Supplementary Materials: A Semantically Impactful Image Manipulation Dataset: Characterizing Image Manipulations using Semantic Significance

Anonymous WACV Algorithms Track submission

Paper ID 1845

A. Implementation Details

List of negative prompts used: lowres, text, error, cropped, worst quality, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, out of frame, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck, username, watermark, signature.

B. Observations and Discussions

B.1. Semantically significant manipulation ranking

From a qualitative standpoint, manipulations that are more semantically significant often cover larger areas of the image compared to minor ones. For example, a burning building occupies much more space than a swapped sign or t-shirt. This size difference likely influences the models, which tend to associate larger modifications with lower similarity scores and higher semantic significance scores.

Another factor affecting the results is the variability and accuracy of the generated textual descriptions. Machine learning models can sometimes produce incorrect or inconsistent descriptions, even for the same image scene. These discrepancies can impact the performance of semantic similarity models that rely on textual inputs. Figure A1 shows an example where the description accurately captures manipulations like a burning building or a flooded street. However, some errors remain, such as incorrectly stating that cars are driving through a flood, when in reality, they are in a dry area of the image.

Figure A2 highlights instances where the generated textual descriptions fail to capture manipulations like swapped signs or broken windows. A qualitative review suggests two key reasons for the discrepancies between successful and failed cases. First, the size of the manipulation is crucial: failed cases typically involve smaller manipulations, while successful ones feature larger changes. Second, the manipulation's impact on the image's overall atmosphere seems to play a role. In successful cases, the manipulations significantly alter the mood or context of the scene, whereas trivial changes in failed cases do not affect the scene's portrayal enough to be included in the description. This indicates that descriptions prioritize manipulations that meaningfully influence the image, often omitting minor alterations.

B.2. Image-level manipulation localization

We also generated quantitative results for detecting image manipulation using image-level classification. This approach identifies whether an image has been manipulated by detecting the presence of manipulated regions in its binary localization mask. Images with such regions are classified as manipulated, while those without are considered pristine. The goal was to evaluate the performance of mainstream methods under less stringent conditions compared to pixel-level analysis.

The results show that image-level classification significantly outperforms pixel-level results. This suggests that conventional methods are able to find manipulation artifacts. Although a significantly lower pixel wise performance suggests that the detected manipulated regions may be incorrectly identified. As a result, this leads to a higher accuracy in image-level classification but a low pixel-based F1 score. Figure A4 provides a qualitative example of instances where mis-detection and over-detection occur.

C. Evaluating image manipulation localization based on the manipulation size and type

To better understand the image manipulation localization results and the newly proposed semantic significance manipulation ranking task, we analyzed the evaluation outcomes based on key factors. Manipulation size is crucial to the success of many mainstream methods, as smaller manipulations are often harder to detect and localize. In our analysis, we examine how the size of manipulations af-

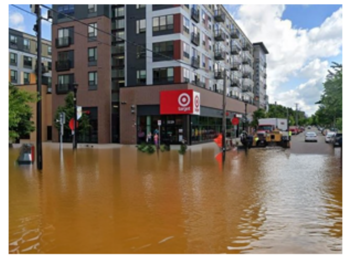

Figure A1. Qualitative example of textual descriptions generated by ChatGPT. Text highlighted in green is text correctly describing the manipulation. Red text is incorrect statements.

Figure A2. Qualitative example of textual descriptions generated by ChatGPT. The textual descriptions do not capture the corresponding manipulations (Swap Sign, Broken Window) for the two images.

Manipulation Localization Methods	**μ Per.**	**μ Recall**	**F1**
ManTraNet [9]	0.50	0.99	0.67
CAT-Net v2 [5]	0.61	0.92	0.74
TruFor [2]	0.75	0.79	0.77
CRCNN [10]	0.49	0.41	0.45
HiFi-Net [3]	0.68	0.54	0.60
IF-OSN [8]	0.72	0.16	0.27
ObjectFormer [7]	0.39	0.17	0.23
PSCC-Net [6]	0.99	0.28	0.44
RRU-Net [1]	0.50	0.99	0.67
SPAN [4]	0.52	0.04	0.08

Table A1. Image level precision, recall, and F1 scores using both manipulated and pristine images from the gold standard set. A fixed threshold of 0.5 is used when converting to a binary mask.

fects the performance of both mainstream and state-of-the-art (SoTA) detection and localization methods. To support this, we included manipulations of varying sizes, with their distribution shown in Figure A3.

Quantitative results show that most SoTA methods struggle to detect manipulations produced by generative stable diffusion models. As a result, we conducted a qualitative evaluation of localization masks from methods that performed well, specifically Cat-Net [5] and TruFor [2]. Observations indicate that small manipulations involving multiple objects often face mislocalization issues, as current methods tend to focus on identifying a single manipulated region in the scene.

Additionally, there are cases where localization models fail to detect any alterations in images with multiple manipulations, regardless of manipulation size or type. Preliminary analysis suggests this issue may be related to the semantic context or metadata of the original image.

Our qualitative analysis reveals two primary reasons

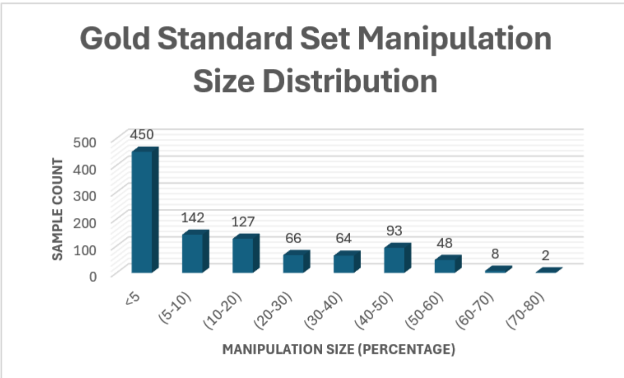

Figure A3. Distribution of manipulated images based on manipulation size

why SoTA methods struggle to localize manipulations generated by generative stable diffusion models: (1) misdetection/over-detection of manipulation regions and (2) complete failure to detect manipulations.

Examples of these localization behaviors are shown in Figure A4, which features a military tank inserted onto a street. Methods like Cat-Net [5] and TruFor [2] successfully localized the manipulated region, while methods such as IF-OSN [8] and RRU-Net [1] mis-detected the manipulated region. Similarly, PSCC-Net [6] and Mantra-Net [9] demonstrated over-detection.

The second common issue involves methods that failed to detect any manipulations at all. Examples include CR-CNN, SPAN, ObjectFormer, and Hifi-Net.

Generative stable diffusion details: We set the sampling size and guidance scale hyperparameters to 20 and 7.5, respectively. The generative stable diffusion model is then tasked with producing the specified manipulation, which is seamlessly integrated into the original image. Figure 2 in the main paper illustrates this process, where a building of interest is selected, and the manipulation simulates a fire, realistically portraying the house engulfed in flames. This example demonstrates our methodology's ability to create realistic and contextually relevant image manipulations.

D. Dataset Visual Examples

Figures A5 to A14 show visual manipulated image samples from the proposed CSI-IMD gold standard set.

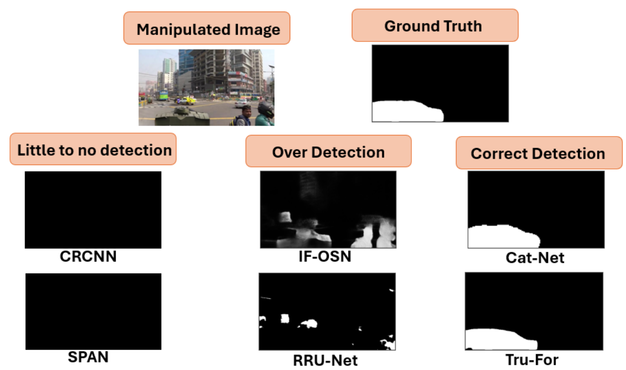

Figure A4. Qualitative example of state of the art method's prediction localization masks. Examples of no detection, over detection, and correct detection's are shown.

WACV #1845

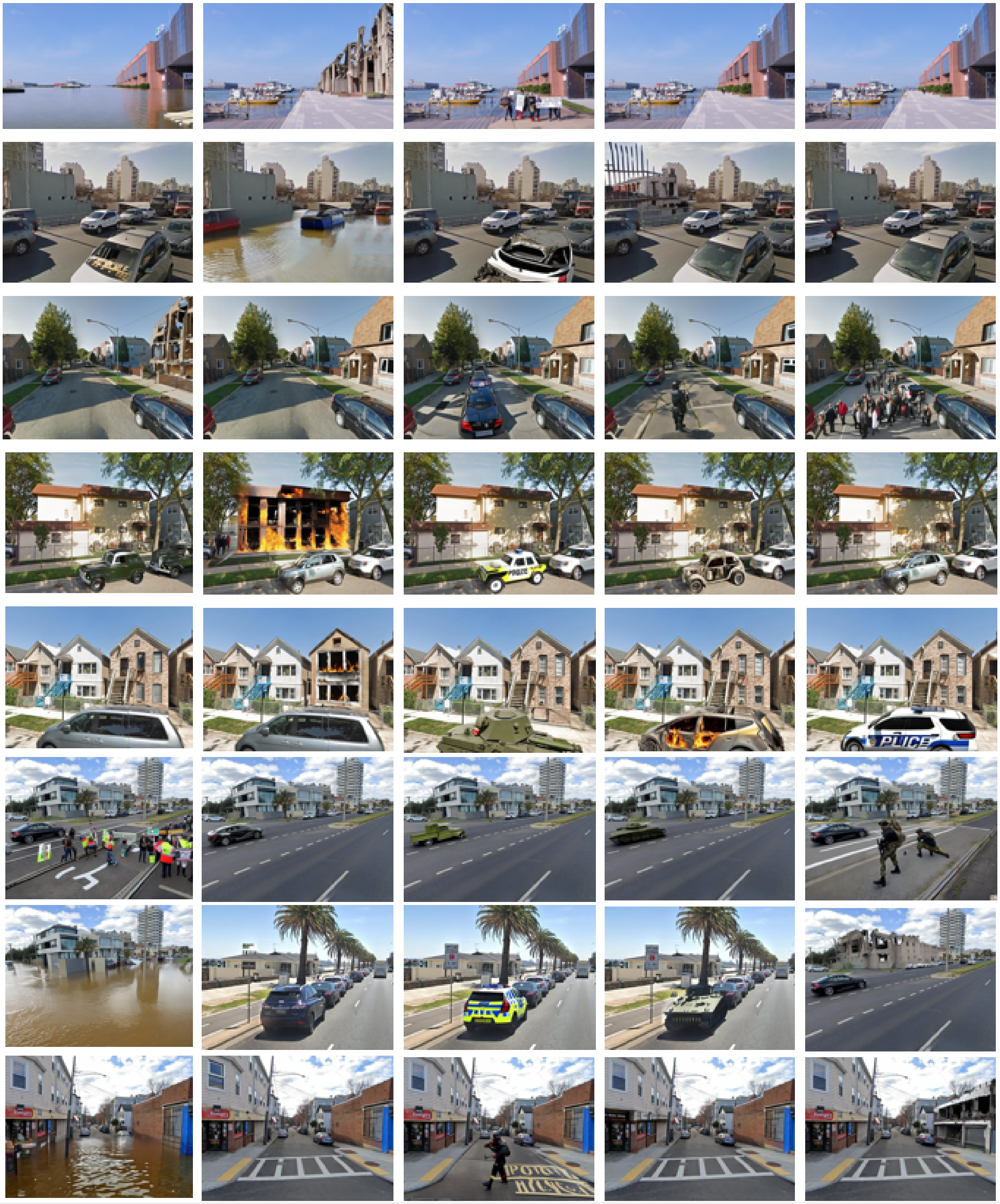

Figure A5. Examples of the manipulated images from our CSI-IMD gold standard set (1 of 10).

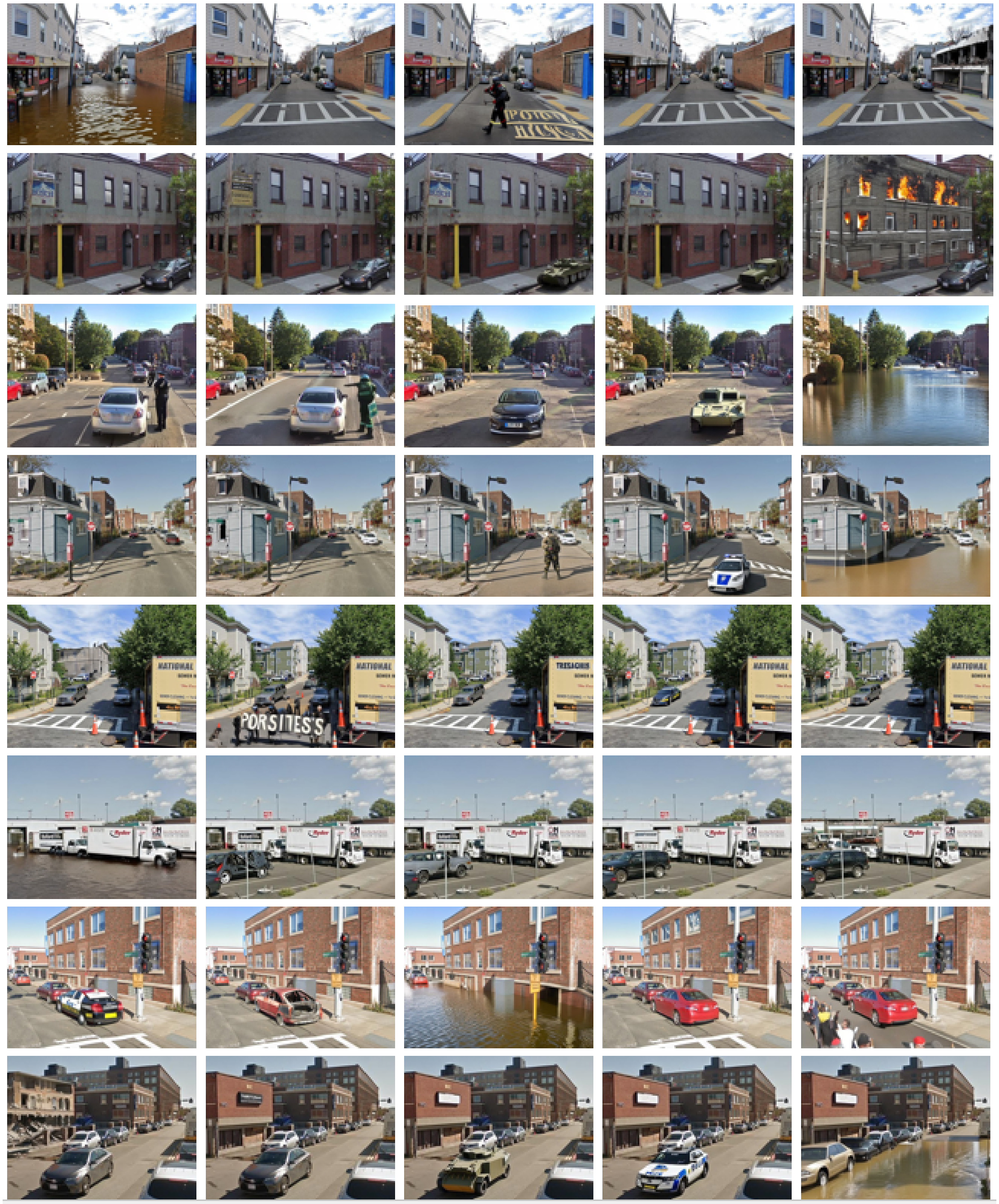

Figure A6. Examples of the manipulated images from our CSI-IMD gold standard set (2 of 10).

Figure A7. Examples of the manipulated images from our CSI-IMD gold standard set (3 of 10).

Figure A8. Examples of the manipulated images from our CSI-IMD gold standard set (4 of 10).

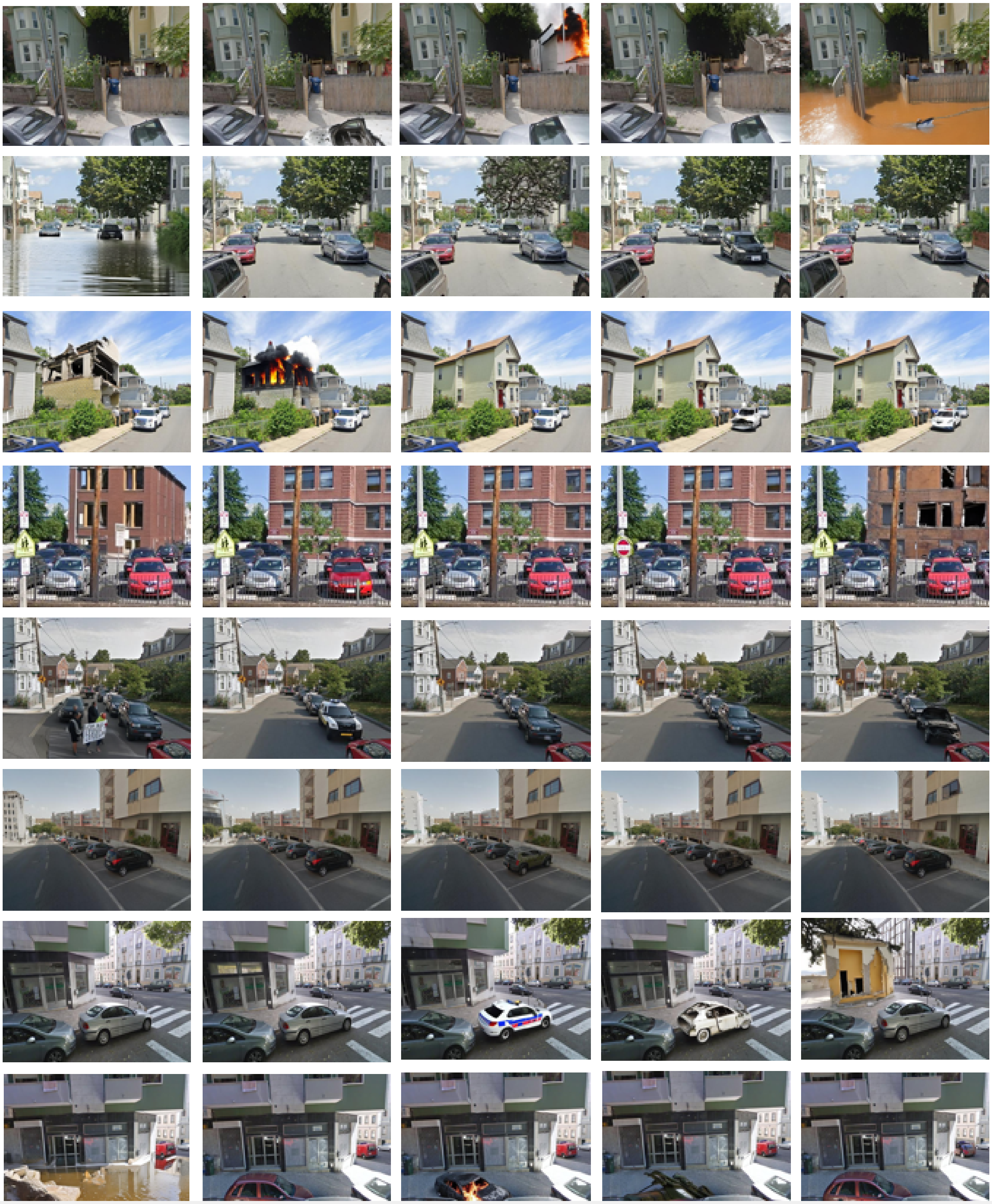

Figure A9. Examples of the manipulated images from our CSI-IMD gold standard set (5 of 10).

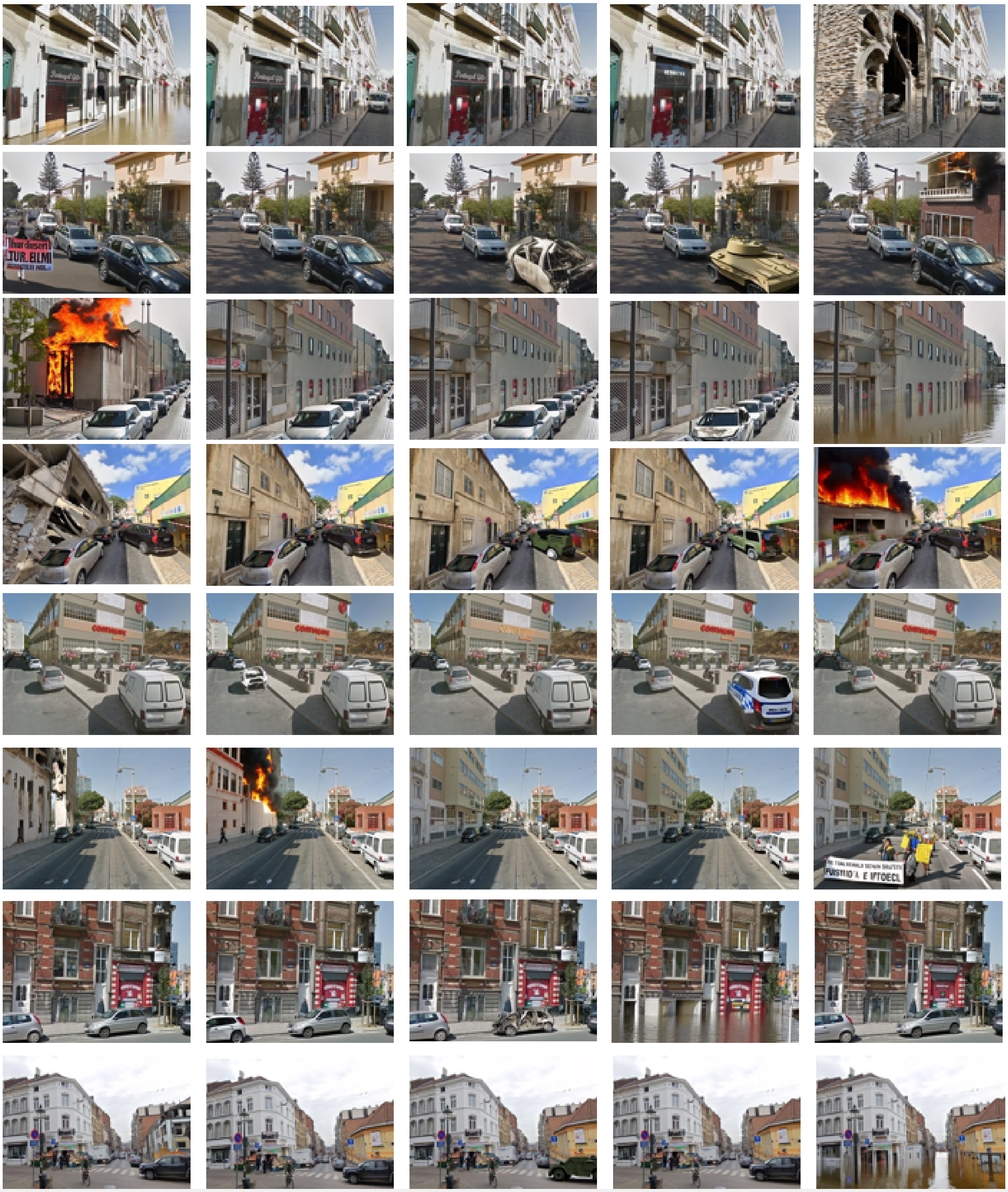

Figure A10. Examples of the manipulated images from our CSI-IMD gold standard set (6 of 10).

WACV
#1845

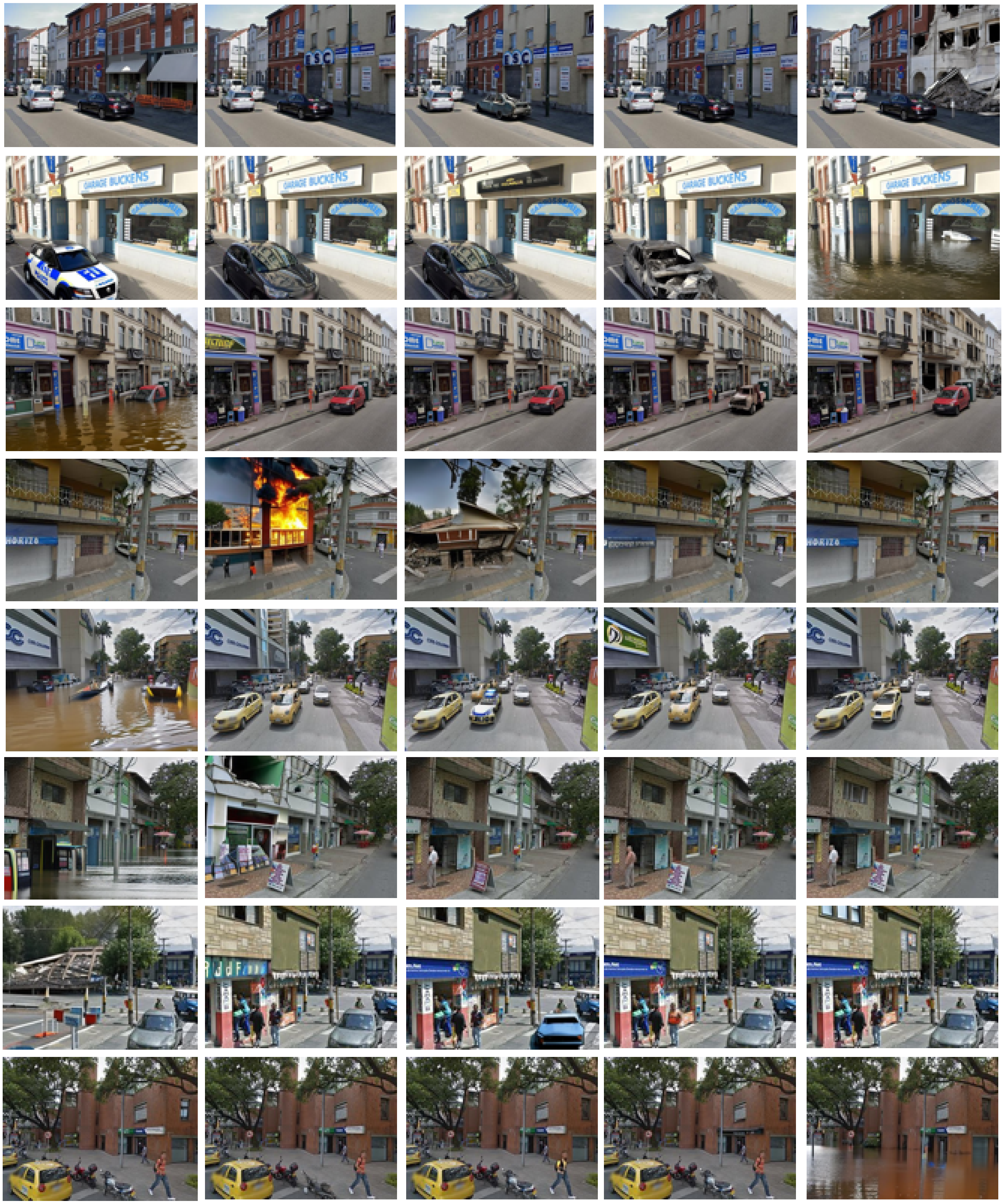

Figure A11. Examples of the manipulated images from our CSI-IMD gold standard set (7 of 10).

WACV
#1845

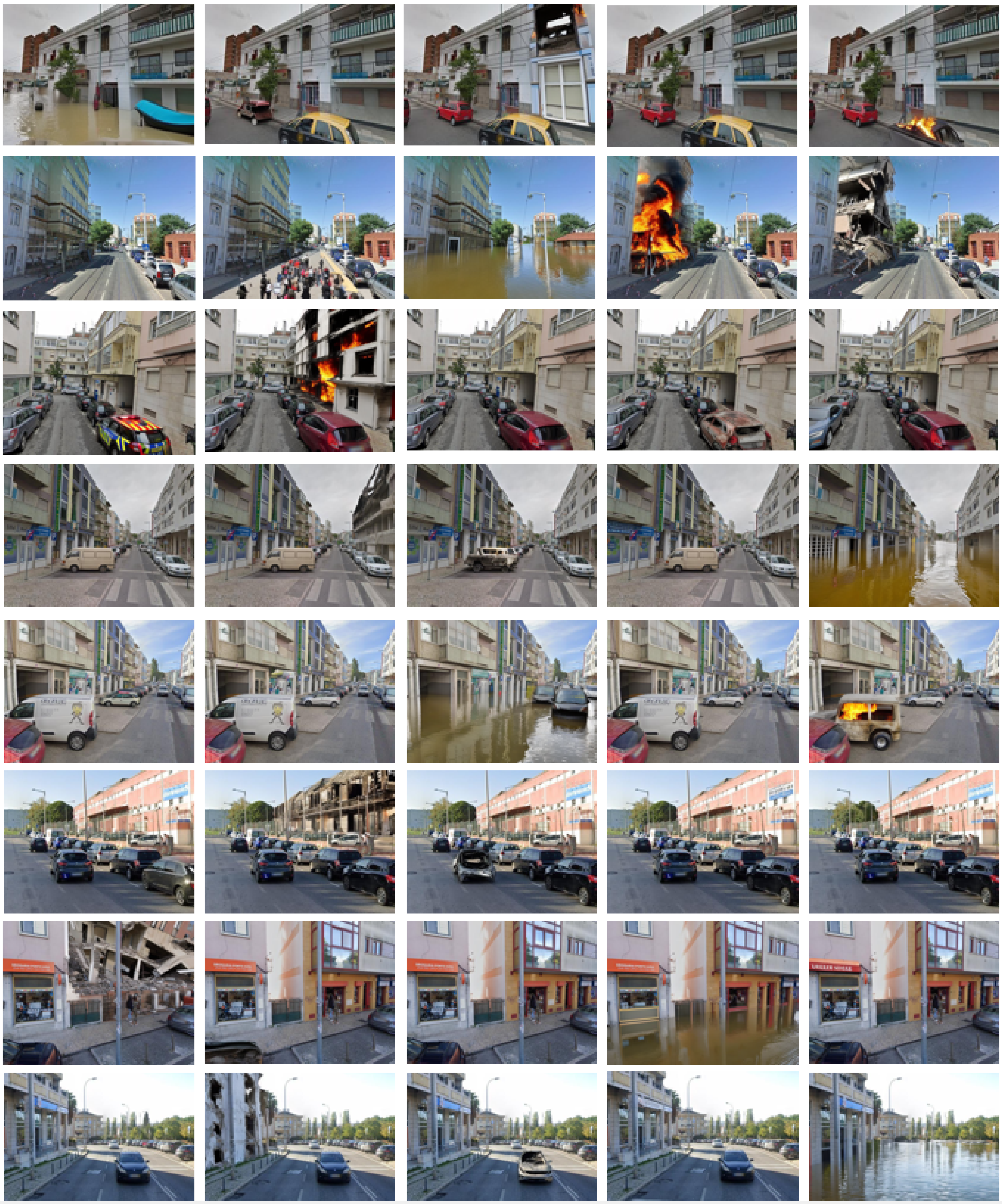

Figure A12. Examples of the manipulated images from our CSI-IMD gold standard set (8 of 10).

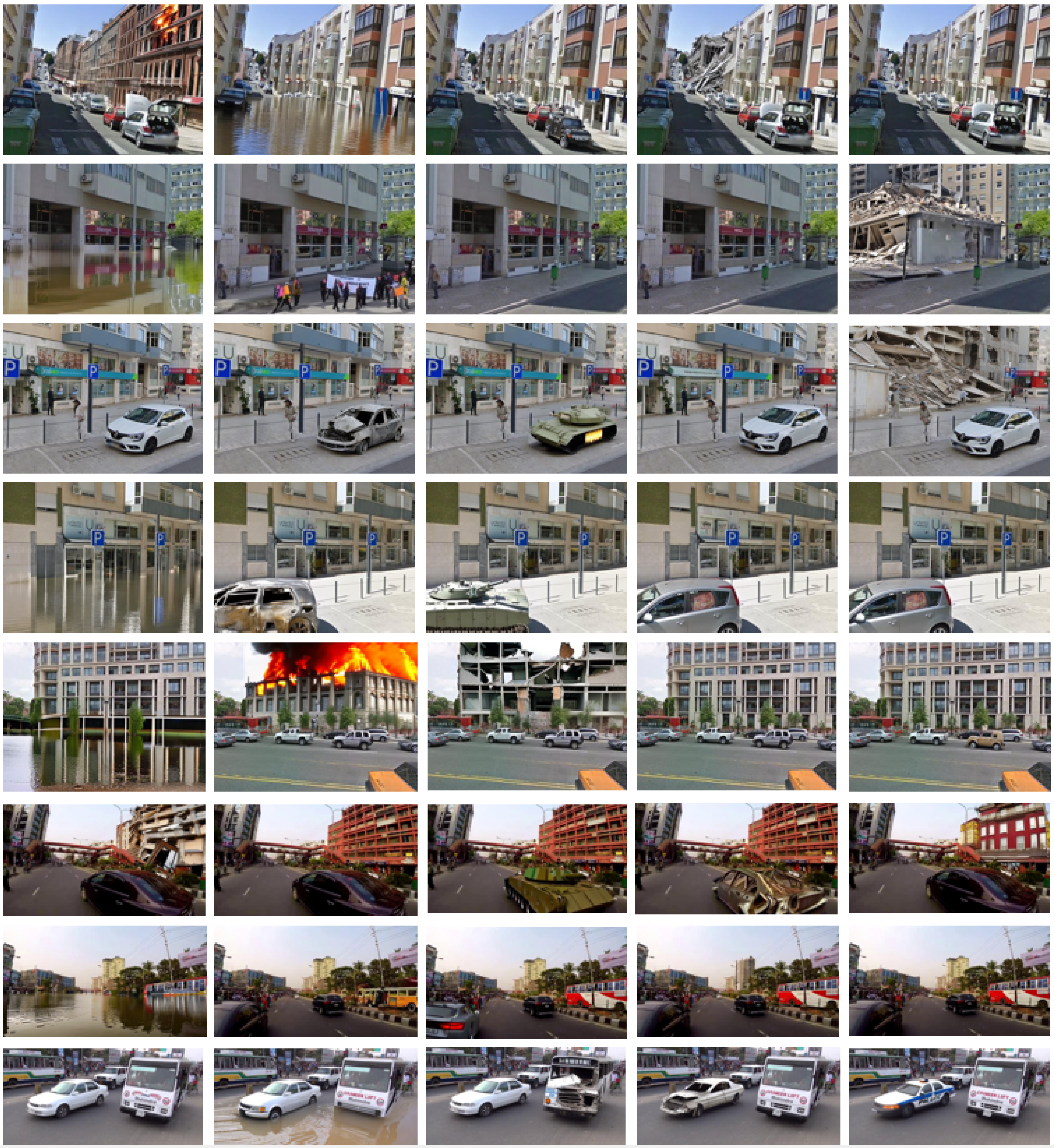

Figure A13. Examples of the manipulated images from our CSI-IMD gold standard set (9 of 10).

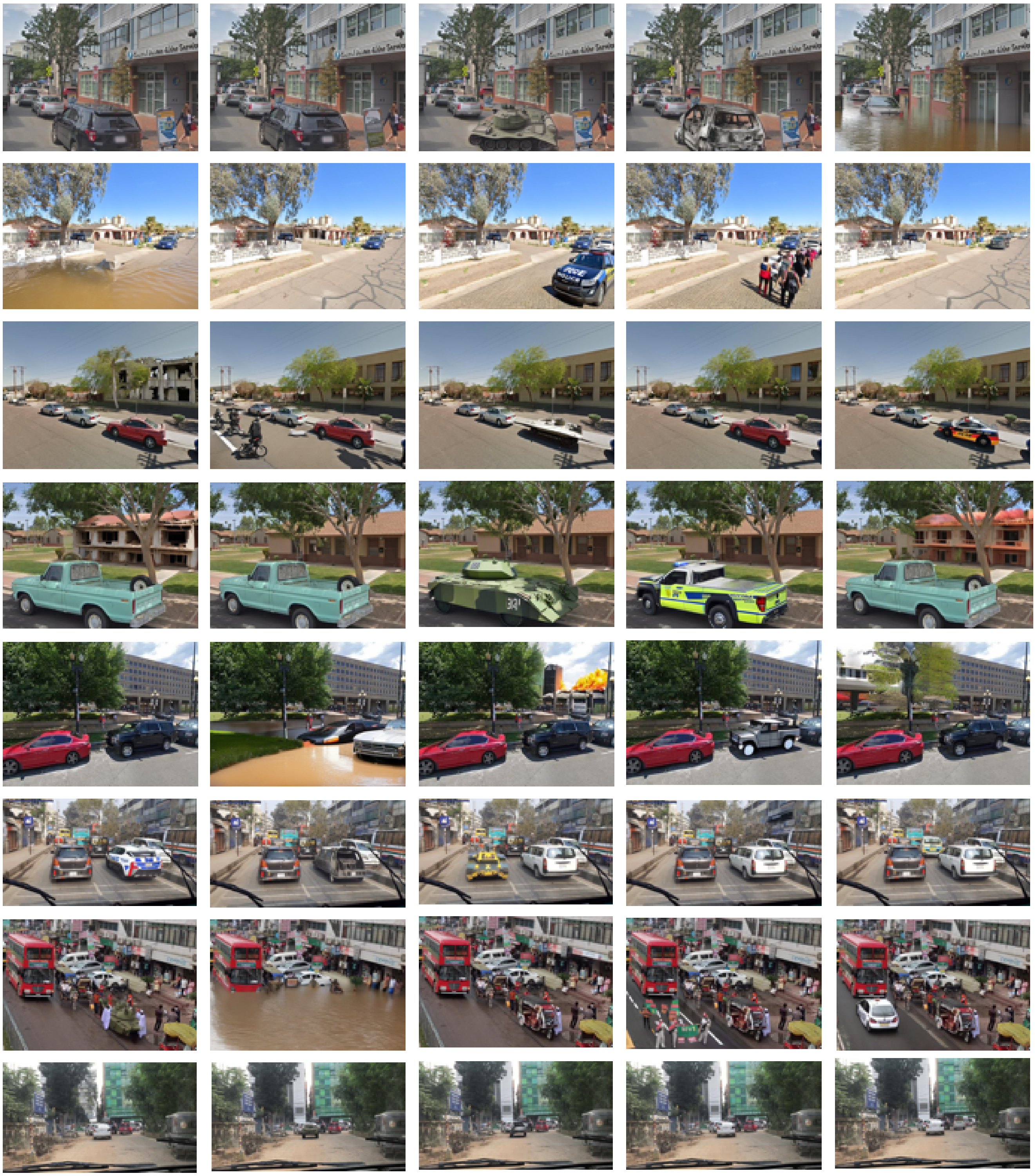

Figure A14. Examples of the manipulated images from our CSI-IMD gold standard set (10 of 10).

References

[1] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2, 3

[2] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[3] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 2

[4] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *European Conference on Computer Vision (ECCV),*, pages 312–328. Springer, 2020. 2

[5] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision (IJCV)*, 2022. 2, 3

[6] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. 2, 3

[7] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 2

[8] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2022. 2, 3

[9] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[10] Chao Yang, Huizhou Li, Fangting Lin, Bin Jiang, and Hao Zhao. Constrained r-cnn: A general image manipulation detection model. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2