

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-



The image shows a building engulfed in flames with a significant amount of smoke billowing into the sky. The fire appears to be intense, with flames visible through the windows and the roof of the structure. There are emergency vehicles, including a fire truck, parked nearby, indicating that firefighters are on the scene to address the situation. The street in front of the building is lined with parked cars, and there are traffic lights and a fire hydrant visible. The sky is partly cloudy, suggesting that the weather is not extreme, which might be a factor in the firefighting efforts. The overall scene is one of an emergency response to a building fire.



The image shows a street scene with significant flooding. The water appears to be a muddy brown color, indicating it might be mixed with sediment or debris. A Target store is visible on the left side of the image, with its sign partially submerged in the water. The street is lined with parked cars that are also partially submerged, and a few cars are driving through the flooded area. There are traffic signs and a construction barrier partially submerged as well. The buildings in the background are multi-story residential or commercial structures, and the sky is partly cloudy. The overall scene suggests that this is a result of heavy rainfall or a nearby waterway overflowing its banks.

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.



The image shows an urban street scene with a large building on the left side. The building appears to be a commercial structure with a red awning and a sign that is partially obscured, but it seems to be a store or a business establishment. The facade is made of a light-colored material, possibly concrete or stone, and it has several windows. On the right side of the image, there is a sidewalk where a person is walking, and a bus stop shelter is visible. The street is paved with what looks like cobblestones or a similar type of paving, and there are a few cars parked along the street. The sky is clear, suggesting it might be a sunny day. There are no visible trees or greenery in the immediate vicinity, which gives the scene an urban, built-up feel.



The image shows an urban street scene with a building that appears to be a commercial or office space, with large windows and a modern architectural style. The facade is beige or light brown, and there are a few visible air conditioning units on the exterior. In front of the building, there is a sidewalk where a person is walking, and a green trash can is placed on the curb. The street is paved with cobblestones, and there are parked cars along the side of the road. The sky is clear, suggesting it might be a sunny day. The overall atmosphere of the image is calm and orderly, typical of a quiet city street during the day. There are no visible people in the immediate vicinity of the building, which gives the impression of a moment of tranquility in an otherwise bustling urban environment.

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.

ffects 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

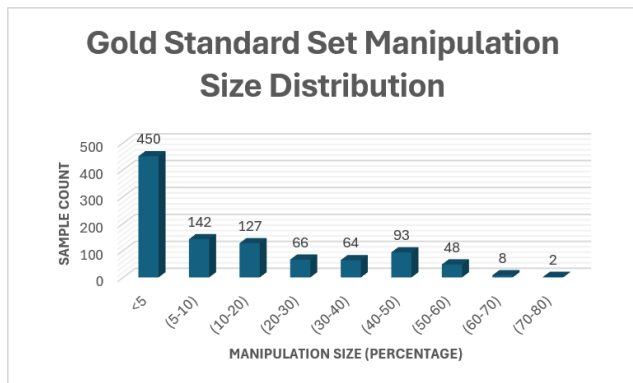


Figure A3. Distribution of manipulated images based on manipulation size

why SoTA methods struggle to localize manipulations generated by generative stable diffusion models: (1) mis-detection/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 CRCNN, 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.

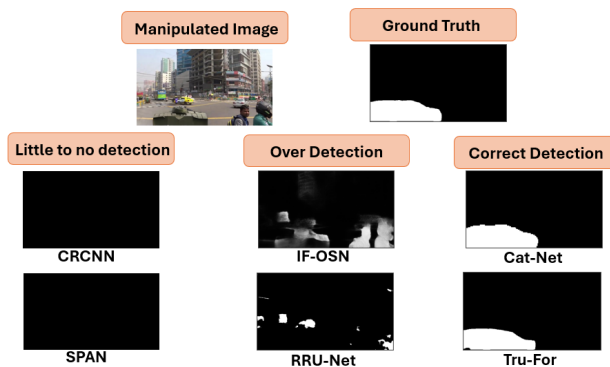


Figure A4. Qualitative example of state of the art method’s prediction localization masks. Examples of no detection, over detection, and correct detection’s are shown.

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

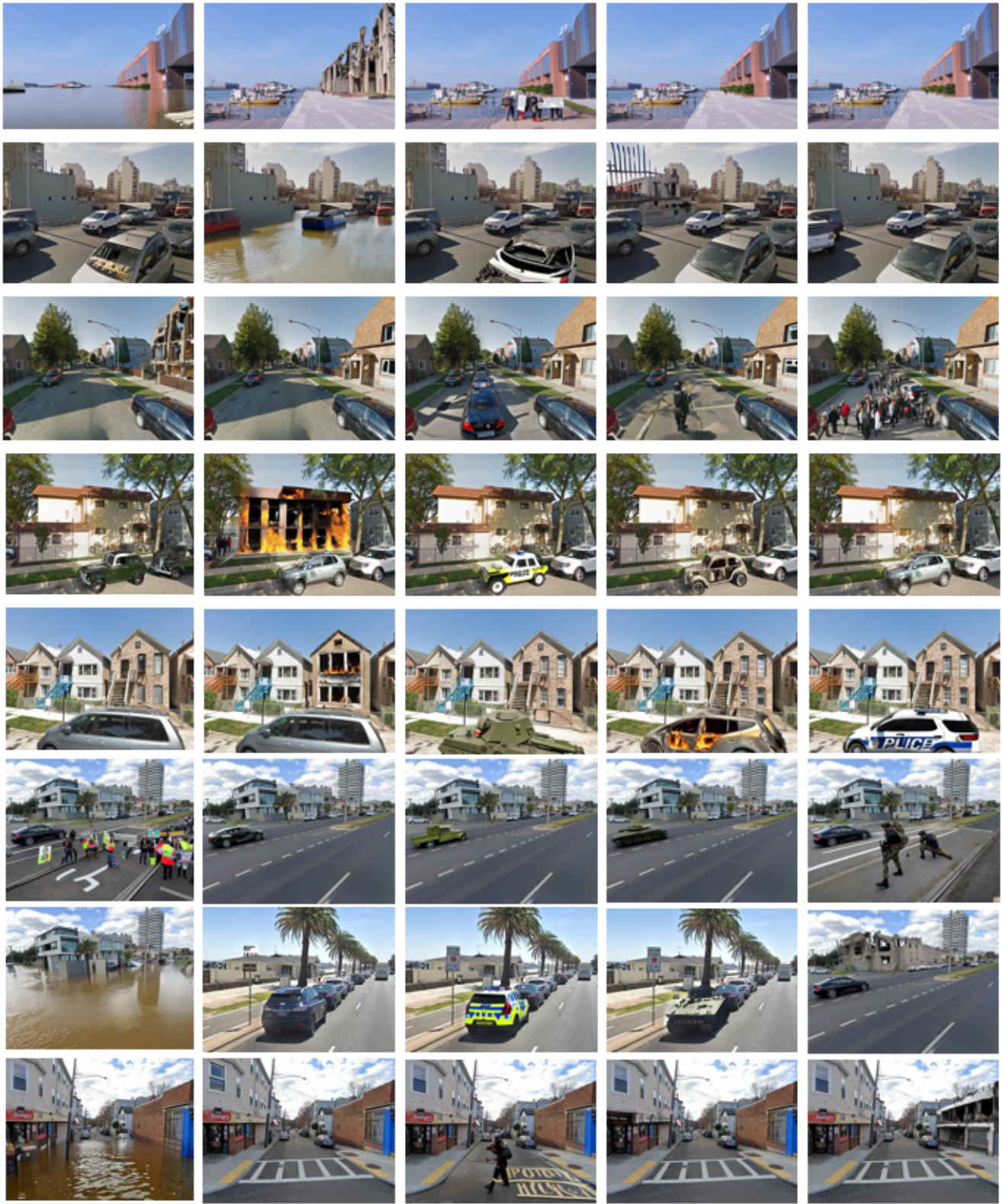
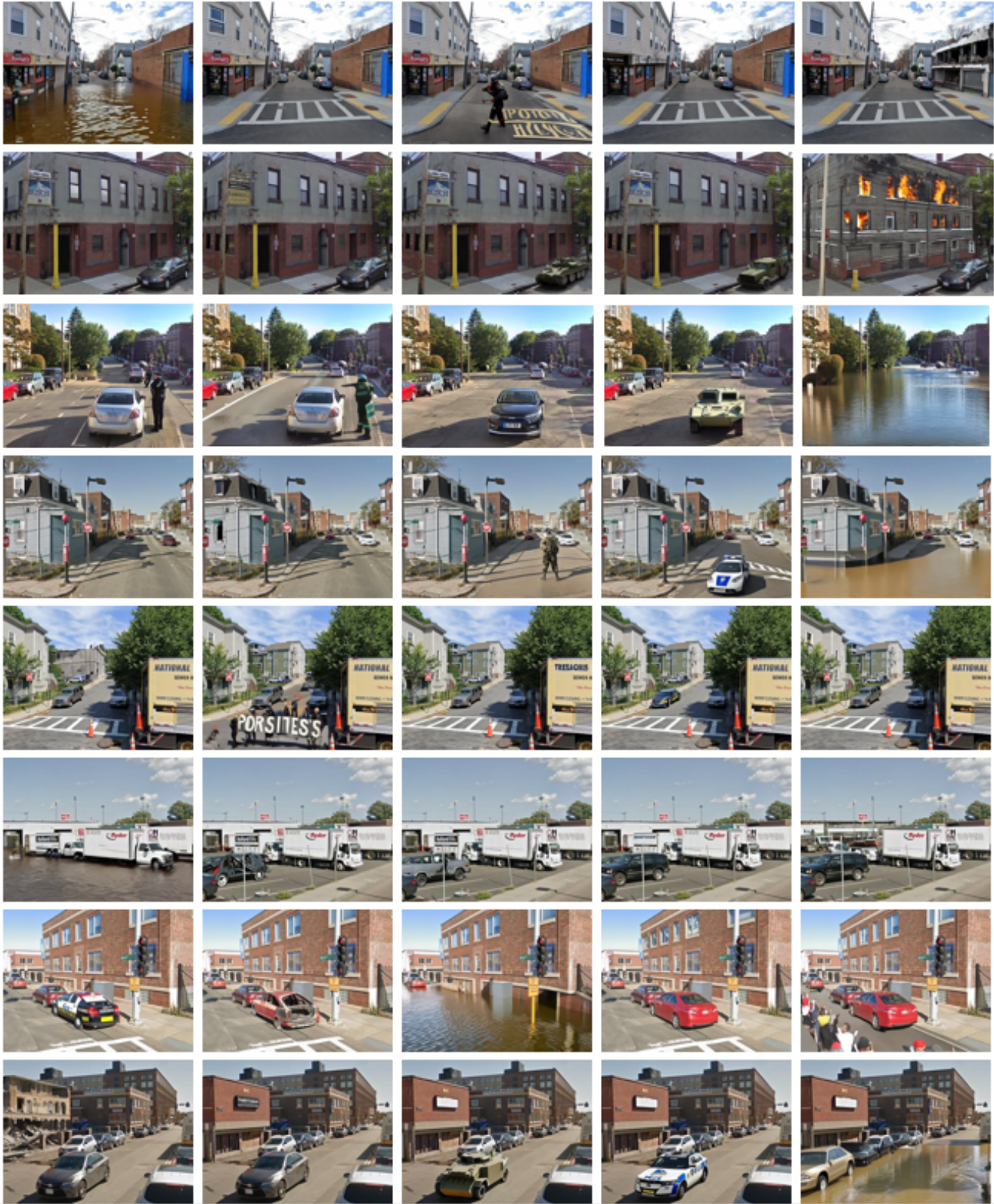


Figure A5. Examples of the manipulated images from our CSI-IMD gold standard set (1 of 10).

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485



486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

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

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

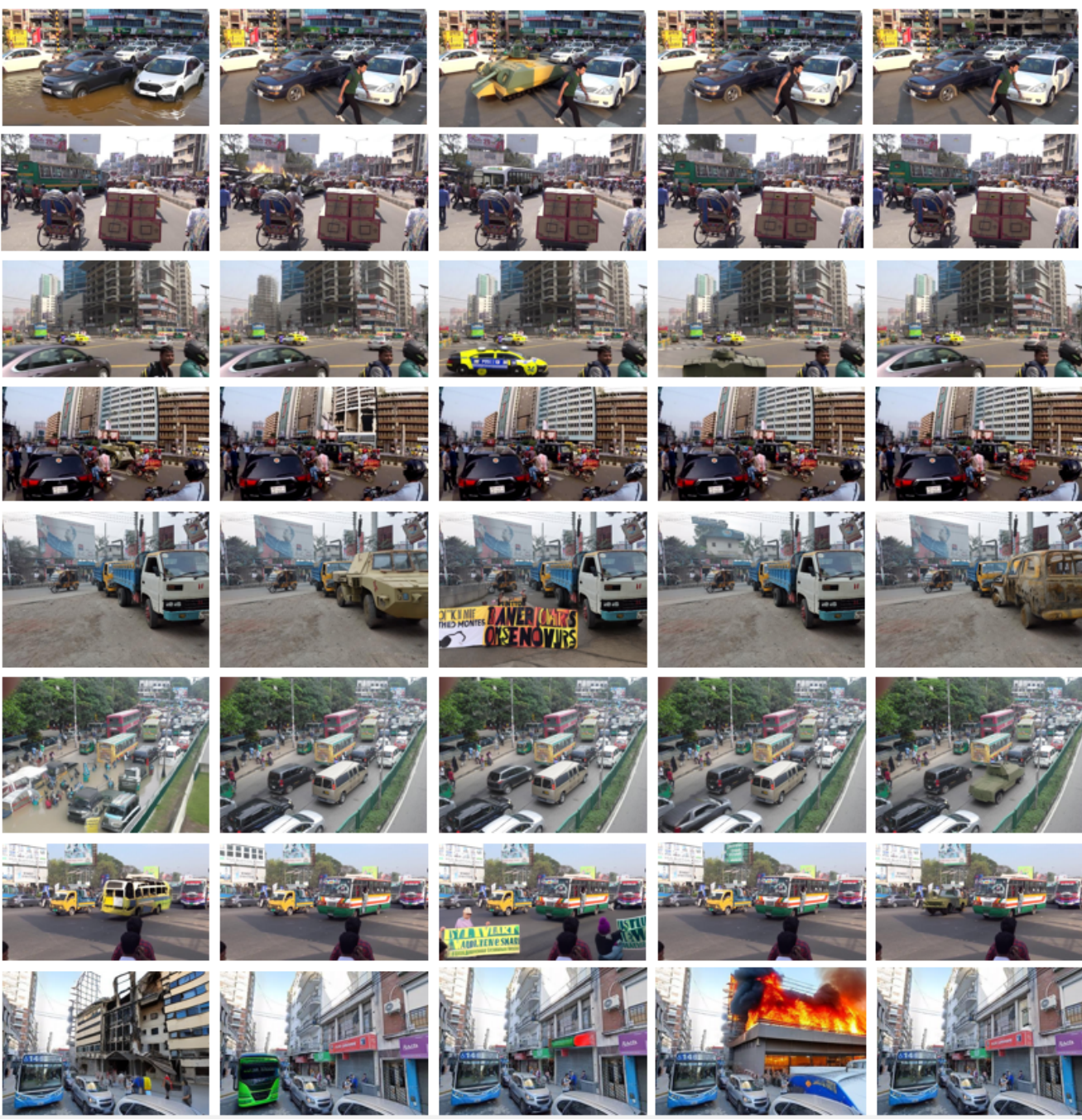
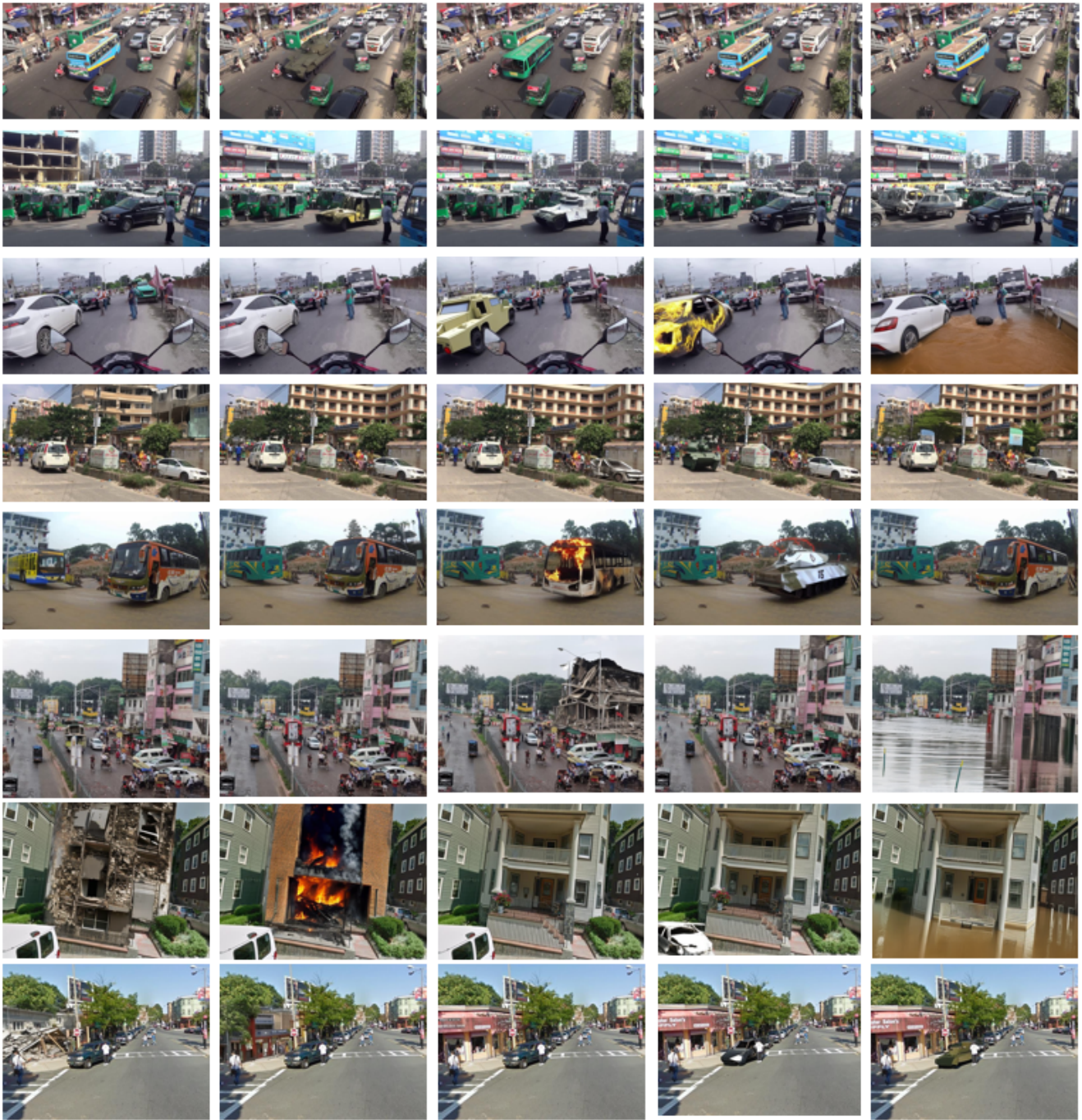


Figure A7. Examples of the manipulated images from our CSI-IMD gold standard set (3 of 10).

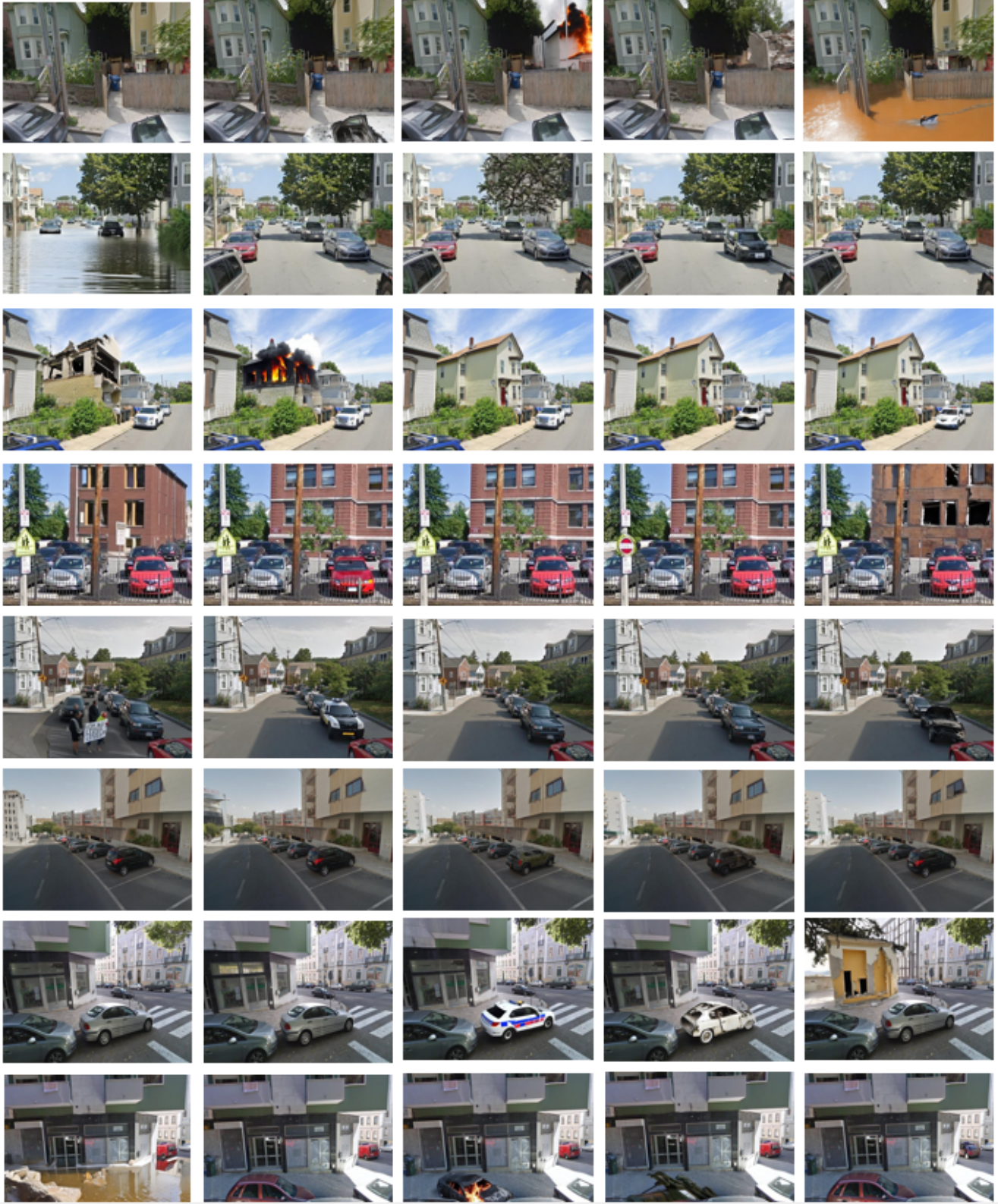
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701



702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

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

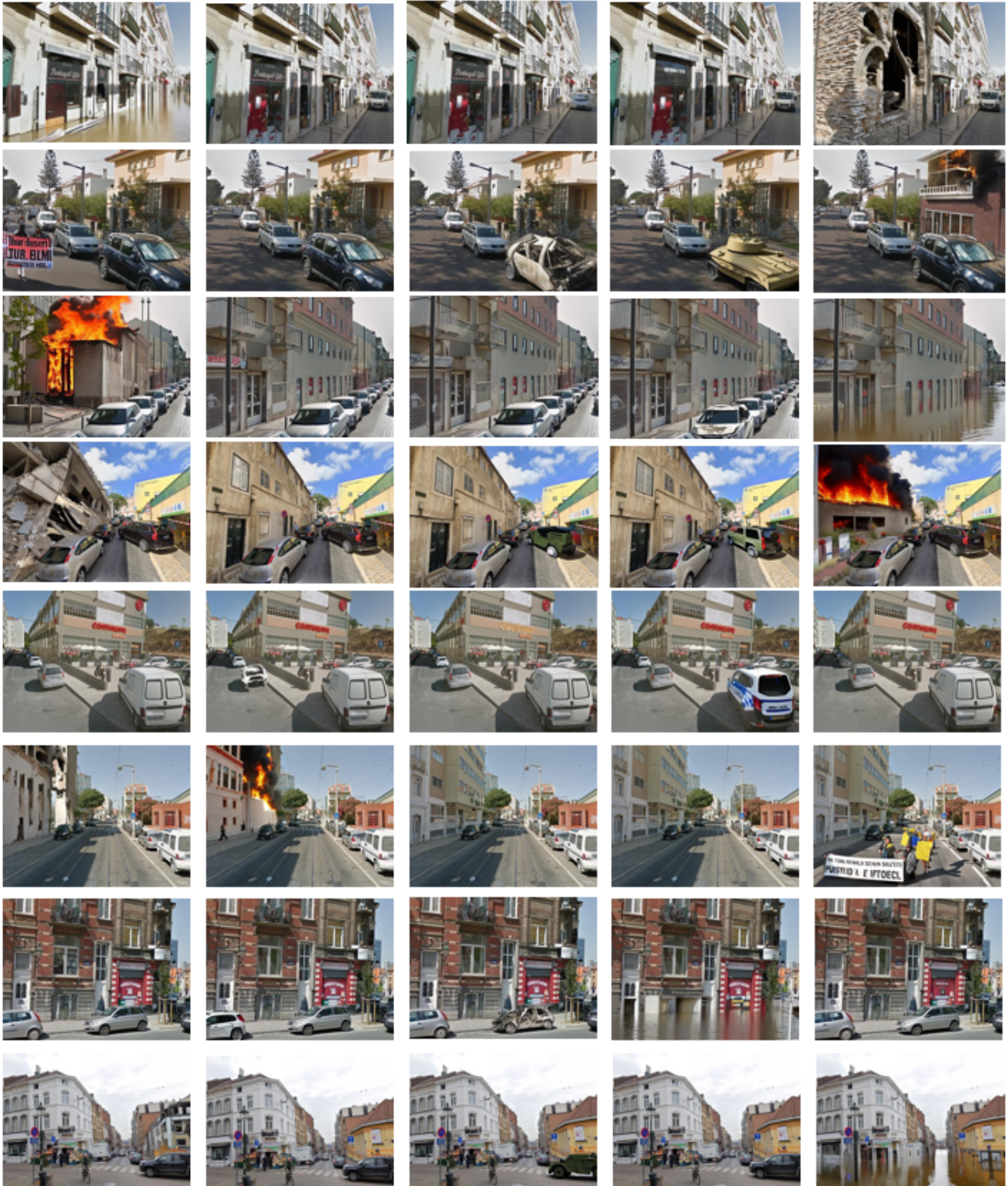
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809



810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

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

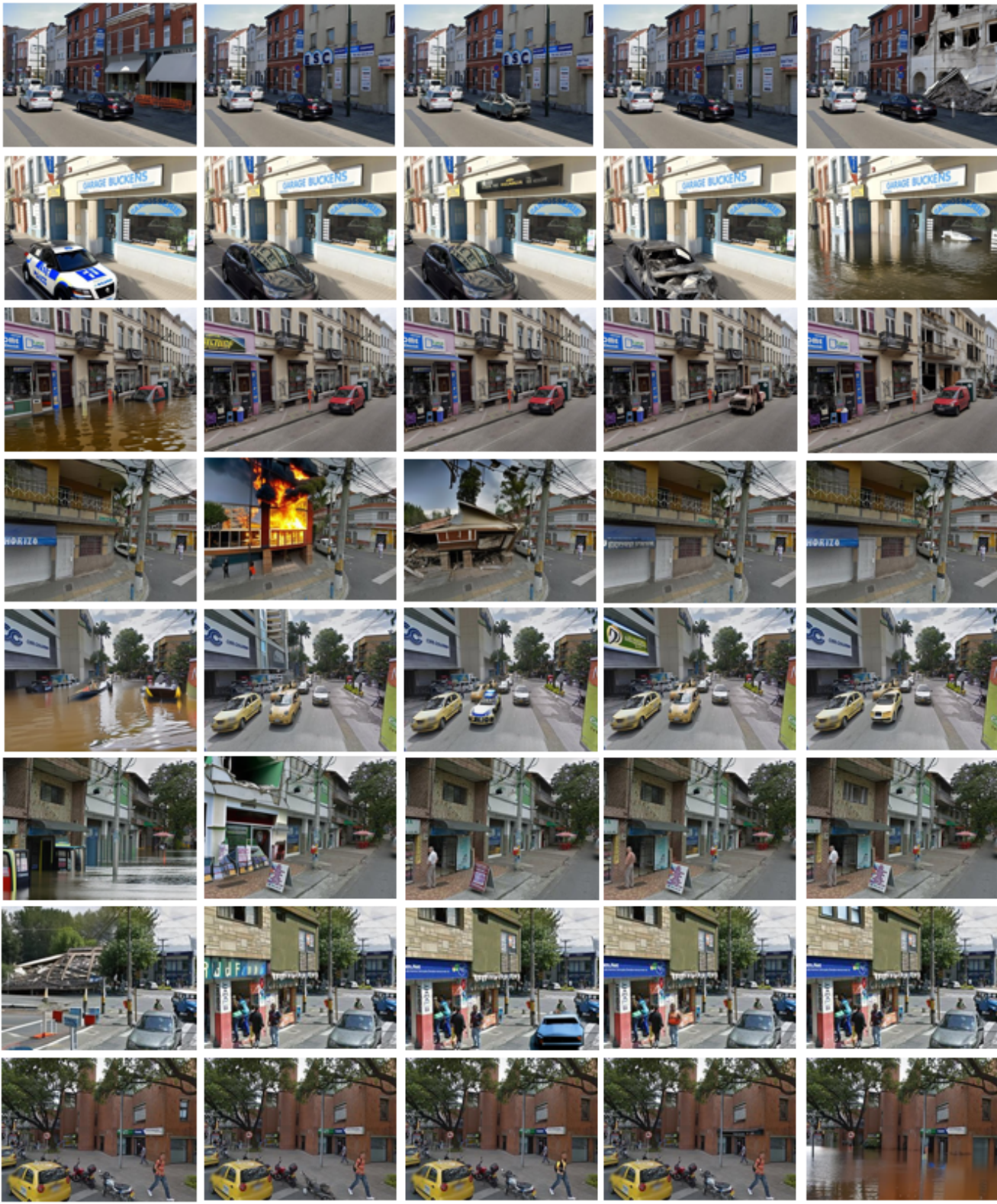
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917



918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

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

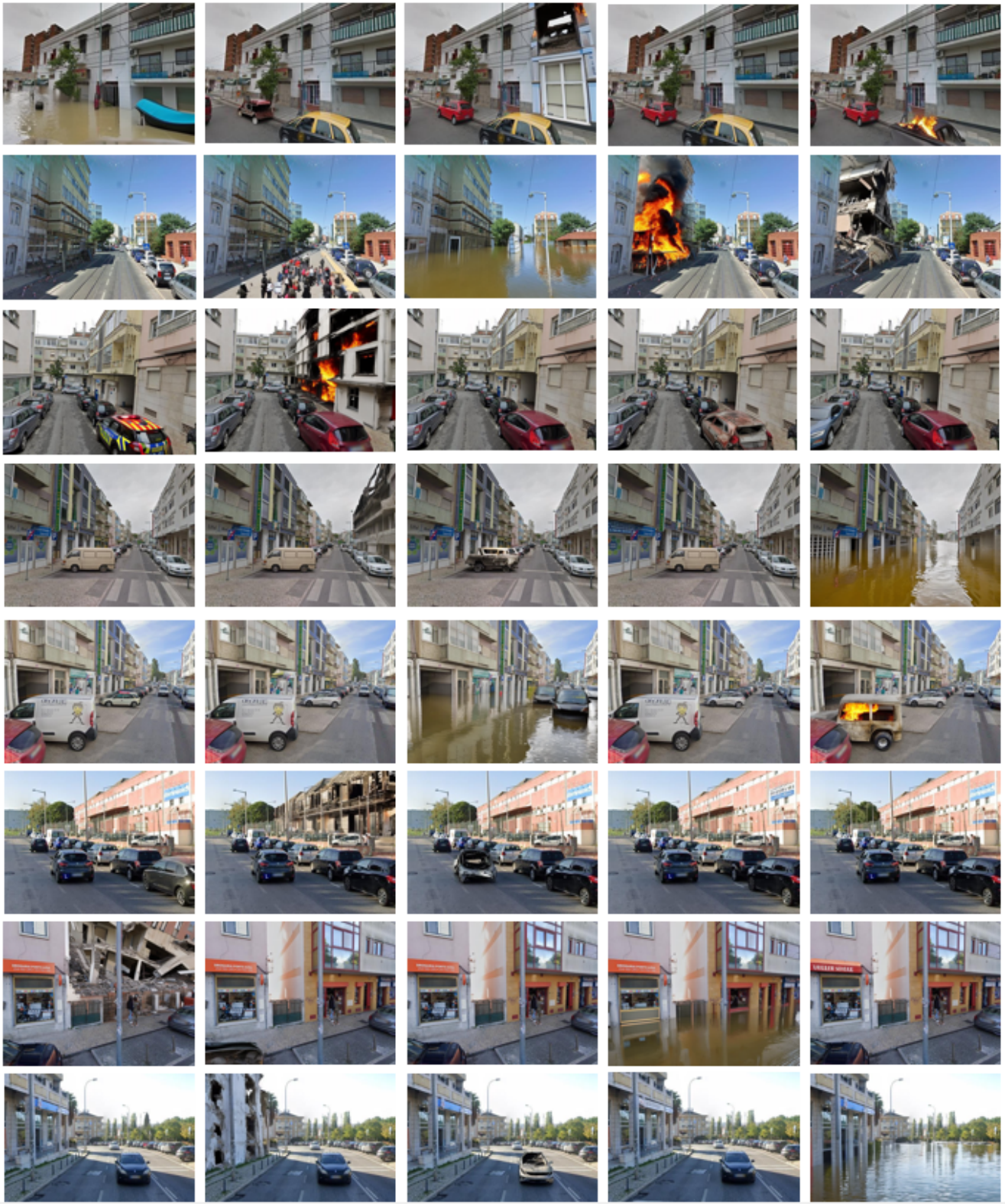
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025



1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

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

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133



1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

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

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

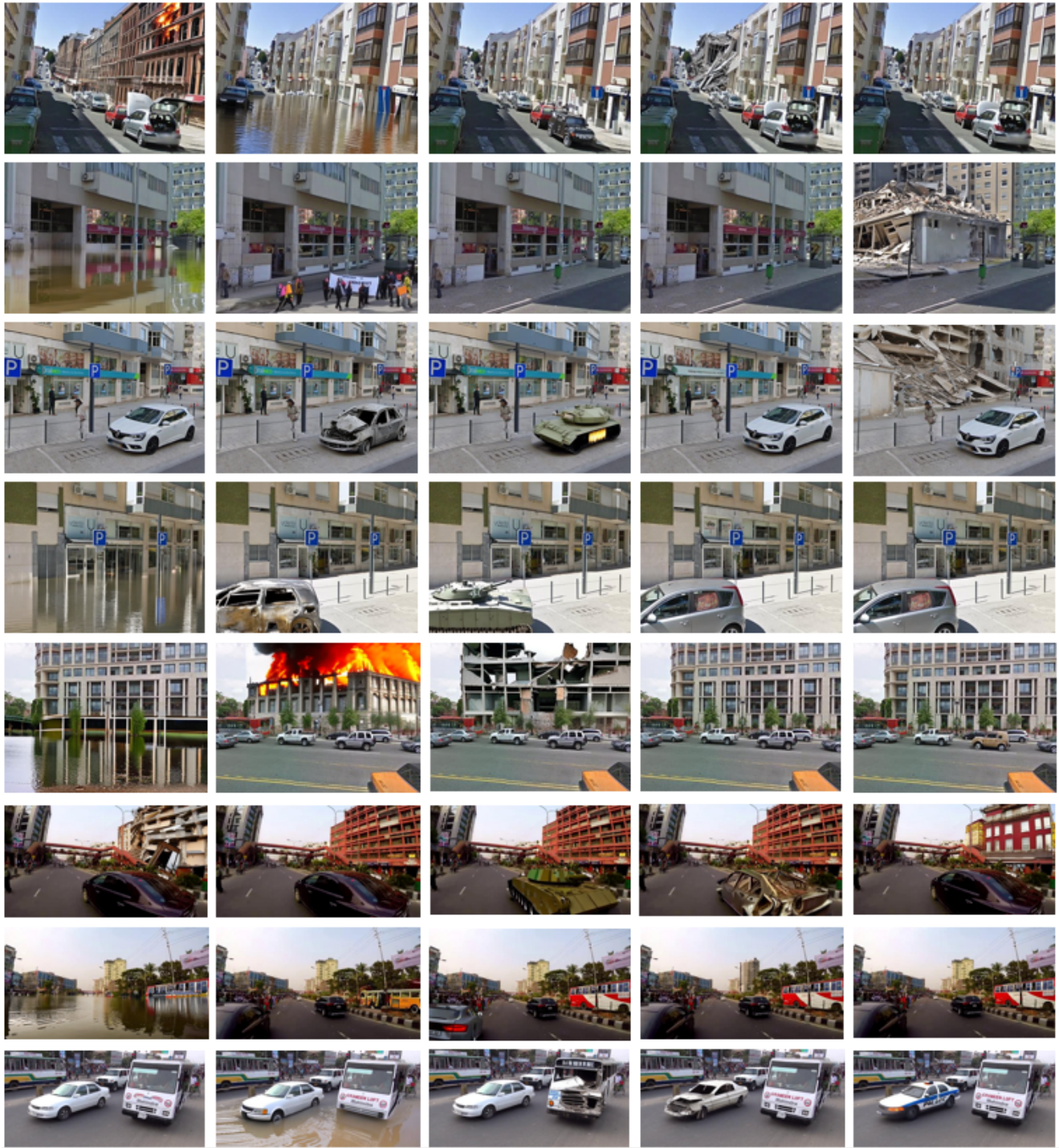


Figure A13. Examples of the manipulated images from our CSI-IMD gold standard set (9 of 10).

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

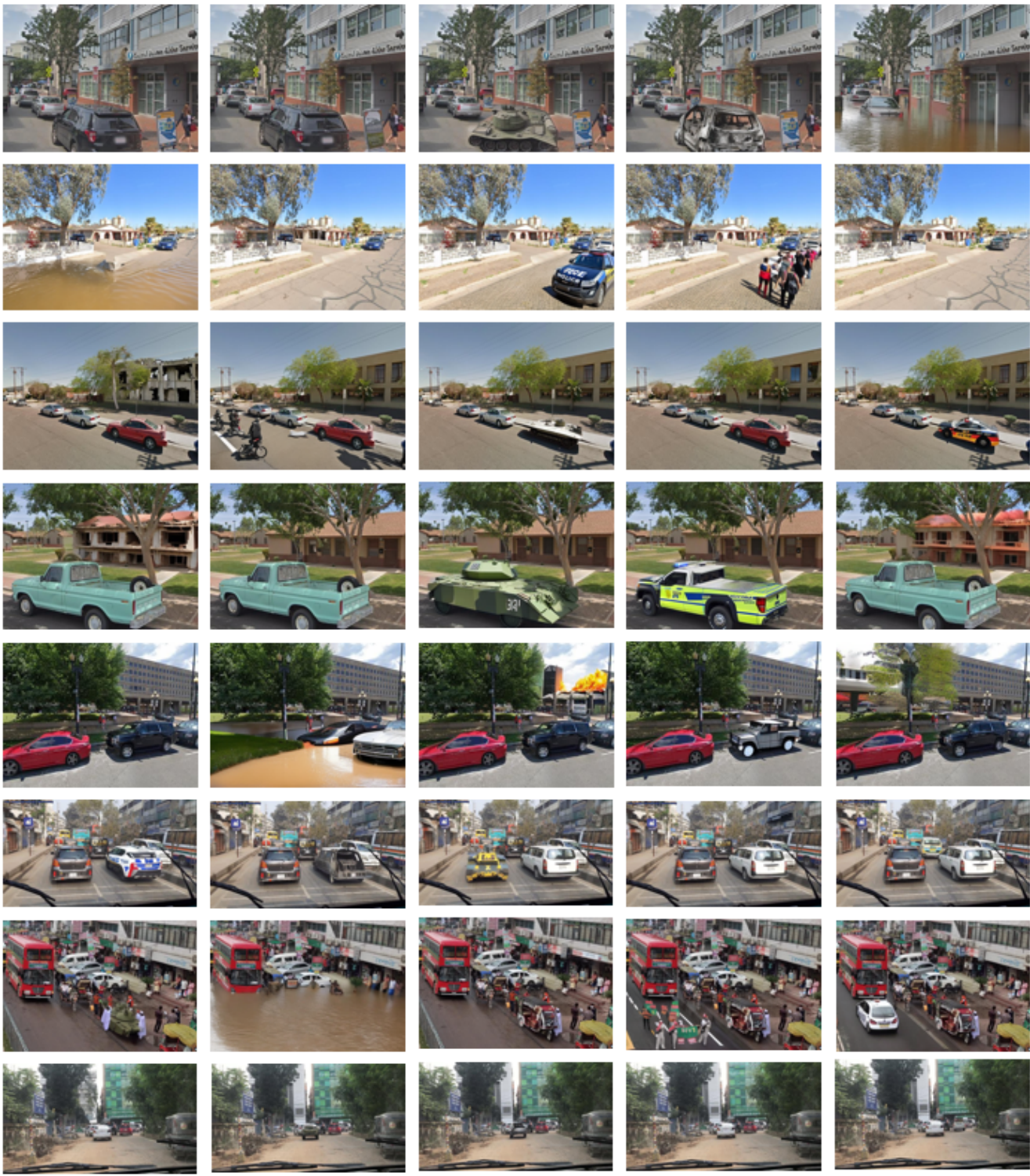


Figure A14. Examples of the manipulated images from our CSI-IMD gold standard set (10 of 10).

References

- 1404
1405
1406 [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 1460
1407
1408
1409
1410
1411 [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 1462
1412
1413
1414
1415
1416 [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 1463
1417
1418
1419
1420
1421 [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 1464
1422
1423
1424
1425
1426
1427 [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 1465
1428
1429
1430
1431
1432
1433 [6] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscn-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 1466
1434
1435
1436
1437
1438
1439 [7] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yungang Jiang. Objectformer for image manipulation detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 2 1467
1440
1441
1442
1443
1444
1445 [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 1468
1446
1447
1448
1449
1450
1451 [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 1469
1452
1453
1454
1455
1456
1457 [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 1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511