Appeared in AIPR, 2017

TraffickCam: Crowdsourced and Computer Vision Based Approaches to Fighting Sex Trafficking

Abby Stylianou (a), Jessica Schreier (a), Richard Souvenir (b), Robert Pless (c)
(a) Washington University in St Louis, (b) Temple University, (c) George Washington University
corresponding author: abby@wustl.edu

Abstract

According to a 2016 study by researchers at the University of New Hampshire, over sixty percent of child sex trafficking survivors were at one point advertised online [13]. These advertisements often include photos of the victim posed provocatively in a hotel room. It is imperative that law enforcement be able to quickly identify where these photos were taken to determine where a trafficker moves their victims. In previous work, we proposed a system to crowdsource the collection of hotel room photos that could be searched using different local feature and image descriptors. In this work, we present the fully realized crowd-sourcing platform, called TraffickCam, report on its usage by the public, and present a production system for fast national search by image, based on features extracted from a neural network trained explicitly for this purpose.

I. Introduction

Victims of sex trafficking are often advertised online using provocative photos often taken in hotel rooms. In order to rescue victims and prosecute traffickerss, law enforcement seek to determine the hotel in the photos of victims. To date, law enforcement do this by performing time consuming manual investigations; for example, they ask individuals who are regular travelers if they recognize the location photographed, and compare the photos to those on travel websites, which can often be out of date or of only the nicest rooms at a hotel with professional photography.

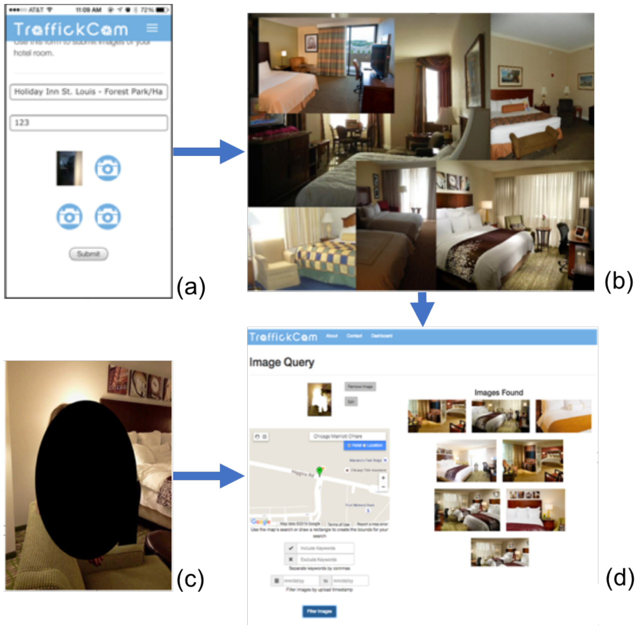
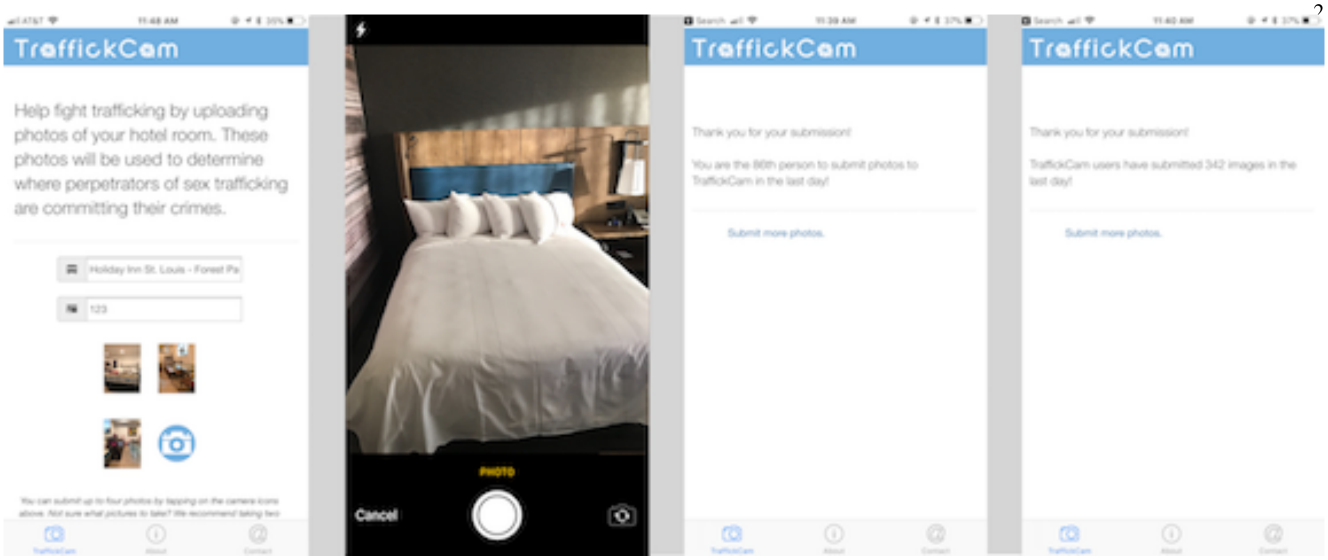


Fig. 1: (a) Travelers use a smart phone application called TraffickCam, which is available for iOS, Android and via any modern browser to anonymously submit photos of their hotel room. (b) Images, and extracted feature representations, are indexed and stored. (c) Law enforcement submit masked photos of victims of sex trafficking to the system. (d) Features from query images are matched with those in the database and results are provided to law enforcement.

In previous work, we proposed a system to crowdsource the collection of representative images of

978-1-5386-1235-4/17/31.00 ©2017*IEEE*



Help fight trafficking by uploading photos of your hotel room. These photos will be used to determine where perpetrators of sex trafficking are committing their crimes.

Holiday Inn St. Louis - Forest Pa

123

You can submit up to four photos by tapping on the camera icons above. Not sure what pictures to take? We recommend taking two

Thank you for your submission!

You are the 80th person to submit photos to TraffickCam in the last day!

Submit more photos.

Thank you for your submission!

TraffickCam users have submitted 342 images in the last day!

Submit more photos.

Fig. 2: Screenshots of the smart phone app, TraffickCam, that allows anyone to contribute to the database. The interface is designed to require minimal user time and to protect the user's identity, requiring the user to provide their hotel name from a list of nearby hotels, their room number, and four photos. The right two images show possible responses the TraffickCam system sends users to provide a sense of community engagement while maintaining the user's anonymity.

hotels around the country using a smart phone application, and presented baseline results on matching hotel room photos from publicly available images of hotel rooms [10]. In this work, we will present:

- the fully realized crowd-sourcing platform and report on its usage;
- a functional, national scale search platform for law enforcement; and
- state of the art hotel room image search results based on features extracted from a neural network trained on images from the TraffickCam system.

II. Background

There is relatively limited work that seeks to use machine learning tools or crowd-sourcing of data to fight sex trafficking. One line of work concentrates on indexing text content of online ads, easily extracted indices like phone numbers and metadata content like the target city of an advertisment to build models of relationships [11], [3] and creating law-enforcement facing search engines to support indexing into these graphs [7].

Other recent work explores approaches to evaluating online advertising to understand if the advertisements for sex or dating are likely to be related to trafficking. Some approaches that focus largely on the text of the advertisements [1] and others focus on multi-modal approaches integrating text and images [12] using a deep learning model. In both cases, there are substantial challenges to create ground-truth datasets because, and use law-enforcement input to assign labels for the training data.

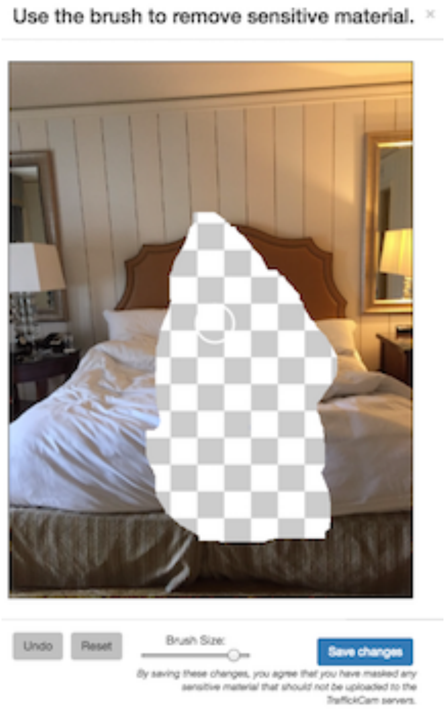
Our work concentrates on the different question of finding the hotel where a given picture was taken. This is a variant of the indoor place recognition problem [8], but most research in this field works to build visual indices at a small scale such as a factory, to help robots navigate in this environment. No work has been done that seeks to recognizes places at the same scale as this paper, trying to identify which of hundreds of thousands of possibly matching hotels is correct.

III. Platform & Dataset

Today, the TraffickCam smart phone app has been downloaded by over 100,000 users on both iOS and Android devices. The application, shown in Figure 2, simply requires users enter their hotel name, room number and take up to four images of their hotel room and bathroom. Users remain anonymous, submitting only their GPS location so that the system can verify that the photos were taken at the same location as the hotel (a protection against both malicious uploads of erroneous photos and accidental uploads due to the user selecting the incorrect hotel). Many crowdsourcing platforms help encourage the ongoing engagement of their community using points or sticker systems, or otherwise providing the user with information about how much they've contributed to the community. Because our users are entirely anonymous, however, we have no details about a user's submission history. Instead, we encourage ongoing engagement by providing the user with insight on how their most recent submission contributed to the dataset as a whole (e.g., the total number of photos including their most recent submission, the number of photos submitted so far on that day, or the number of photos previously submitted at that hotel). These types of feedback messages are seen in Figure 2.

Since TraffickCam was publicly released for iOS and Android in June 2016, TraffickCam users have uploaded over 188,000 images from nearly 27,000 hotels around the world. In addition to this ongoing collection of photos, the dataset also includes images from publicly available sources of hotel room photos, such as those available via the Expedia Affiliate Network (http://developer.ean.com/). As of October 2017, there are over 2.85 million images from over 254,000 hotels represented in the TraffickCam dataset.

Figure 4 demonstrates the importance of collecting images from the TraffickCam application. Images from publicly available sources such as Expedia are professionally photographed and showcase the nicest, staged rooms at a hotel. Images posted in ads for sex services are often taken with a smart phone by the victim themselves in less impressive hotel rooms. The TraffickCam images are often more representative of the types of photos seen in these ads. In addition, the TraffickCam photos provide an ever-growing archive of what the hotel looked like at a giving time, capturing renovations and changes that may not be present in the images on travel websites.



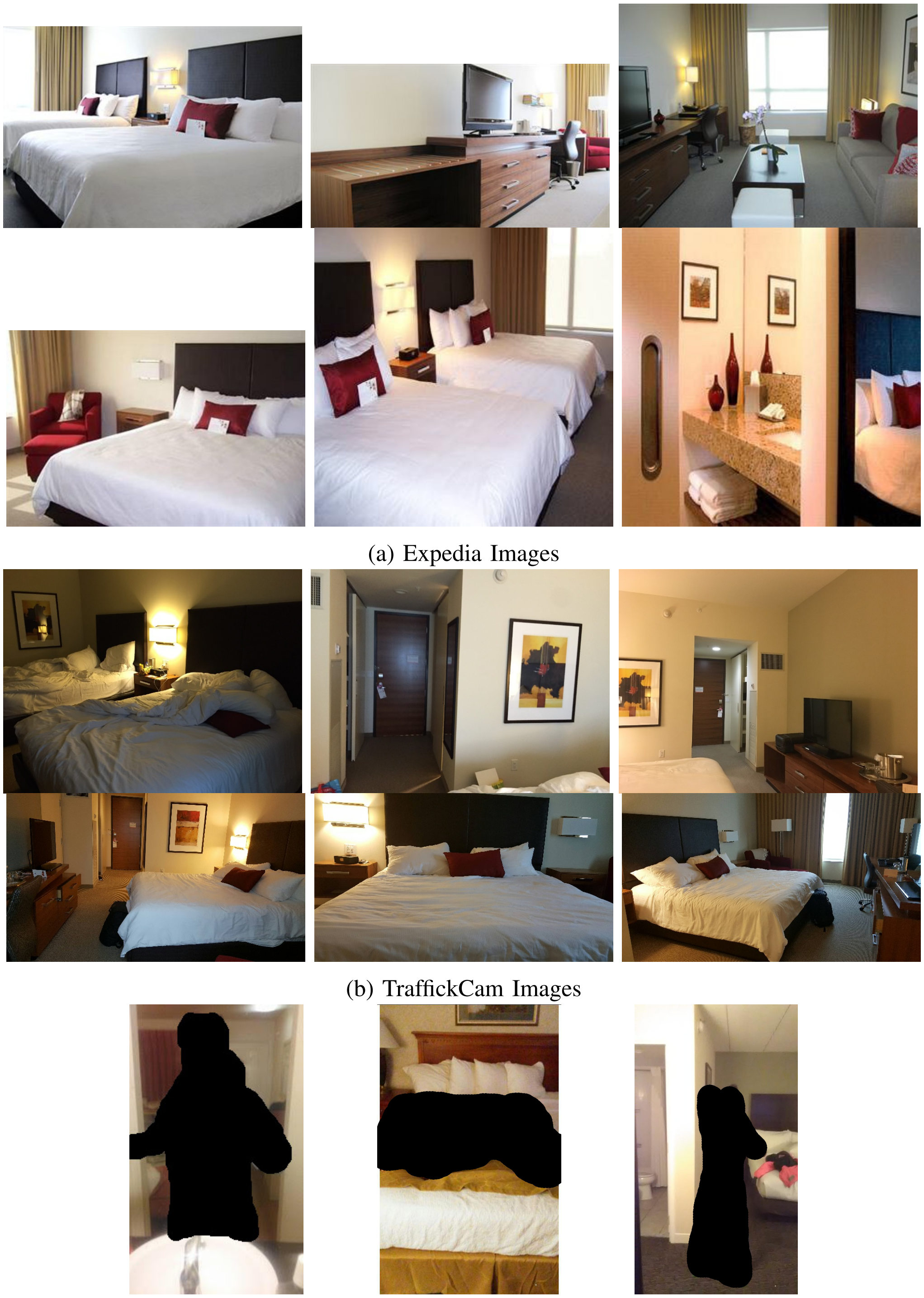
Use the brush to remove sensitive material.

By saving these changes, you agree that you have masked any sensitive material that should not be uploaded to the TraffickCam servers.

Fig. 3: Members of law enforcement mask off any sensitive content from their query images prior to the image being submitted to the server.

Law Enforcement Interface: Members of law enforcement who have been verified as working on sex trafficking cases can be granted access to the TraffickCam law enforcement portal. The law enforcement portal allows investigators to either browse all of the hotel room images that fit a text or geographic query, or to browse the images that are most similar to a query image. When an investigator provides a query image, they first mask off any sensitive regions of the image, as in Figure 3. This masking occurs before the image content ever leaves the investigator's computer; therefore, sensitive data is never transmitted or stored by the TraffickCam system. The masked image is then submitted to the TraffickCam server, where image features are extracted and compared to the database of TraffickCam images in order to provide the investigator with a list of similar images and hotels.

The TraffickCam search interface allows law enforcement to retrieve these search results for a masked query image at a national scale in a matter of seconds. This search is based on features learned from a neural network trained on the TraffickCam dataset, as described in Section IV, and the search



(a) Expedia Images

(b) TraffickCam Images

(c) Example Censored Query Images from Law Enforcement

Fig. 4: The top set of images are from Expedia and the middle set of images taken by TraffickCam users at the same hotel. The bottom set of images are censored versions of the types of images that might be provided by law enforcement. These examples demonstrate the discrepancy in the types of photos provided by Expedia, by the TraffickCam app and by law enforcement.

index is implemented using the Faiss library for efficient search [6].

IV. Methods

Our goal is to learn an embedding that differentiates between images from different hotels, given training data with less than 12 images per hotel on average. While this problem would often be posed as a fine-tuning problem, it is not clear that any existing pre-trained networks are the appropriate starting place for an indoor scene classification problem such as hotel matching. Instead, we train a model from scratch on a training set of 152,464 images from 13,246 different hotels, using the TensorFlow-slim implementation of the ResNet-50 [4] architecture. Our loss function is a variant of triplet loss described in [5].

In standard margin based triplet loss, a batch consists of $batch\ size/3$ anchor-positive example-negative example triplets, and the loss is formulated as:

$$mean(max(0, m + d_{a-p} - d_{a-n})) \tag{1}$$

where m is a margin, d_{a-p} is the distance in feature space between an anchor image and a positive example, and d_{a-n} is the distance between an anchor image and a negative example (so we want to learn an embedding such that the anchor-positive pair and the anchor-negative pair are at least m apart).

In this combinatorial variant of triplet loss, the margin based loss is the same, but batches are structured to include an equal number of examples from K different classes. For each class, there are $batch\ size/K$ possible anchor images. For each of those anchor images, there are $batch\ size/K - 1$ positive examples, and $(K - 1) \times batch\ size/K$ negative examples. This means that in a batch of size 120, for example, you see only 40 possible triplets per iteration in standard triplet loss, whereas in the combinatorial approach, if you have 10 classes represented equally in the batch, you see $(120/10) \times ((120/10) - 1) \times (120 - (120/10)) = 14256$ possible triplets per iteration.

Our training dataset is selected to include only hotel classes from the TraffickCam dataset that have at least four images, and at minimum two each from the TraffickCam smart phone application and from external sources. While we could significantly increase our training set size and the number of images that we see per hotel per iteration if we included locations with no TraffickCam smart phone images, we choose to only include hotels that have both TraffickCam smart phone images as well as images from external sources so that we learn to embed images from disparate sources to the same location. As a result of these dataset choices, we selected a batch size of 120, with 30 different hotel classes per batch and four images per class per batch, and report test results below after training for 85,000 iterations.

V. Experimental Design and Results

We compare the performance of our learned feature with representations learned from the ResNet-50 convolutional neural network (CNN) architecture [4], as implemented in TensorFlow-slim. We use publicly-available, pre-trained models, which we call *Places-365*, trained on the Places-365 Database [15], and *ILSVRC2012*, trained on a subset of the ImageNet dataset ([9],[2]).

We extract features from the 2048-dimensional global average pooling (GAP) layer prior to the final, fully connected layer as implemented in [14] (we get comparable performance with features extracted from the fully connected layer, but using the GAP layer provides more insight into feature localization).

Results: We evaluate the accuracy of the different feature types on a test set of 10,000 hotel images from 320 randomly selected hotel classes, where no images from any of the hotels were seen during training. We only use images from the TraffickCam smart phone application as query images, matching into a database of images from both TraffickCam and publicly available travel photos. This is because the images uploaded from the TraffickCam smart phone application resemble the types of query images law enforcement would use more closely than publicly available images from travel websites, as shown in Figure 4.

The Top-K accuracy reported in Figure 5 states for each feature type (ours, ILSVRC2012 and Places-365) whether the query image matched to another image from the same hotel in the top K results. The features trained on images from the TraffickCam dataset perform the best with a top-1 accuracy of

19%, top-10 accuracy of 48%, and top-100 accuracy of 80%, compared to (14%, 31%, 68%), (3%, 13%, 49%) and (0.2%, 4%, 27%) for ILSVRC, Places and random chance respectively.

Example query images and their top results for each of the different feature types are shown in Figure 6. The features that we have learned appear to better encode local structural information (e.g., a particular headboard or carpet pattern) and color information with some lighting invariance, while the ILSVRC2012 and Places-365 features maintain more global structural information (e.g., "a photo frame over the center of a queen sized bed on the left side of the room") and are more sensitive to strong lighting cues (e.g., a completely saturated window or strong lighting pattern projected by a lamp onto a wall). The ILSVRC2012 and Places-365 features also often appear to focus on "clutter" seen in the TraffickCam images, while our network has learned that that is not an important differentiating feature. Given the improved accuracy that we observe, the features that we have learned generalize well to photos of other rooms in the same hotel taken in different lighting conditions. This is important when extending these features to images not just of hotel rooms from TraffickCam or travel websites, but rather to images of victims of sex trafficking posed in hotel rooms.

VI. Discussion and Future Work

Working to make a practical tool in this problem domain requires some additional work beyond the classic machine learning problem discussed so far. We have discovered that it is very important that our training data include images from both TraffickCam and Expedia. Figure 4 shows example images from both domains and illustrates that the Expedia images have both better and more consistent lighting that the Traffickcamn images. Better and more consistent lighting is bad in our case because the Deep Learning approach needs examples of relevant lighting, viewpoint, and image quality variation in order to learn to generalize across those variations.

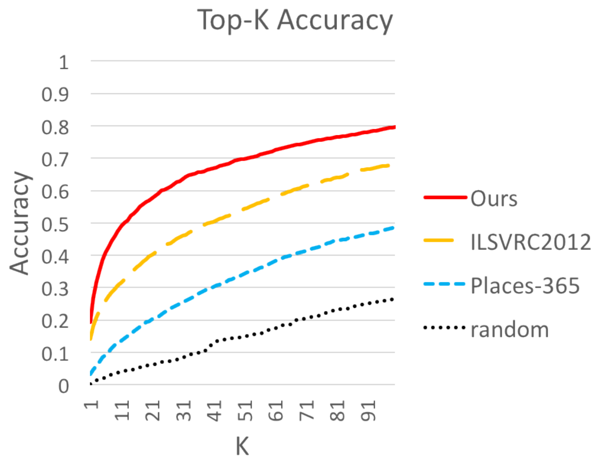


Fig. 5: We compare search accuracy for features extracted from our network and features extracted from ResNet-50 networks pre-trained on the ILSVRC2012 and Places-365 datasets. Top-K accuracy easures whether there are any instances of the correct answer between 1 and K. The best performance is achieved by our model trained from scratch using a variant of triplet loss described in Sec IV.

In future work we plan to address additional variations that appear in query images that are not common in the database of hotel room images. Figure 3 shows two of these features, the images used to query the database are often masked and are (from our experience) more likely to have the camera rotated slightly so the walls are not vertical. Additionally, when the query images are advertising images online, often information such as names and phone numbers are integrated into the image.

These issues can be addressed with a combination of training data augmentation approaches (our training images can be rotated, cropped, and have regions masked off) to create better approximations of the query images with known correct answers.

We expect additional testing with law enforcement to highlight additional issues related both to technical issues of our search and interface issues with how the results are presented.

References

[1] H. Alvari, P. Shakarian, and J. K. Snyder. Semi-supervised learning for detecting human trafficking. *Security Informatics*, 6(1):1, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[3] A. Dubrawski, K. Miller, M. Barnes, B. Boecking, and

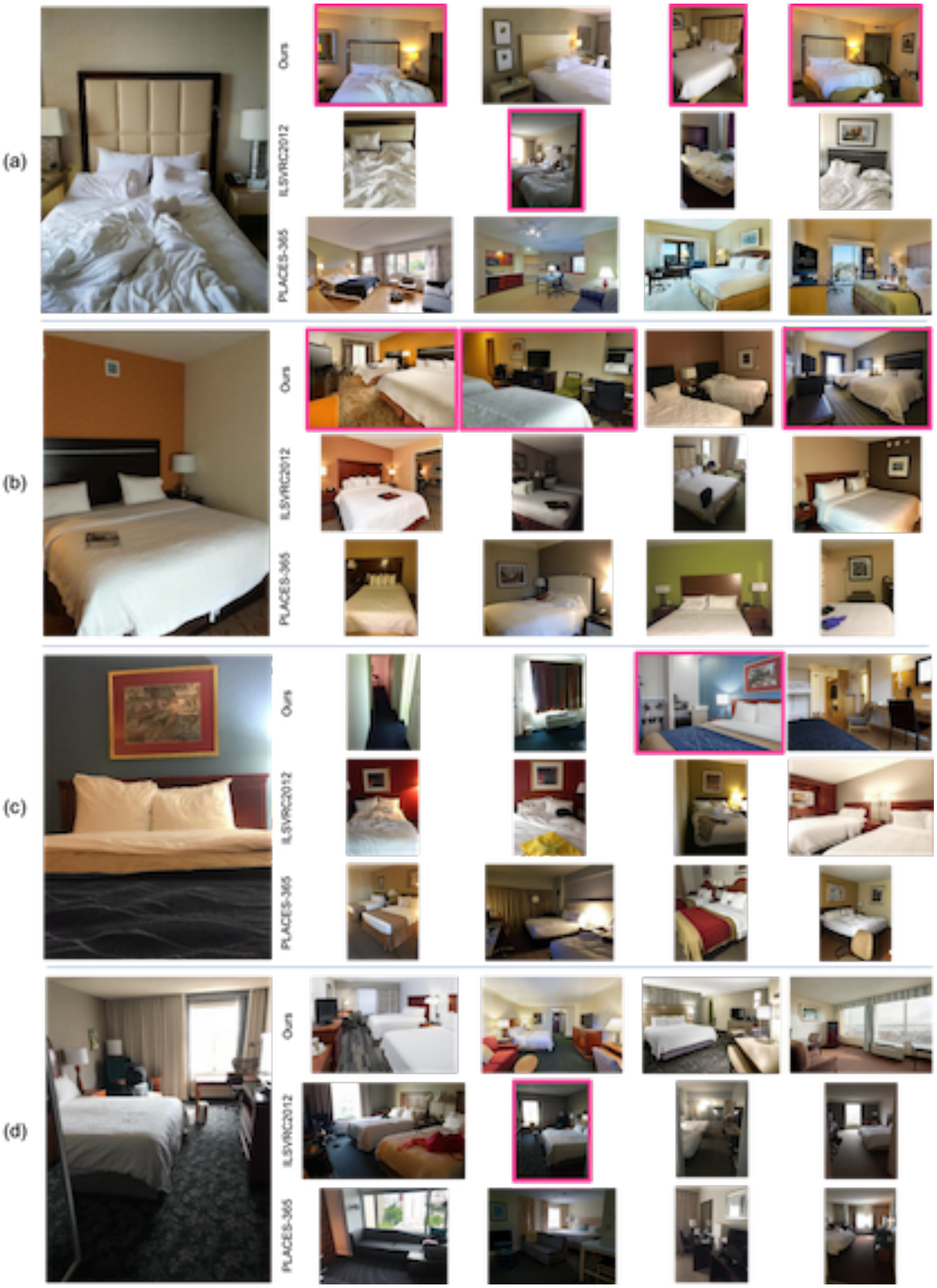


Fig. 6: Four example query images and the top four results using our features (top), ILSVRC-2012 features (middle) and Places-365 features (bottom). Our features appear to better encode local structural information (e.g., headboard shape, as seen in (a)) and color information with some lighting invariance (as seen in (b)), while the other features maintain more global structural information (e.g., ”a photo frame over a bed”, as seen in (c)) and are more sensitive to strong lighting cues (e.g., a completely saturated window, as seen in (d)).

E. Kennedy. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1(1):65–85, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[5] A. Hermans*, L. Beyer*, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017.

[6] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[7] M. Kejriwal and P. Szekely. An investigative search engine for the human trafficking domain. In *International Semantic Web Conference*, pages 247–262. Springer, 2017.

[8] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen. A realistic benchmark for visual indoor place recognition. *Robot. Auton. Syst.*, 58(1):81–96, Jan. 2010.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[10] A. Stylianou, A. Norling-Ruggles, R. Souvenir, R. Pless, undefined, undefined, undefined, and undefined. Indexing open imagery to create tools to fight sex trafficking. *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 00:1–6, 2015.

[11] P. Szekely, C. A. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, K. Knight, et al. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference*, pages 205–221. Springer, 2015.

[12] E. Tong, A. Zadeh, C. Jones, and L.-P. Morency. Combating human trafficking with deep multimodal models. *arXiv preprint arXiv:1705.02735*, 2017.

[13] J. Wolak and D. Finkelhor. Sectortion: Findings from a Survey of 1,631 Victims. Technical report, University of New Hampshire, 06 2016.

[14] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

[15] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.