TARGET DETECTION AND CLASSIFICATION PERFORMANCE ENHANCEMENT USING SUPER-RESOLUTION INFRARED VIDEOS

Chiman Kwan, David Gribben and Bence Budavari

Applied Research, LLC, Rockville, Maryland, USA

ABSTRACT

Long range infrared videos such as the Defense Systems Information Analysis Center (DSIAC) videos usually do not have high resolution. In recent years, there are significant advancement in video super-resolution algorithms. Here, we summarize our study on the use of super-resolution videos for target detection and classification. We observed that super-resolution videos can significantly improve the detection and classification performance. For example, for 3000 m range videos, we were able to improve the average precision of target detection from 11% (without super-resolution) to 44% (with 4x super-resolution) and the overall accuracy of target classification from 10% (without super-resolution) to 44% (with 2x super-resolution).

KEYWORDS

Deep learning, mid-wave infrared (MWIR) videos, target detection and classification, contrast enhancement, YOLO, ResNet

1. INTRODUCTION

For infrared videos, two groups of target detection algorithms are normally used in the literature. The first group is to apply supervised machine learning algorithms. For example, some conventional target tracking methods [1]-[4] normally require target locations in the first frame of the videos to be known. Another group uses deep learning algorithms such as You Only Look Once (YOLO) for optical and infrared videos [5]-[20]. Target locations in the first frame are not needed. However, training videos are required in these algorithms. Some of those deep learning algorithms [6]-[16] are using compressive measurements directly for target detection and classification, meaning that no time consuming reconstruction of compressive measurements is needed and hence fast target detection and classification can be achieved.

In long range infrared videos such as the DSIAC dataset, which has videos from 1000 m to 5000 m, the target size is small, the resolution is low, and the video quality is also low. It is therefore extremely important to apply practical methods that can improve the detection and classification performance using deep learning methods. In recent years, there have been huge progress in image super-resolution development. Many high performance algorithms have been developed. It will be interesting to investigate the incorporation of super-resolution videos for target detection and classification.

In this paper, we present some results on target detection and classification in infrared videos. The objective is to see how much gain one can get if one integrates video super-resolution with target detection and classification algorithms. Our approach consists of the following steps. First, we propose to apply a state-of-the-art video super-resolution algorithm to enhance the resolution of

the videos by two to four times. We have compared two deep learning algorithms for video super-resolution. After that, we customize YOLO for target detection and a residual network (ResNet) for classification. The YOLO is responsible for target detection and the target locations will be passed to ResNet for classification. All the deep learning algorithms were customized using videos from one particular range and other videos from other ranges were used for testing.

In the experiments of this paper, we focus on the DSIAC MWIR videos, which do not have high resolution and hence it is very challenging. Due to long ranges, the vehicle size is quite small, which then seriously affects the target detection and classification performance. Our contributions are as follows:

- We investigated the application of a recent deep learning based video super-resolution (VSR) algorithm to enhance the resolution of DSIAC videos. Two and four times resolution enhanced videos were generated.

- We built a YOLO model for target detection and a ResNet model for target classification using only original resolution videos at 1500 m. We then fed the enhanced resolution videos into the trained YOLO and ResNet models and generated the detection and classification statistics using videos at 2000 m, 2500 m, 3000 m, and 3500 m ranges. We observed that detection and classification statistics using enhanced videos have improved tremendously. In particular, the intersection of union (IoU) and average precision (AP) statistics have improved quite a lot in almost all ranges.

Our paper is organized as follows. Section 2 describes the VSR algorithm, target detection and classification algorithms, performance metrics, and infrared videos. Section 3 summarizes the experimental results. Finally, some remarks are included in Section 4.

2. METHODS, PERFORMANCE METRICS, AND DATA

The proposed approach consists of three steps. First, we apply VSR to improve the image resolution by two to four times. Second, we apply YOLO to locate the targets. Third, we apply ResNet to the YOLO detected outputs and then classify the vehicles. All three steps require training. We used 1500 m range videos to train the above algorithms.

2.1. Video Super Resolution (VSR) Using Deep Learning

In [26], researchers developed a deep learning video super resolution (VSR) method known as Zooming-Slow Motion (ZSM) in 2020. The ZSM can improve both the resolution of the frames and the frame rate. There are three key components in ZSM: feature temporal interpolation network, a deformable ConvLSTM, and a deep construction network [27]. The feature temporal interpolation network is to perform frame interpolation. The bidirectional deformable ConvLSTM is for aligning and aggregating temporal information. The deep construction network is to predict and generate super resolution video frames. This ZSM architecture is depicted in Figure 1.

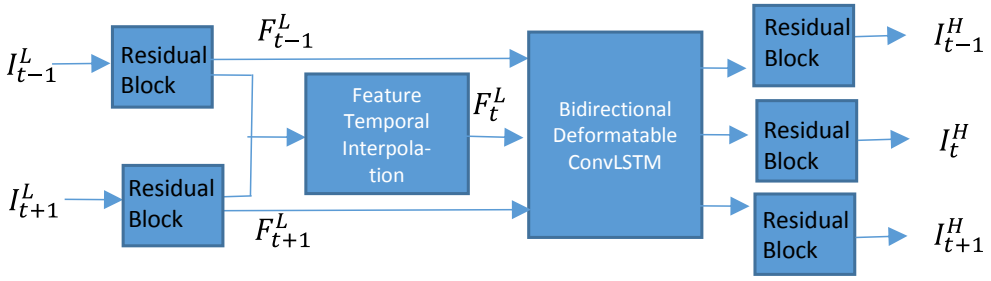

Figure 1. ZSM Architecture.

The visual performance of the various super-resolution methods is shown in Figure 2. We compared ZSM with Dynamic Upsampling Filter (DUF) [25] and bicubic methods. One can see that ZSM method yielded better results than the other methods in terms of clarity.

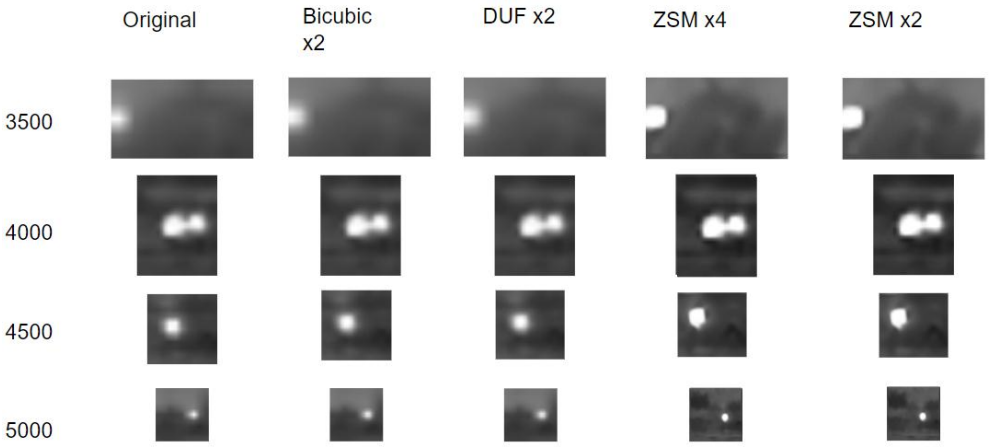

Figure 2. Comparison of the image qualities of different videos at different ranges and super-resolution methods.

2.2. YOLO for Target Detection

YOLO and Faster R-CNN are some deep learning based object detectors that do not require initial bounding boxes and can simultaneously detect objects. YOLO [22] and Faster R-CNN [23] have comparable performance. The input image is normally resized to 448x448. There are 24 convolutional layers and two fully connected layers. The output is 7x7x30. We have used YOLOv2 because it is more accurate than YOLO version 1. The training of YOLO is simple where images with ground truth target locations are used. We re-trained the last layer of YOLO using the 1500 m range videos. YOLO took approximately 2000 epochs to complete the training.

Although YOLO has a built-in classification module, the classification accuracy using YOLO's built-in module is not good as compared to ResNet [5]-[6].

2.3. ResNet for Target Classification

YOLO has been widely used for object detection such as humans, traffic signs, vehicles, buses, etc. Its built-in classifier is, however, not so good for intra-class (e.g. BTR70 vs. BMP2) discrimination. The ResNet-18 model [24] is an 18-layer convolutional neural network (CNN) that can avoid performance saturation when training deeper layers.

The relationship between YOLO and ResNet in our paper can be explained in this way. YOLO [22] was used to determine where the vehicles were located in each frame. YOLO generated bounding boxes for those vehicles and that data were used to crop the vehicles from the image. The cropped vehicles will be fed into the ResNet-18 for classification. To be specific, ResNet-18 is used directly after bounding box information is obtained from YOLO.

Training of ResNet requires target patches. The targets are cropped from training videos. Mirror images are then created. Data augmentation using scaling, rotation (every 45 degrees), and illumination variations was used to generate more training data. For each cropped target, we created a data set with 64 more images. We re-trained the last layer of the ResNet model, which was until the validation score reached a steady state value.

2.4. Performance Metrics for Assessing Target Detection and Classification Performance

The six different performance metrics to quantify the detection performance are: Center Location Error (CLE), Distance Precision at 10 pixels (DP@10), Estimates in Ground Truth (EinGT), Intersection over Union (IoU), Average Precision (AP), and number of frames with detection. These metrics are as follows:

- Center Location Error (CLE): This is the error between the center of the bounding box and the ground-truth bounding box. Smaller means better. CLE is calculated by measuring the distance between the ground truth center location $(C_{x,gt}, C_{y,gt})$ and the detected center location $(C_{x,est}, C_{y,est})$. Mathematically, CLE is given by

$$CLE = \sqrt{\left(C_{x,est} - C_{x,gt}\right)^2 + \left(C_{y,est} - C_{y,gt}\right)^2}. \quad (1)$$

- Distance Precision (DP): This is the percentage of frames where the centroids of detected bounding boxes are within 10 pixels of the centroid of ground-truth bounding boxes. Close to 1 or 100% indicates good results.
- Estimates in Ground Truth (EinGT): This is the percentage of the frames where the centroids of the detected bounding boxes are inside the ground-truth bounding boxes. It depends on the size of the bounding box and is simply a less strict version of the DP metric. Close to 1 or 100% indicates good results.
- Intersection over the Union (IoU): It is the ratio of the intersected area over the union of the estimated and ground truth bounding boxes.

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \quad (2)$$

- Average Precision (AP): AP is the ratio between the intersection area and the area of the estimated bounding box and the value is between 0 and 1, with 1 or 100% being the perfect overlap. The AP being used can be computed as

$$AP = \frac{Area\ of\ Intersection}{Area\ of\ estimated\ bounding\ boxes}. \quad (3)$$

- Number of frames with detection: This is the total number of frames that have detection.

The confusion matrices were used for evaluating vehicle classification performance using ResNet. From the confusion matrix, we can also evaluate overall accuracy (OA), average accuracy (AA), and kappa coefficient.

2.5. DSIAC Data

We selected five vehicles in the DSIAC videos for detection and classification. There are optical and mid-wave infrared (MWIR) videos collected at distances ranging from 1000 m to 5000 m with 500 m increments. The five types of vehicles are shown in Figure 3. These videos are challenging for several reasons. First, the target sizes are small due to long ranges. This is very different from some benchmark datasets such as MOT Challenge [21] where the range is short and the targets are big. Second, the target orientations also change drastically. Third, the illuminations in different videos are also different. Fourth, the cameras also move in some videos.

In this research, we focus mainly on MWIR night-time videos because MWIR is more effective for surveillance during the nights. The video frame rate is 7 frames/second and the image size is 640x512. The total number of frames is 1800 per video. Each pixel is represented by 8 bits. All frames are contrast enhanced using some reference frames in the 1500 m range videos.

(a) (b) (c)

(d) (e)

Figure 3. Five vehicles in DSIAC: (a) BTR70; (b) BRDM2; (c) BMP2; (d) T72; and (e) ZSU23-4.

3. EXPERIMENTS

3.1. Baseline Results Using Original Resolution Videos

Here, baseline means that the YOLO and ResNet models trained using 1500 m videos were tested using the original resolution videos. The baseline performance metrics will be used as a baseline to compare the results of using SR videos. There are five different distances that have results: 1500 m, 2000 m, 2500 m, 3000 m, and 3500 m. Table 1 contains the YOLO detection statistics for the each distance while Table 2 contains the ResNet confusion matrices and classification statistics. There is an obvious deterioration in accuracy as the vehicle distance moves from 1500 meters, the distance the model was trained on.

From Table 1 and Table 2, each metric trends worse as it moves further away from the trained 1500 meter distance. This is a trend that is seen across both detection and classification statistics. The overall degradation in accuracy as distances move from the trained distances is quite extreme. For example, with detection the AP value, measuring the amount of overlap between ground truth and detected bounding box, halves with each increase of 500 meters. The final distance, 3500, is one-fifth the previous distances value.

Table 1. Baseline YOLO detection results using original resolution videos. The metrics are named as follows: Center Location Error (CLE), Distance Precision (DP), Estimates in Ground Truth (EinGT), Intersection over Union (IoU), Average Precision (AP), and Detection Percentage (% det.).

1500 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	1.201	100.00%	100.00%	70.56%	70.60%	91.17%
BRDM2	1.279	100.00%	100.00%	78.54%	78.69%	91.06%
BMP2	1.092	100.00%	100.00%	87.70%	88.96%	91.06%
T72	1.497	100.00%	100.00%	85.21%	86.25%	91.11%
ZSU23-4	1.233	100.00%	100.00%	77.58%	77.75%	90.00%
Avg	1.260	100.00%	100.00%	79.92%	80.45%	90.88%

2000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	1.861	100.00%	100.00%	30.64%	30.64%	93.44%
BRDM2	3.023	100.00%	100.00%	37.74%	37.75%	90.50%
BMP2	3.542	100.00%	100.00%	58.01%	58.69%	41.83%
T72	2.276	100.00%	100.00%	39.80%	39.80%	98.44%
ZSU23-4	8.953	97.83%	97.83%	38.11%	38.11%	84.56%
Avg	3.931	99.57%	99.57%	40.86%	41.00%	81.76%

2500 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	2.374	100.00%	100.00%	13.43%	13.43%	91.17%
BRDM2	3.68	100.00%	99.26%	16.44%	16.44%	90.50%
BMP2	24.834	89.79%	89.79%	23.42%	23.42%	75.61%
T72	3.253	99.95%	99.95%	20.99%	20.99%	78.89%
ZSU23-4	2.688	100.00%	100.00%	17.37%	17.37%	73.39%
Avg	7.366	97.95%	97.80%	18.33%	18.33%	81.91%

3000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	1.995	100.00%	99.55%	7.09%	7.09%	49.00%
BRDM2	3.499	100.00%	99.38%	10.33%	10.33%	27.06%
BMP2	3.91	100.00%	98.48%	17.31%	17.31%	3.61%
T72	4.541	100.00%	77.02%	11.86%	11.86%	32.56%
ZSU23-4	2.18	100.00%	99.67%	11.25%	11.25%	15.67%
Avg	3.225	100.00%	94.82%	11.57%	11.57%	25.58%

3500 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	1.689	100.00%	83.52%	2.16%	2.16%	36.89%
BRDM2	3.112	99.48%	74.55%	2.65%	2.65%	42.56%
BMP2	1.886	100.00%	100.00%	3.80%	3.80%	0.17%
T72	3.975	99.86%	55.59%	2.84%	2.84%	36.39%
ZSU23-4	2.541	100.00%	69.80%	2.73%	2.73%	39.17%
Avg	2.641	99.87%	76.69%	2.84%	2.84%	31.03%

Table 2. Baseline ResNet classification results using original resolution videos. Confusion matrices with Overall Accuracy (OA), Average Accuracy (AA), and kappa.

1500 m	5	6	9	11	12	
BTR70	1849	0	0	0	2	
BRDM2	0	1808	0	0	0	
BMP2	0	0	1800	0	0	
T72	0	0	0	1829	0	
ZSU23-4	0	0	0	0	1882	
Class Stats	OA	99.98%	AA	99.98%	kappa	1.00

2000 m	5	6	9	11	12	
BTR70	1511	49	167	56	84	
BRDM2	0	1834	18	12	37	
BMP2	7	30	715	0	2	
T72	15	272	159	1739	95	
ZSU23-4	0	90	191	0	1472	
Class Stats	OA	84.99%	AA	86.50%	kappa	0.85

2500 m	5	6	9	11	12	
BTR70	43	608	313	516	434	
BRDM2	0	1800	19	44	20	
BMP2	0	19	1347	47	105	
T72	2	98	561	938	378	
ZSU23-4	65	236	391	662	554	
Class Stats	OA	50.89%	AA	52.61%	kappa	0.51

3000 m	5	6	9	11	12	
BTR70	1	69	271	12	540	
BRDM2	0	13	59	1	414	
BMP2	0	3	52	0	11	
T72	0	461	68	27	136	
ZSU23-4	2	56	56	29	156	
Class Stats	OA	10.22%	AA	27.53%	kappa	0.10

3500 m	5	6	9	11	12	
BTR70	48	31	368	65	192	
BRDM2	36	6	331	113	284	
BMP2	0	0	2	0	1	
T72	8	218	113	354	5	
ZSU23-4	4	10	491	172	78	
Class Stats	OA	16.66%	AA	27.06%	kappa	0.17

3.2. YOLO and ResNet Performance Using 2x Super-Resolution Videos

Here, we investigate the detection and classification performance of using 2x super-videos in the testing stage. The goal is to see if there is an improvement over the baseline results, which are the cases using the original resolution videos. 1500 m range videos (original resolution) were used for training and other ranges for testing. Table 3 shows the detection results of distances 2000 m through 3500 m and Table 4 shows the confusion matrices and classification results for the same distances.

Detection results shown in Table 3 are certainly improved by a good margin as compared to those using the original resolution videos. The largest improvement is seen at the further distances. This improvement is seen mostly in the IoU and AP results. In general, this makes sense for the same reasons mentioned when looking at the VSR cropped results shown in Figure 2. The other thing to compare is whether there is any change to the down-sampled video metrics at 3000 meters. The reduction in accuracy is a little less than two times in the IoU and AP values. There is no reduction in CLE—when taking the pixel difference into account—DP or EinGT. Unfortunately, percent detection is very hurt.

Classification present differently than detection results. In Table 4, the OA, AA, and kappa, the largest improvements are seen at the further distances. Compared with the baseline classification results, a slight improvement is seen at 2000 m but by 3000 m there is a 4 times improvement and 3500 there is a slightly greater than 2 times improvement. It is also interesting that for this version of VSR it is even better than the full resolution. The confusion matrix for 3000 meters has fewer data points due to the decrease in detection percentage but the metrics are around 30 percent more accurate.

Table 3. YOLO detection metrics for distances 2000, 2500, 3000, and 3500 meters using videos that have two times better resolution.

2000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	2.596	100.00%	100.00%	70.39%	72.30%	99.39%
BRDM2	6.459	99.93%	100.00%	61.81%	64.27%	71.78%
BMP2	4.347	99.65%	100.00%	67.15%	70.55%	46.72%
T72	3.626	100.00%	100.00%	65.27%	67.00%	85.17%
ZSU23-4	3.791	100.00%	100.00%	70.96%	76.58%	87.06%
Avg	4.164	99.92%	100.00%	67.12%	70.14%	78.02%

2500 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	4.652	99.89%	99.67%	24.89%	24.89%	49.06%
BRDM2	7.250	97.03%	95.91%	32.88%	32.92%	13.78%
BMP2	3.810	100.00%	100.00%	53.68%	54.46%	79.28%
T72	5.198	99.85%	100.00%	35.46%	35.70%	33.50%
ZSU23-4	3.829	100.00%	100.00%	45.13%	45.28%	56.17%
Avg	4.948	99.35%	99.12%	38.41%	38.65%	46.36%

3000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	4.143	100.00%	99.49%	11.93%	11.93%	19.78%
BRDM2	5.367	99.27%	93.82%	19.67%	19.67%	13.44%
BMP2	2.838	100.00%	100.00%	39.86%	39.86%	11.83%
T72	6.636	99.28%	99.28%	20.77%	20.78%	14.28%
ZSU23-4	2.512	100.00%	100.00%	22.31%	22.31%	30.72%
Avg	4.299	99.71%	98.52%	22.91%	22.91%	18.01%

3500 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	3.276	100.00%	87.18%	4.66%	4.66%	47.83%
BRDM2	4.167	99.64%	94.71%	6.86%	6.86%	43.39%
BMP2	3.447	100.00%	95.87%	10.96%	10.96%	19.44%
T72	3.721	99.90%	80.93%	8.14%	8.14%	42.89%
ZSU23-4	2.739	100.00%	93.63%	7.48%	7.48%	76.78%
Avg	3.470	99.91%	90.46%	7.62%	7.62%	46.07%

Table 4. ResNet Confusion matrices and classification metrics for 2000 through 3500 meters using 2x super-resolution videos.

2000 m	5	6	9	11	12	
BTR70	1430	0	90	89	326	
BRDM2	4	1275	13	11	134	
BMP2	0	8	482	44	316	
T72	17	189	14	1296	152	
ZSU23-4	1	54	32	63	1649	
Class Stats	OA	88.92%	AA	90.23%	kappa	0.8892

2500 m	5	6	9	11	12	
BTR70	163	176	197	140	243	
BRDM2	0	237	9	17	6	
BMP2	0	0	1461	2	43	
T72	0	25	188	321	125	
ZSU23-4	3	6	59	114	886	
Class Stats	OA	69.40%	AA	66.90%	kappa	0.694

3000 m	5	6	9	11	12	
BTR70	22	21	63	8	275	
BRDM2	15	29	65	1	165	
BMP2	4	0	215	0	5	
T72	0	143	51	48	35	
ZSU23-4	0	20	74	56	484	
Class Stats	OA	44.36%	AA	41.17%	kappa	0.4436

3500 m	5	6	9	11	12	
BTR70	120	108	679	13	71	
BRDM2	6	75	427	0	323	
BMP2	1	3	359	0	0	
T72	0	379	260	252	100	
ZSU23-4	0	119	447	204	893	
Class Stats	OA	35.11%	AA	39.83%	kappa	0.3511

3.3. YOLO Performance Enhancement Using 4x Video Super-Resolution

Table 5 and Table 6 are for the full resolution VSR videos cropped to original video size. We only focused on 3000 m range because it is difficult to discern any objects beyond this range. It is observed that while the CLE value looks worse for the 3000 m distance versus the baseline results, the VSR method has four times as many pixels. When divided by four, the average CLE is an average of 1.905 pixels, which is almost half the 3.225 average for the baseline. The same can be said for the DP value, which has a static 20 pixel distance. Otherwise, the results are greatly improved for both detection and classification with IoU and AP being four times more accurate. Compared with baseline results, the OA and kappa are three times more accurate.

Table 5. 3000 m range YOLO target detection results using 4x super-resolution videos.

3000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	9.22	93.22%	98.18%	47.95%	48.91%	30.00%
BRDM2	6.403	99.30%	100.00%	55.49%	56.42%	54.83%
BMP2	6.481	98.67%	99.67%	40.12%	40.36%	59.61%
T72	7.221	100.00%	100.00%	33.11%	33.22%	57.61%
ZSU23-4	8.719	99.87%	100.00%	43.18%	43.45%	40.33%
Avg	7.61	98.21%	99.57%	43.97%	44.47%	48.48%

Table 6. 3000 m range ResNet confusion matrix and classification metrics using 4x super-resolution videos.

3000	5	6	9	11	12	
BTR70	139	259	44	363	390	
BRDM2	3	260	115	28	375	
BMP2	2	38	543	265	287	
T72	3	391	16	143	52	
ZSU23-4	2	260	135	260	542	
Class Stats	OA	30.04%	AA	31.40%	kappa	0.3004

3.4. Performance Comparison between Original Resolution and Super-Resolution Videos

Here, we would like to compare the YOLO detection results in several cases for the 3000 m range scenario. Table 7(a), (b), and (c) show the baseline YOLO results, YOLO results using 2x super-resolution videos, and YOLO results using 4x super-resolution videos, respectively. Comparing Table 7(a) and Table 7(b), one can see than the YOLO results with 2x super-resolution videos are two times better in terms of IoU and AP. The 4x SR videos are even better. This clearly demonstrates that SR videos can help YOLO performance.

For comparing the ResNet classification results, we focus on the 3000 m videos. Table 8 (a), (b), and (c) show the results of baseline ResNet, ResNet using 2x SR videos, and ResNet using 4x SR videos, respectively. The best performing one is the 2x SR videos case. However, even in the 2x SR case, the OA and AA values are still quite low. This means that SR can only improve the performance to a certain extent. For further improvement, the camera will need to be improved.

Table 7. Comparison of YOLO results for several cases using 3000 m MWIR nighttime videos with different resolutions.

(a) YOLO baseline

3000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	1.995	100.00%	99.55%	7.09%	7.09%	49.00%
BRDM2	3.499	100.00%	99.38%	10.33%	10.33%	27.06%
BMP2	3.91	100.00%	98.48%	17.31%	17.31%	3.61%
T72	4.541	100.00%	77.02%	11.86%	11.86%	32.56%
ZSU23-4	2.18	100.00%	99.67%	11.25%	11.25%	15.67%
Avg	3.225	100.00%	94.82%	11.57%	11.57%	25.58%

(b) YOLO results using 2x VSR

3000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	4.143	100.00%	99.49%	11.93%	11.93%	19.78%
BRDM2	5.367	99.27%	93.82%	19.67%	19.67%	13.44%
BMP2	2.838	100.00%	100.00%	39.86%	39.86%	11.83%
T72	6.636	99.28%	99.28%	20.77%	20.78%	14.28%
ZSU23-4	2.512	100.00%	100.00%	22.31%	22.31%	30.72%
Avg	4.299	99.71%	98.52%	22.91%	22.91%	18.01%

(c) YOLO results using 4x VSR

3000 m	CLE	DP	EinGT	IoU	AP	% det.
BTR70	9.22	93.22%	98.18%	47.95%	48.91%	30.00%
BRDM2	6.403	99.30%	100.00%	55.49%	56.42%	54.83%
BMP2	6.481	98.67%	99.67%	40.12%	40.36%	59.61%
T72	7.221	100.00%	100.00%	33.11%	33.22%	57.61%
ZSU23-4	8.719	99.87%	100.00%	43.18%	43.45%	40.33%
Avg	7.61	98.21%	99.57%	43.97%	44.47%	48.48%

Table 8. Comparison of 3000 m ResNet confusion matrix and classification metrics for videos with different resolutions.

(a) Baseline ResNet

3000 m	5	6	9	11	12	
BTR70	1	69	271	12	540	
BRDM2	0	13	59	1	414	
BMP2	0	3	52	0	11	
T72	0	461	68	27	136	
ZSU23-4	2	56	56	29	156	
Class Stats	OA	10.22%	AA	27.53%	kappa	0.10

(b) ResNet 2x VSR

3000 m	5	6	9	11	12	
BTR70	22	21	63	8	275	
BRDM2	15	29	65	1	165	
BMP2	4	0	215	0	5	
T72	0	143	51	48	35	
ZSU23-4	0	20	74	56	484	
Class Stats	OA	44.36%	AA	41.17%	kappa	0.4436

(b) ResNet 4x VSR

3000 m	5	6	9	11	12	
BTR70	139	259	44	363	390	
BRDM2	3	260	115	28	375	
BMP2	2	38	543	265	287	
T72	3	391	16	143	52	
ZSU23-4	2	260	135	260	542	
Class Stats	OA	30.04%	AA	31.40%	kappa	0.3004

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have presented a framework for target detection and classification using long range and low quality infrared videos. The framework consists of video super-resolution using a state-of-the-art deep learning algorithm, a proven target detector using YOLO, and a customized target classifier using ResNet. The integrated framework significantly improved the target

detection and classification performance using actual infrared videos. In particular, we have observed that both the 2-times and 4-times super-resolution videos improved the YOLO detection and ResNet classification performance.

One future direction is to investigate the use of YOLO-v3, which is a new version of YOLO, for target detection. We will also investigate better training strategies to handle long range target detection and classification applications.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by the US Army under contract W909MY-20-P-0024. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

[1] C. Kwan and B. Budavari, "Enhancing Small Moving Target Detection Performance in Low Quality and Long Range Infrared Videos Using Optical Flow Techniques," Remote Sensing, 12(24), 4024, December 9, 2020.

[2] Y. Chen, G. Zhang, Y. Ma, J. U. Kang, and C. Kwan, "Small Infrared Target Detection based on Fast Adaptive Masking and Scaling with Iterative Segmentation," IEEE Geoscience and Remote Sensing Letters, January 2021.

[3] C. Kwan and B. Budavari, "A High Performance Approach to Detecting Small Targets in Long Range Low Quality Infrared Videos," arXiv:2012.02579, 2020.

[4] H. S. Demir and A. E. Cetin, "Co-difference based object tracking algorithm for infrared videos," IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 434-438

[5] C. Kwan, D. Gribben, and T. Tran, "Tracking and Classification of Multiple Human Objects Directly in Compressive Measurement Domain for Low Quality Optical Videos," IEEE Ubiquitous Computing, Electronics & Mobile Communication Conference, New York City. 2019.

[6] C. Kwan, B. Chou, J. Yang, and T. Tran, "Deep Learning based Target Tracking and Classification Directly in Compressive Measurement for Low Quality Videos," Signal & Image Processing: An International Journal (SIPIJ), November 16, 2019.

[7] C. Kwan, D. Gribben, A. Rangamani, T. Tran, J. Zhang, R. Etienne-Cummings, "Detection and Confirmation of Multiple Human Targets Using Pixel-Wise Code Aperture Measurements," J. Imaging. 6(6), 40, 2020.

[8] C. Kwan, B. Chou, J. Yang, and T. Tran, "Deep Learning based Target Tracking and Classification for Infrared Videos Using Compressive Measurements," Journal Signal and Information Processing, November 2019.

[9] C. Kwan and D. Gribben, "Target Detection and Classification Improvements using Contrast Enhanced 16-bit Infrared Videos," Signal & Image Processing: An International Journal (SIPIJ), February 28, 2021.

[10] S. Lohit, K. Kulkarni, and P. K. Turaga, "Direct inference on compressive measurements using convolutional neural networks," Int. Conference on Image Processing. 2016. 1913-1917.

[11] A. Adler, M. Elad, and M. Zibulevsky, "Compressed Learning: A Deep Neural Network Approach," arXiv:1610.09615v1 [cs.CV]. 2016.

[12] Y. Xu and K. F. Kelly, "Compressed domain image classification using a multi-rate neural network," arXiv:1901.09983 [cs.CV]. 2019.

[13] Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, and S. B. Kang, "Privacy-Preserving Action Recognition Using Coded Aperture Videos," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2019.

[14] H. Vargas, Y. Fonseca, and H. Arguello, "Object Detection on Compressive Measurements using Correlation Filters and Sparse Representation," 26th European Signal Processing Conference (EUSIPCO). 1960-1964, 2018.
[15] A. Değerli, S. Aslan, M. Yamac, B. Sankur, and M. Gabbouj, "Compressively Sensed Image Recognition," 7th European Workshop on Visual Information Processing (EUVIP), Tampere, 2018.
[16] P. Latorre-Carmona, V. J. Traver, J. S. Sánchez, and E. Tajahuerce, "Online reconstruction-free single-pixel image classification," Image and Vision Computing, 86, 2018.
[17] C. Li and W. Wang, "Detection and Tracking of Moving Targets for Thermal Infrared Video Sequences," Sensors, 18, 3944, 2018.
[18] Y. Tan, Y. Guo, and C. Gao, "Background subtraction based level sets for human segmentation in thermal infrared surveillance systems," Infrared Phys. Technol., 61: 230–240, 2013.
[19] A. Berg, J. Ahlberg, and M. Felsberg, "Channel Coded Distribution Field Tracking for Thermal Infrared Imagery," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; Las Vegas, NV, USA. pp. 1248–1256, 2016.
[20] C. Kwan, D. Gribben, B. Chou, B. Budavari, J. Larkin, A. Rangamani, T. Tran, J. Zhang, R. Etienne-Cummings, "Real-Time and Deep Learning based Vehicle Detection and Classification using Pixel-Wise Code Exposure Measurements," Electronics, June 18, 2020.
[21] MOT Challenge, motchallenge.net/
[22] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arxiv, 2018.
[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Advances in Neural Information Processing Systems, 2015.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
[25] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3224-3232), 2018.
[26] X. Xiang, "Mukosame/Zooming-Slow-Mo-CVPR-2020." GitHub, github.com/Mukosame/Zooming-Slow-Mo-CVPR-2020, 2020.
[27] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming Slow-Mo: Fast and Accurate One-Stage Space-Time Video Super-Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3370-3379, 2020.

AUTHORS

Chiman Kwan received his Ph.D. degree in electrical engineering from the University of Texas at Arlington in 1993. He has written one book, four book chapters, 15 patents, 75 invention disclosures, 380 technical papers in journals and conferences, and 550 technical reports. Over the past 25 years, he has been the PI/Program Manager of over 120 diverse projects with total funding exceeding 36 million dollars. He is also the founder and Chief Technology Officer of Signal Processing, Inc. and Applied Research LLC. He received numerous awards from IEEE, NASA, and some other agencies and has given several keynote speeches in several international conferences. He is an Associate Editor of IEEE Trans. Geoscience and Remote Sensing.

David Gribben received his B.S. in Computer Science and Physics from McDaniel College, Maryland, USA, in 2015. He is a software engineer at ARLLC. He has been involved in diverse projects, including mission planning for UAVs, target detection and classification, and remote sensing.

Bence Budavari received his B.S. in Audio Engineering from Belmont University in 2015. He is a software developer at ARLLC.