ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414461

SEMI-SUPERVISED FEATURE EMBEDDING FOR DATA SANITIZATION IN REAL-WORLD EVENTS

Bahram Lavi, José Nascimento, and Anderson Rocha

Institute of Computing, University of Campinas (Unicamp), Campinas, São Paulo, Brazil

ABSTRACT

With the rapid growth of data sharing through social media networks, determining relevant data items concerning a particular subject becomes paramount. We address the issue of establishing which images represent an event of interest through a semi-supervised learning technique. The method learns consistent and shared features related to an event (from a small set of examples) to propagate them to an unlabeled set. We investigate the behavior of five image feature representations considering low- and high-level features and their combinations. We evaluate the effectiveness of the feature embedding approach on five collected datasets from real-world events.

Index Terms— Image sanitization; semi-supervised learning; feature embedding; data relevance analysis, forensic application.

1. INTRODUCTION

Typically, when an event happens, a large amount of information will quickly appear online and in social media, showing different aspects and vantage points of such an event. However, separating what is relevant to understand the event is difficult, especially for visual information. A large proportion of the available data often contains irrelevant or redundant contents. This high volume of uncleaned data is difficult to manage and understand. In such situations, forensic investigators struggle to analyze and fact-check an event, especially during crisis and terrorist attacks.

Notably, the task of cleaning an image data pool, also referred to as *image data sanitization*, is an essential task to exclude irrelevant/undesired data from a dataset. Properly selecting relevant items can enable forensic investigators to understand an event and the scale of damage caused.

In the context of machine learning for determining data relevancy, a robust model usually requires a broad set of labeled examples for its training step. In this case, human experts are needed to manually label a large set of data samples to train a supervised machine-learning model; this is an expensive and time-consuming task, which we seek to avoid in this research.

Some studies have addressed the data sanitization issue through deep learning solutions. In particular, most recent works have focused on deep convolutional neural networks (DCNNs) and supervised learning for filtering irrelevant images collected through social media. Johnson et al. [1] relied upon five multi-layer perceptron classifiers, each of which specialized in a particular set of image features to identify relevant data items. In [2], the authors proposed a method for indicating the level of damage in a disaster by training a standard linear Support Vector Machine (SVM) classifier over feature maps obtained through convolutional networks. They further proposed a domain-adaptation technique [3] to learn data relevancy over a set of image feature representations. The method relied upon a pre-trained VGG16 network [4] as its backbone. A domain adaptation approach was also proposed in [5] along with an adversarial learning method to learn relevant features from a source domain and map them to a target domain. Damage severity assessment was considered in [6, 7, 8] in which DCNN models were used for capturing and filtering relevant images out of some data collections. Some events considered include earthquakes and floods, for which many images were available through social media. In some cases, the estimation of the level of damages was also an objective. As these kind of events are unfortunately common, many times images from one flood may appear in a dataset representing another.

Some methods have focused on identifying related information by including alternative domain sources (e.g., text-modality) for a target event. In [9], a combination of long-short-term-memory (LSTM) and CNNs was used to explore text and visual information when determining data relevancy. A modality-agnostic shared representation was proposed in [10] to learn a joint representation of text and images, simultaneously. Taking a different path, the authors in [11] investigated supervised and unsupervised learning techniques to reduce the training set size required for determining relevant items by selecting representative samples (data seeds) identified through clustering.

A detailed analysis of the prior art has shown one crucial limitation: the typical need for large volumes annotated examples for training. In other words, to learn representative examples of a flood taking place on a particular date and place, typically many examples are needed for training the method. In contrast, we investigate a robust solution that requires just a small set of training samples. This is a significant contribution, especially in forensics investigations for which annotated data is scarce.

We address the task of image data sanitization as a binary semi-supervised classification problem by embedding generated features from different sources. It is noted that we refer to *embedding* as the feature space (manifold) which is generated in an unsupervised manner (no training needed to obtain/generate them), and further is used in a semi-/supervised technique to discern its performance variation. Our goal is to learn consistent features present in a reduced set of examples and look for those features in a large (unlabeled) set. Data sanitization is of particular interest for social media data, typically characterized by strong imbalance (relevant items are just a fraction of the collected data items). Once we extract representative features for a small training set, we rely upon a semi-supervised learning (SSL) algorithm [12] to learn local and global consistency (LGC) among those features through a regularization framework. This SSL method is responsible for propagating the consistent features to unlabeled examples and, thus, determining relevant items.

This research was supported by São Paulo Research Foundation (FAPESP), under the thematic project "*DéjàVu: Feature-Space-Time Coherence from Heterogeneous Data for Media Integrity Analytics and Interpretation of Events*" with grant numbers {2017/12646-3, 2018/05668-3, and 2020/02241-9} and CNPq #304497/2018-5.

The features themselves play a pivotal role in reaching a robust performance, especially for semi-supervised methods. Finding a discriminant image feature representation that can describe local and global patterns of a given image is a hard task. In this regard, we investigate low- and high-level features that better capture the various aspects of a given event. We evaluate the proposed method considering different real-world events with data collected from various sources (e.g., Twitter, GooglePlus, Flickr, Youtube, among others).

We organize the remaining of this paper into four sections. Sec. 2 presents the proposed methodology while Sec. 3 shows the performed experiments. Finally, Sec. 4 concludes the paper.

2. PROPOSED METHOD FOR DATA SANITIZATION

In this section, we describe the methodology we propose for image data sanitization. It comprises two major elements: a richer image representation combining handcrafted and data-driven features and a semi-supervised learning method for propagating annotations (based on such features) to unlabelled examples. The task starts with a small set of labeled data samples for learning the initial classifier.

Image Characterization. In this task, we are interested in a discriminant descriptor that generates local and global information from a given image. Finding a universal descriptor is a hard task due to the scene's natural complexity and the spatial-temporal properties of events, which vary from an event to another.

Based on the nature of real-world events, we can observe that shape, pattern, and texture are the most apparent cues to describe an event semantically. Moreover, background-clutter, illumination changes, and pose variations are other factors that must be taken into account while seeking an image descriptor. Data-driven (high-level) features are now a staple to generate robust feature maps for different image tasks given the success of deep convolutional neural networks in various computer vision problems. However, although recently relegated as old-fashioned, handcrafted (low-level) features may allow great flexibility toward enabling a bird's eye view of an event in multiple scales. Given that an event is complex and spans different moments in time and space, we believe it is appropriate to explore a combination of data-driven and handcrafted methods to obtain both rich low-level and high-level features. Such combinations have been explored in the literature for different problems and had shown significant advances [13, 14, 15].

In this vein, we consider five image descriptors to generate image feature representations. Among them, we used three popular dCNN techniques to generate rich high-level features for a given image from an event: VGG16 [4], Inception-V4 [16], and Xception [17]. These are well-known CNN models for capturing shape, entangled patterns, texture, and color information. As we seek a model capable of analyzing different events, we do not fine-tune such networks for each specific event. We opt to use their models trained on the ImageNet [18] dataset and use them as feature extractors to obtain final feature maps.

Complementarily, we adopted two descriptors to encode detailed information related to color, local patterns, and texture: *gBiCov* [19] and *HOG* [20] features. The first descriptor, gBicov, was previously proposed for the challenging task of person re-identification. It relies upon biologically-inspired features (BIF) generated by Gabor filters with different scales over the HSV color channels. For a given image, it is subdivided into overlapping regions, and each region is represented by a covariance descriptor that encodes shape, location, and color information in multi-scales. In turn, the HOG feature can capture detailed information on different viewpoints. Table 1 presents the configurations adopted for each feature set.

Table 1. Feature dimensionality and the input image size for each group of image descriptors adopted in the investigations.

Level	Descriptor	Feature Dimensionality	Image input size
H	VGG16	4096	224×224
	InceptionV4	1536	299×299
	Xception	2048	299×299
L	gBiCov	1536	150×150
	HOG	648	150×150

Image Sanitization via Label Spreading. After representing all images with the feature sets described previously, we need to proceed with the spreading of available labels to unlabelled examples in order to properly filter what is relevant to understand an event of interest.

Consider we have a dataset X with a set of n data points $X = \{x_1, x_2, \ldots, x_h, \ldots, x_n\}$. Also, assume that we have a $n \times 2$ labeled matrix y, where the first h points are labeled as $y_i = 0$ if x_i is irrelevant to the event, and $y_i = 1$ and if x_i is relevant. The remaining $x_i(h + 1 \leq i \leq n)$ are unlabeled points and denoted as $y_{i \geq h+1} = -1$. Each data point, $x_i \in R^N$, represents a feature vector in N-dimensional space image.

Given an initial set of labeled examples (*seeds*), we adopt a learning graph-based semi-supervised method to propagate labels. The method relies upon local and global consistency (LGC) checks and aims to learn the function f from a small portion of labeled data and propagates labels on unlabeled data samples. In this work, we consider a binary classification problem to separate relevant data samples from irrelevant ones.

The algorithm learns a ranking function f, $f : X \to R^2$, and assigns a ranking value of f_i to each data point $x_i \in X$. The function is treated as a vector $f = [f_1, \ldots, f_n]^T.y$ on dataset X for classifying the corresponding class label of data point x_i as a label $y_i = \arg\max_j f_{ij}$.

A pairwise relationship on data points $x_{ij} \in X$ is defined as a graph $G = (V, E)$, where the nodes V is dataset X and the edges E is a weighted affinity matrix W constructed by the kNN kernel.

$$w_{ij} = \begin{cases} 1, & \text{if } x_j \in kNN(x_i, n\_neighbors) \\ 0, & \text{if } i = j \text{ or } x_j \notin kNN(x_i, n\_neighbors) \end{cases} \tag{1}$$

considering that $n\_neighbors$ is a hyper-parameter. A normalized Laplacian matrix is computed by the equation:

$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{2}$$

where D is diagonal matrix $D = \{d_1, \ldots, d_n\}$ and is computed as

$$d_i = max(\sum_j^n N_{nodes}(i, j), 1) \tag{3}$$

in which $N_{nodes}(i, j)$ returns 1 if i is a nearest neighbour from j, and 0 otherwise. Assuming that $f(0) = y$, the function f is iterated for a number of iterations until it converges, and can be defined as

$$f(k + 1) = \alpha S f(k) + (1 - \alpha) y \tag{4}$$

where α is a value in $(0, 1)$. The optimum ranking scores of function f are defined as a classifying function

$$f^* = \arg\max_j F \tag{5}$$

We refer the reader to [12] for a detailed explanation of the method.

Table 2. Datasets details.

Group	Event	Location	Year	Number of images			Source	Description
				Positive	*Negative*	*Unlabeled*		
A	Notre-Dame Cathedral	Paris, France	2019	1660	22023	0	Twitter (93.2% of the images) Flickr (6.8% of the images)	A remarkable incident, in which the historical Notre-Dame Cathedral in Paris was partially destroyed by a devastating fire. The fire broke out close to the Cathedral's spire and, as a result, the spire collapsed and most of the roof was destroyed. Since this event made the news all over the world, we collected a large amount of data through keywords related to the event.
	Grenfell Tower	London, UK	2017	14161	0	0	Forensic Architecture team [21]	An unprecedented fire that broke out in a 24-story residential building in London, injuring and killing several people. The catastrophe was captured and shared live by thousands of Londoners, with their cameras and smartphones, leading to a profusion of images and videos. This dataset contains roughly 150 videos of the event, from which we extracted the available frame.
	Marathon Bombing	Boston, US	2013	19092	0	0	YouTube video frames	A terrorist attack carried out using two pressure cooking bombs that caused explosions near the finishing line of the race. The attack killed three persons and injured several hundred others [22].
B	Bangladesh Fire	Dhaka, BD	2019	125	125	709	Twitter (96.0% of the images) Flickr (4.0% of the images)	A fast-moving fire in a densely populated district in Dhaka, Bangladesh. The fire started when a car's gas cylinder exploded after the collision of two cars. The fire spread over nearby buildings being used to store chemicals and also other ones. This tragic incident killed at least 80 people and injured dozens of others.
	National Museum	Rio de Janeiro, Brazil	2018	125	125	440	Twitter (82.5% of the images) Flickr (16.7% of the images) GooglePlus (0.8% of the images)	A devastating fire that destroyed the National Museum in Rio de Janeiro, Brazil. The fire started in an air conditioning equipment on the first floor of the building and quickly spread over the next three floors of the building uncontrollably. The entire building was damaged, and the ceiling collapsed. Almost the entire Museum's collection was destroyed.

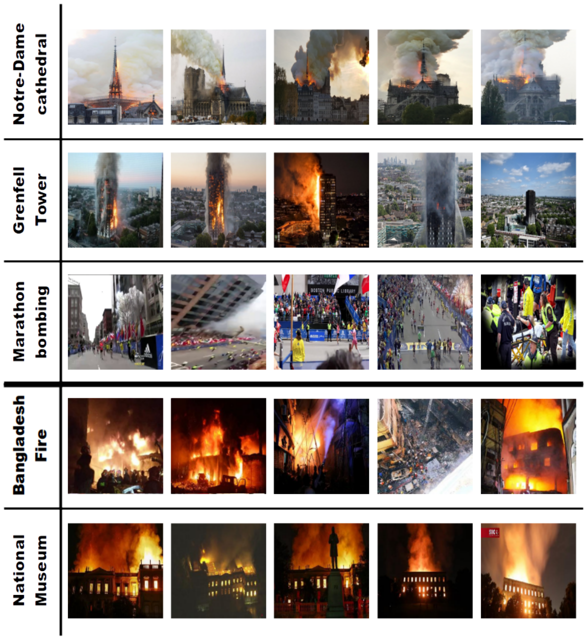

Fig. 1. Examples of the five events we consider in this work.

3. EXPERIMENTS AND ANALYSIS

In this section, we present the experimental settings to evaluate the proposed method for image data sanitization tasks. Our experiments were conducted using five image descriptors to generate image feature representations, and we evaluated the semi-supervised learning algorithm on our five datasets.

Datasets. Our primary objective is to understand events under a forensics vantage point. Therefore, we collect data from social media and the internet related to public events that caught the media attention over the past years. Table 2 shows detailed information for each dataset we adopt in this work. We grouped the datasets into two sets: group (A) comprises fully-labeled data samples, while group (B) comprises a small proportion of labeled data samples and many examples without annotations. While labeling relevant items for an event, we considered samples representing moments right before, during, and slightly after an event of interest to better capture its rich semantic components. Figure 1 depicts some images of each dataset. All datasets will be freely available upon the request.

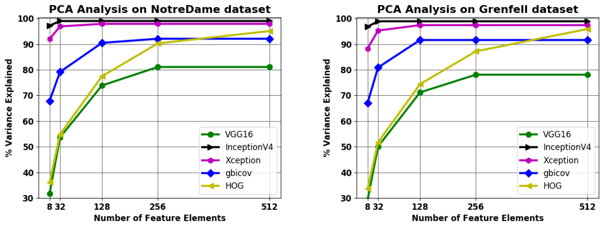

Fig. 2. The variance explained by PCA over feature representations on two target datasets. Most discriminative features are retained in the first 128 dimensions.

Dataset Split. The available datasets contain relevant examples (related to the event in question) and irrelevant ones. However, sometimes, only a few examples are annotated in each dataset. In the experiments, we also inject data samples collected from *Flickr* media source containing $35K$ image samples from 20 categories as junk to better evaluation the sanitization process. As we pointed out previously, it is undeniable that the broad set of collected media contents for the desired event includes irrelevant data and causes imbalance distribution. For each experiment, we selected β randomly labeled data samples from the target dataset, so that half of the samples are positive, and the other half are negative.

Implementation Details. We adopted a k Nearest Neighbours (kNN) [23] algorithm to construct an affinity graph. We chose the appropriate value of k for kNN through grid search. We achieved the best tradeoff between the speed of the algorithm and its performance accuracy when $k = 16$ neighbors. Furthermore, kNN performs well with a large number of observations in a low-dimensional feature space. The LGC algorithm was iterated up to 300 times to find the optimal solution for label spreading. To normalize the different feature representations, we adopted the principal component analysis method (PCA) as a pre-processing step. The number of dimensions of PCA was calculated based on the energy variance. We reduced the dimension of features obtained by each descriptor to 128 elements. Fig. 2 depicts the energy variance obtained by the Principal Component Analysis (PCA) application on each descriptor. The rationale for using PCA was that many descriptors carry information not necessary for the actual relevance classification. In addition, PCA also works as a proper normalization allowing the combination of different representations. As we can see, with 128 dimensions, we can already capture most of the variance of the signals.

Afterward, labels are propagated using the LGC learning algorithm previously described. We obtained the feature maps from the

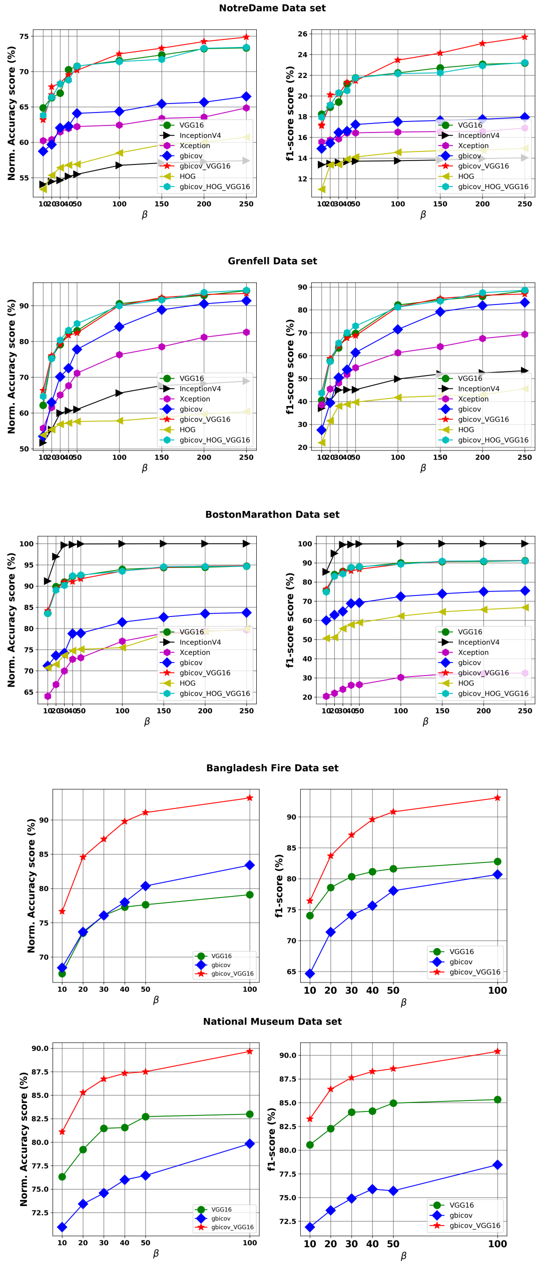

Fig. 3. Normalized accuracy and f1-score curves for different value of β on the target datasets. First three rows show the evaluation on the datasets in group A, while last two rows present group B datasets as so we considered three feature representations: gBicov, VGG16, and gBiCov_VGG16, the most accurate ones.

last layer for the adopted CNN models before the classification layer.

Feature Fusion. We empirically investigated a combination of low- and high-level features by applying a simple concatenation of one CNN as high-level features with the adopted low-level features. We compared the performance of the combined features with other single features. The feature fusion step takes place after the dimensionality reduction step.

Results. To assess how well we identify relevant images from an event, we consider two metrics: *normalized accuracy*, which reports the percentage of correctly classified data samples in a set, and *f1-score* (or F-measure) that represents the harmonic mean of relevant and irrelevant items relative to an event (in terms of precision and recall measures). For each classification score, we report the average value of over ten independent runs. It is noted that we only present the best performances achieved by each conducted feature. Fig. 3 shows the results for the target datasets in group A and B. The x-axis represents the number of labeled samples for different β values, and the y-axis is the corresponding classification score.

The combined *gBiCov_VGG16* features outperformed the classification accuracy when $\beta = 100$ for the *Notre-Dame Cathedral* case. The best performance achieved on *Grenfell Tower* case was with *VGG16* feature maps, and *gBiCov_VGG16* feature has 6% higher classification accuracy when $\beta = 10$. The image features obtained from *InceptionV4* yielded the best performances only in the case of *Marathon Bombing* dataset. Semantically, this dataset represents more color information (mostly the contents conveyed from the runners) and less content about the surrounding environment. It is evident from this dataset that the color information is more possible to be captured rather than other information such as local patterns and textures. However, this feature type showed poor performance on the remaining datasets.

From the overall experimental evaluation showed on the datasets in group A, the label spreading method showed optimum performance after increasing $\beta = 100$. Moreover, the feature representations of *gBiCov*, *VGG16*, and *gBiCov_VGG16* have more stable behavior than other features over the experiments on datasets in group A. Based on this observation, we only considered these three best methods for feature representation to evaluate the classification of the datasets in group B. The best performance was obtained with *gBiCov_VGG16* features.

4. CONCLUSION

Training a supervised learning method for image sanitization is daunting as each event might have its dynamics, space-time constraints, and, possibly, just a few labeled examples. It is wise not to expect that just one image descriptor will capture all the nuances of such an event. In this paper, we addressed both questions by proposing the combination of handcrafted and data-driven features that can be combined in a straightforward way to feed a semi-supervised method. Label spreading has shown to be adequate to this problem properly propagating the labels in five events of interest. Obtained results showed that considering a combination of VGG16 and gBiCov is the right choice. This combination achieved the best performance accuracy in a range between 65% and 95% over the adopted datasets.

Exploring semi-supervised algorithms hold promise for the applications that are highly expensive on annotating data process. In future work, we would like to explore a different set of graph-based semi-supervised techniques that fit with complex data structure. Another future work avenue is to explore self-supervised learning algorithms which can be trained jointly with multiple-feature instances to further enhance the performance of the image sanitization problem.

5. REFERENCES

[1] Matthew Johnson, Dhiraj Murthy, Brett Roberstson, Roth Smith, and Keri Stephens, "Disasternet: Evaluating the performance of transfer learning to classify hurricane-related images posted on twitter," in *Proceedings of the Hawaii Intl. Conf. on System Sciences*, 2020.

[2] Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra, "Damage assessment from social media imagery data during disasters," in *Proceedings of the 2017 IEEE/ACM Intl. Conf. on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 569–576.

[3] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises," *arXiv preprint arXiv:1704.02602*, 2017.

[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] Xukun Li, Doina Caragea, Cornelia Caragea, Muhammad Imran, and Ferda Ofli, "Identifying disaster damage images using a domain adaptation approach," in *Proceedings of the Intl. Conf. on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain. Academic Press*, 2019.

[6] Firoj Alam, Muhammad Imran, and Ferda Ofli, "Image4act: Online social media image processing for disaster response," in *Proceedings of IEEE/ACM Intl. Conf. on Advances in Social Networks Analysis and Mining*, 2017, pp. 601–604.

[7] Firoj Alam, Ferda Ofli, and Muhammad Imran, "Processing social media images by combining human and machine computing during crises," *Intl. Journal of Human–Computer Interaction*, vol. 34, no. 4, pp. 311–327, 2018.

[8] Ashish Kumar Layek, Amreeta Chatterjee, Debanjan Chatterjee, and Samit Biswas, "Detection and classification of earthquake images from online social media," in *Computational Intelligence in Pattern Recognition*, pp. 345–355. Springer, 2020.

[9] Abhinav Kumar, Jyoti Prakash Singh, Yogesh K Dwivedi, and Nripendra P Rana, "A deep multi-modal neural network for informative twitter content classification during emergencies," *Annals of Operations Research*, pp. 1–32, 2020.

[10] Ferda Ofli, Firoj Alam, and Muhammad Imran, "Analysis of social media data using multimodal deep learning for disaster response," *arXiv preprint arXiv:2004.11838*, 2020.

[11] Moacir A Ponti, Gabriel B Paranhos da Costa, Fernando P Santos, and Kaue U Silveira, "Supervised and unsupervised relevance sampling in handcrafted and deep learning features obtained from image collections," *Applied Soft Computing*, vol. 80, pp. 414–424, 2019.

[12] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems*, 2004, pp. 321–328.

[13] Tengfei Bao, Chenqin Fu, Tao Fang, and Hong Huo, "Ppcnet: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[14] P Jende, Z Hussnain, M Peter, S Oude Elberink, M Gerke, and G Vosselman, "Low-level tie feature extraction of mobile mapping data (mls/images) and aerial imagery.," *Intl. Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 40, 2016.

[15] Xianwang Wang, Tong Zhang, Daniel R Tretter, and Qian Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2035–2045, 2013.

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI Conf. on Artificial Intelligence*, 2017.

[17] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[19] Bingpeng Ma, Yu Su, and Frederic Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.

[20] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*. IEEE, 2005, vol. 1, pp. 886–893.

[21] Forensic Architecture team, "Grenfell media archive," http://www.grenfellmediaarchive.org, 2017.

[22] C. M. Rodrigues, L. Pereira, A. Rocha, and Z. Dias, "Image semantic representation for event understanding," in *IEEE Intl. Workshop on Information Forensics and Security (WIFS)*, 2019, pp. 1–6.

[23] Jun Wang, Shih-Fu Chang, Xiaobo Zhou, and Stephen TC Wong, "Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels," in *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.