual, we use violent crime articles and their street images.

(a) Without interest point matching.

(b) With interest point matching.

Figure 8: Visualization of attention maps. DevianceNet with interest point matching maintains a focus on discriminative visual attributes even with significant viewpoint changes in Google street view images.

their geo-tagged data. We then categorize the subjective perceived safety into 5 deviance classes. The result is reported in Figure 6. DevianceNet shows promising results on the perceived safety as well.

Lastly, to validate a generality of DevianceNet for developing countries which do not have publicly available incident record data, we collect sequential street images for developing countries, and infer deviant classes on them. Our DevianceNet predicts all places as deviance class 1 correctly, and check on the facts with news articles in Figure 7.

3.4 Ablation Study

We conduct an extensive ablation study to examine the effects of different components on DevianceNet performance. The results are summarized in Table 4.

Interest Point Matching. We compare DevianceNet with and without SuperPoint (Sinha et al. 2020) as an interest point matching. The interest point matching network extracts image features and performs a nearest neighbor matching among interest points of sequential images. As shown in Table 4, the interest point matching achieves performance improvement over the stacking of input images, which is commonly used for video understanding tasks.

To better understand its effectiveness, we visualize the attention maps for two variants (with and without interest point matching). As shown in Figure 8, DevianceNet with

Frames	SEA	DIA	MAE
4 frames (25 m)	40.82	81.03	1.06
8 frames (50 m)	43.82	87.13	1.05
12 frames (75 m)	44.19	86.56	0.98
16 frames (100 m)	**48.17**	**89.77**	**0.88**
20 frames (125 m)	42.88	85.83	0.96
24 frames (150 m)	45.26	86.03	1.00

Table 5: Ablation study on the number of input frames. We test the performance changes in accordance with the number of input images. When we use 16 frames as input, DevianceNet shows the best performances. Note that $(\cdot)$ indicates the coverage range of the input sequence.

interest point matching consistently focuses on discriminative parts such as the building and the skyscraper, even with the significant viewpoint changes.

(2+1)D Convolution. In Table 4, we compare the use of (2+1)D convolution (Tran et al. 2018) against 3D convolution in DevianceNet. Decomposing the 3D convolution layer leads to an additional nonlinear rectification between 2D spatial convolution and 1D temporal convolution, which enables representing more sophisticated functions with the same number of parameters as those of 3D convolution. The results verify that the (2+1)D block is better for learning spatio-temporal representation from our sequential images.

The Number of Input Frames. Lastly, we find the optimal number of frames required for deviance identification and severity estimation. Since two consecutive frames usually cover about 6 meters, we test DevianceNet with input sequential images consisting of from 4 frames (25 meters) to 24 frames (150 meters). As shown in Table 5, DevianceNet taking 16 frames (100 meters) achieves superior performance over those trained on shorter and longer frames.

There is a trade-off between the recognition performance and the computational complexity like (Tran et al. 2018). Although more frames provide more information for recognizing deviant places, it becomes difficult to learn context information of areas as the complexity increases, which causes the performance drop.

3.5 Analysis of Visual Attributes

We provide visual and statistical analyses to better understand what DevianceNet learns. As shown in Figure 1, it is difficult for humans to recognize which attributes are associated with deviance. This naturally raises an interesting question: what are the visual elements that affect deviance?

We first examine the statistical distributions of visually distinct elements by counting the number of objects with high attention values. To do this, we perform semantic segmentation (Chen et al. 2018) and infer attention maps using Grad-CAM (Selvaraju et al. 2017) for input images. Here, we normalize a scale of the attention maps by the number of pixels for each element (i.e., car, building and road, etc.) in the semantic segmentation result. We then count the number of elements with the highest attention value for each input image. As shown in Figure 9, we display visual elements

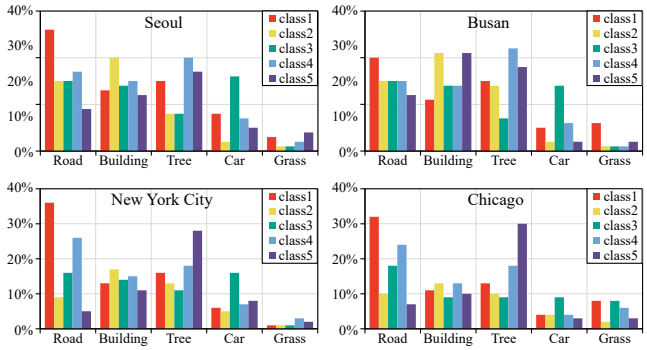
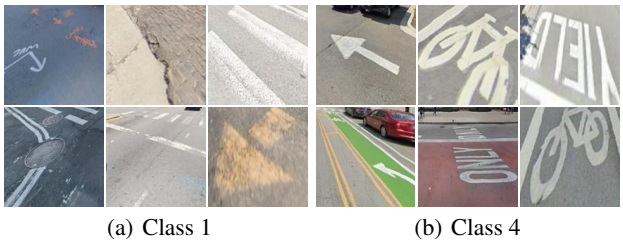


Figure 9: Distribution of visual elements. We report the distribution for two South Korea (Seoul and Busan) and two US cities (New York and Chicago).



(a) Class 1 (b) Class 4

(c) Class 2 (d) Class 4

Figure 10: Analysis on visual elements. Example visual elements in streets according to the deviance classes. The upper rows show road images of Chicago and lower rows are images of building in Seoul.

from four different cities (i.e., Seoul, Busan, Chicago and, New York) in both South Korea and US.

Since our dataset is obtained from Google street view, its elements consist mainly of roads, trees, and buildings, etc. Although their statistical distributions are similar across countries, it is interesting to see that the road is a common element in most cities with the highest attention value for deviance class 1, which is related to murders and sexual assaults. In Figure 10-(a) and (b), we can observe that there are illegal road markings and cracks in places for deviance class 1, compared to the places for deviance class 3. It is notable that as sociologist Kruger and his colleagues also demonstrate, road signs and damaged roads are associated with crimes in (Kruger and Landman 2008). Another interesting element is the building, which is known for its relation to crime occurrences (Cozens, McLeod, and Matthews 2018). As shown in Figure 10-(c) and (d), deviance classes vary depending on building styles. In particular, (Katyal 2001) suggests that the design of building has an effect on a crime pre-

Figure 11: Safety check with DevianceNet. We infer deviance class of street images around an abandoned building, called Ghost Tower, in Bangkok. Since the streets next to the dilapidated building are considered as dangerous place, it is predicted that there may be a deviation. Specifically, DevianceNet predicts clips corresponding to red route (i.e., shortest path) as class 1 and blue route (i.e., a detour path) as class 4.

vention. We observe that the exteriors of most of the buildings in places of deviance class 4 are made of glass blocks as well. We provide additional visual attributes analysis on the relationship between a variety of objects (e.g., road, building, tree, pavement and car) and deviance classes in our supplementary material.

Through the analysis of visual attributes, we note that there is a main difference between our work and broken windows theory-based approaches (Arietta et al. 2014; Naik et al. 2014; Porzi et al. 2015). Our DevianceNet selectively gives more weight to encoded visual attributes in surrounding areas, while the approaches focus on local disordered visual elements such as graffiti and dustbin.

4 Discussion

Applications in real-world scenarios. Our work supports policymakers in planning cities and individual users visiting unfamiliar areas. Urban safety plans can be established with factors that affect deviance through simple streetview images, rather than specific GPS-level crime records. Our research can be particularly useful in developing countries that rely on sociodemographic information covering too broad a range. We also expect that Crime Prevention Through Environmental Design (CPTED), one of the social science studies related to deviance occurrence, can also be replaced by our data-driven model which can transfer visual features of other cities with ease.

Additionally, individual users can identify potential risks of routes when visiting unfamiliar places. We show an application to safety way-guidance. As shown in Figure 11, a person would like to safely walk from the starting point to the destination. A path-finding like Google Maps directs

Prediction	Class 4	Class 5	Class 2
GT	Class 2	Class 3	Class 1

Figure 12: Failure cases. DevianceNet often fails the deviance prediction for uncommon places.

the shortest path. However, the route can be considered as a dangerous path when the high-level deviance is predicted. For this case, our DevianceNet is applicable for alternative path-finding to suggest a detour around the unsafe place.

Inspiration of network design. Deviance, including violent crimes, is an issue in our daily lives and many related theories are studied in social sciences (e.g., broken window theory, CPTED, and deivance theory). Among them, our network design is partially inspired by a symbolic interaction approach of the deviance theory that people learn deviance from their neighborhoods (Burgess and Akers 1966). In other words, individuals' deviant actions are associated with their surrounding environments (e.g., streets and actual cases of crime). Therefore, we design DevianceNet which learns holistic representations of streets from sequential images with its corresponding incident reports. However, other approaches (i.e. structural-functional and social-conflict) in deviance theory which highlight the relationship between deviance and social structure, are not covered in this work. We expect that the performance improvement can be achieved if the whole deviance theory is incorporated into a design of a CNN framework.

5 Conclusion

We have developed a CNN framework for the deviance prediction, whose design is inspired by the concept of deviance which includes formal and informal social norms.

We also collect a large-scale and geo-tagged sequential images of deviant places based on objective incident report data. Moreover, we have proposed the severity-aware loss based on Heinrich's law, which shows better performances in both deviance identification and severity estimation than the traditional cross-entropy loss.

However, there is still a limitation. It is sensitive to uncommon visual appearances in Google street view images such as dusk, tunnel and coastal roadways as displayed in Figure 12. An effective incorporation of domain adaptation (Wu et al. 2018; Vu et al. 2019) within DevianceNet is expected to minimize the gap of the visual appearances.

Lastly, our DevianceNet is designed as an assistive method that primarily identifies potential deviant places. In our dataset, one deviance class contains at least 11 incident types, which is comprehensive. In other words, its prediction results do not indicate specific crime occurrences. Therefore, we would like to highlight that DevianceNet can be used as a supporting tool to provide a primary guidance to policy makers and researchers.

Acknowledgements

We would like to thank the Police Science Institute of Korean National Police University for providing well-processed data. This work is in part supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST), No.2014-3-00077, AI National Strategy Project, No.2021-0-02068, Artificial Intelligence Innovation Hub), Vehicles AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea (NIPA) funded by the Ministry of Science and ICT(No. S1602-20-1001), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2020R1C1C1012635).

References

Alves, L. G.; Ribeiro, H. V.; and Rodrigues, F. A. 2018. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505: 435–443.

Andersson, V. O.; Birck, M. A.; and Araujo, R. M. 2017. Investigating crime rate prediction using street-level images and Siamese convolutional neural networks. In *Latin American Workshop on Computational Neuroscience*.

Arietta, S. M.; Efros, A. A.; Ramamoorthi, R.; and Agrawala, M. 2014. City forensics: Using visual elements to predict non-visual city attributes. *IEEE transactions on visualization and computer graphics (TVCG)*, 20(12): 2624–2633.

Burgess, R. L.; and Akers, R. L. 1966. A differential association-reinforcement theory of criminal behavior. *Social problems*, 14(2): 128–147.

Carine, J. Y.; and Park, T. 2019. Functional and Technical Methods of Information and Risk Communication. *Perspectives on Risk, Assessment and Management Paradigms*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Cozens, P.; McLeod, S.; and Matthews, J. 2018. Visual representations in crime prevention: exploring the use of building information modelling (BIM) to investigate burglary and crime prevention through environmental design (CPTED). *Crime Prevention and Community Safety*, 20(2): 63–83.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Diba, A.; Fayyaz, M.; Sharma, V.; Paluri, M.; Gall, J.; Stiefelhagen, R.; and Van Gool, L. 2020. Large Scale Holistic Video Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; and Hidalgo, C. A. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *Proceedings of the European conference on computer vision (ECCV)*.

Facil, J. M.; Olid, D.; Montesano, L.; and Civera, J. 2019. Condition-invariant multi-view place recognition. *arXiv preprint arXiv:1902.09516*.

Fu, K.; Chen, Z.; and Lu, C.-T. 2018. Streetnet: preference learning with convolutional neural network on urban crime perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*.

Gau, J. M.; and Pratt, T. C. 2010. Revisiting broken windows theory: Examining the sources of the discriminant validity of perceived disorder and crime. *Journal of criminal justice*, 38(4): 758–766.

Gorgul, E.; Chen, C.; Wu, K. K.; and Guo, Y. 2019. Measuring street enclosure and its influence to human physiology through wearable sensors. *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 65–68.

Hara, K.; Kataoka, H.; and Satoh, Y. 2017. Learning spatio-temporal features with 3D residual networks for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

Hayhurst, E. R. 1932. *Industrial accident prevention, a scientific approach*. American Public Health Association.

Huey, L. 2016. Harm-focused policing. *Journal of Community Safety and Well-Being*, 1(3): 83–85.

Katyal, N. K. 2001. Architecture as crime control. *Yale Lj*, 111: 1039.

Keizer, K.; Lindenberg, S.; and Steg, L. 2008. The spreading of disorder. *Science*, 322(5908): 1681–1685.

Kelling, G. L.; Wilson, J. Q.; et al. 1982. Broken windows. *Atlantic monthly*, 249(3): 29–38.

Koss, M. P.; Woodruff, W. J.; and Koss, P. G. 1990. Relation of criminal victimization to health perceptions among women medical patients. *Journal of Consulting and Clinical Psychology*, 58(2): 147.

Kruger, T.; and Landman, K. 2008. Crime and the physical environment in South Africa: Contextualizing international crime prevention experiences. *Built environment*, 34(1): 75–87.

Kwan, Y. K.; Ip, W. C.; and Kwan, P. 2000. A crime index with Thurstone's scaling of crime severity. *Journal of Criminal Justice*, 28(3): 237–244.

Maharana, A.; Nguyen, Q. C.; and Nsoesie, E. O. 2019. Quantifying the Impact of the Built Environment on Neighborhood Crime Rates. In *International Conference on Machine Learning Workshop on AI for Social Good*.

Naik, N.; Philipoom, J.; Raskar, R.; and Hidalgo, C. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Najjar, A.; Kaneko, S.; and Miyanaga, Y. 2018. Crime Mapping from Satellite Imagery via Deep Learning. *arXiv preprint arXiv:1812.06764*.

Özbil, A.; Yeşiltepe, D.; and Argin, G. 2015. Modeling walkability: The effects of street design, street-network configuration and land-use on pedestrian movement. *A— Z ITU Journal of the Faculty of Architecture*, 12: 189–207.

Porzi, L.; Rota Bulò, S.; Lepri, B.; and Ricci, E. 2015. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia, (ACM Multimedia)*.

Ratcliffe, J. H. 2015. Towards an index for harm-focused policing. *Policing: A journal of policy and practice*, 9(2): 164–182.

Salesses, P.; Schechtner, K.; Hidalgo, C. A.; et al. 2013. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLOS ONE*, 8(7): 1–12.

Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*.

Sinha, A.; Murez, Z.; Bartolozzi, J.; Badrinarayanan, V.; and Rabinovich, A. 2020. DELTAS: Depth Estimation by Learning Triangulation And densification of Sparse points. In *Proceedings of the European conference on computer vision (ECCV)*.

Suel, E.; Polak, J. W.; Bennett, J. E.; and Ezzati, M. 2019. Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific reports*, 9(1): 1–10.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Z.; Han, X.; Lin, Y.-L.; Uzunbas, M. G.; Goldstein, T.; Lim, S. N.; and Davis, L. S. 2018. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.