Improvement of Deep Learning-based Human Detection using Dynamic Thresholding for Intelligent Surveillance System

Wahyono[1*], Moh. Edi Wibowo[2], Ahmad Ashari[3], Muhammad Pajar Kharisma Putra[4]
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta Indonesia[1, 2, 3]
Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Lampung, Indonesia[4]

***Abstract*—Human detection plays an important role in many applications of the intelligent surveillance system (ISS), such as person re-identification, human tracking, people counting, etc. On the other hand, the use of deep learning in human detection has provided excellent accuracy. Unfortunately, the deep-learning method is sometimes unable to detect objects that are too far from the camera. It is because the threshold selection for confidence value is statically determined at the decision stage. This paper proposes a new strategy for using dynamic thresholding based on geometry in the images. The proposed method is evaluated using the dataset we created. The experiment found that the use of dynamic thresholding provides an increase in F-measure of 0.11 while reducing false positives by 0.18. This shows that the proposed strategy effectively detects human objects, which is applied to the ISS.**

***Keywords*—Human detection; YOLO; dynamic thresholding; intelligent surveillance system**

I. INTRODUCTION

Currently, the use of cameras as surveillance media is growing very fast. Usually, cameras are widely used for security purposes in public areas such as schools, offices, stations, airports, highways, and even private homes. Supported by artificial intelligence (AI) development, surveillance with cameras is no longer carried out manually by officers who have many drawbacks such as fatigue, limited staff, etc. Instead, camera-based surveillance allows it to be carried out automatically by utilizing AI-based modules, known as the Intelligent Surveillance System (ISS) [1].

Human detection plays an important role in many applications of ISS, such as person re-identification [2], human tracking [3], people counting [4], human action recognition [5], unattended baggage detection [1], etc. Even though many studies have been carried out for human detection, research on this topic still faces many challenges to overcome real problems constantly changing. One of the popular studies on human detection is the Histogram of Oriented Gradient (HOG), which was first proposed by Dalal [6]. In this study, the gradient in the image is extracted using the Sobel operator, and then histograms are formed. This method is straightforward but produces very good accuracy. Unfortunately, this method cannot handle various kinds of human poses and requires long processing times. Therefore, many new techniques have been proposed to improve HOG, such as HOG+LBP [7] for handling occlusion, efficient HOG [8], SHOG [9], Rotation-Invariant HOG [10], etc.

The development of deep learning also has a good effect on human detection research. Because of this, many methods for human detection are based on deep learning [11][12]. Martinson and Yalla proposed to use CNN for human detection in mobile robots [11]. Kim and Moon proposed to use deep convolutional neural networks (DCNNs) for human detection on Dopler radar [12]. The use of deep learning provides a significant increase in accuracy compared to handcrafted methods such as HOG. If we assume humans as objects, then many deep-learning-based object detection methods can be used to detect humans. One of the popular deep learning-based methods is YOLO [13]. YOLO produces high accuracy in detecting various kinds of objects, one of which is human objects, and also won the Real-Time Object Detection on PASCAL VOC 2007 competition. Unfortunately, the deep-learning method is sometimes unable to detect objects that are too far from the camera. Thus, this paper proposes to solve this issue.

In many cases, human detection is used at an early stage and significantly affects the accuracy of ISS applications [1-5]. If the human detection accuracy is good, these applications will also produce a good performance and vice versa. Therefore, a reliable human detection module is needed. One method that produces good accuracy is based on deep learning, namely the YOLO Network. However, selecting the confidence value threshold produced by deep learning in the decision stage is very challenging. Using a large threshold will cause human objects far from the camera not to be detected correctly. Conversely, if we use a small threshold, it will result in a lot of false positives. In this study, we propose a new strategy using dynamic thresholding based on the location of potential objects that have been detected. Thus, it is expected to be able to detect small objects while reducing false positives.

Overall, this paper provides the following major contributions: (1) Utilize deep learning for the human detection method. (2) Propose a new strategy to use dynamic thresholding in the decision stage of human detection. (3) Provide a more detailed investigation regarding the effect of threshold selection.

*Corresponding Author.

II. The Proposed Strategy

A. Data Collection

In this research, we use the open images dataset v4[1] for training purposes. This dataset contains more than 9 million images with unified annotations for image classification, object detection, and visual relationship detection [14]. We only used person images with object detection annotations. The example of the open images dataset is shown in Fig. 1. We used two CCTV videos from FMIPA Universitas Gadjah Mada, one CCTV video from traffic, and two random CCTV videos for testing purposes. The examples of a testing video are shown in Fig. 2. Table 1 shows the characteristics of video data for evaluation with the various scenario.

B. Data Labeling and Preprocessing

For testing purposes, the ground truth labeling process starts automatically and then manually verifies using labeling software LabelImg[2]. The annotation process is carried out by following procedure:

1) The automatic labeling process is repeated until the best results are obtained subjectively assessed by the annotator.

2) The image and label, which is the output of the previous process, will be loaded on the labelImg application.

3) The validation process is conducted manually by the annotator.

4) If there are wrong or missing bounding boxes, the annotator can add them.

Every second of the video will be divided into twelve images then the coordinates of the bounding box of each image will be saved in a CSV file. The example of a saved bounding box is shown in Fig. 3.

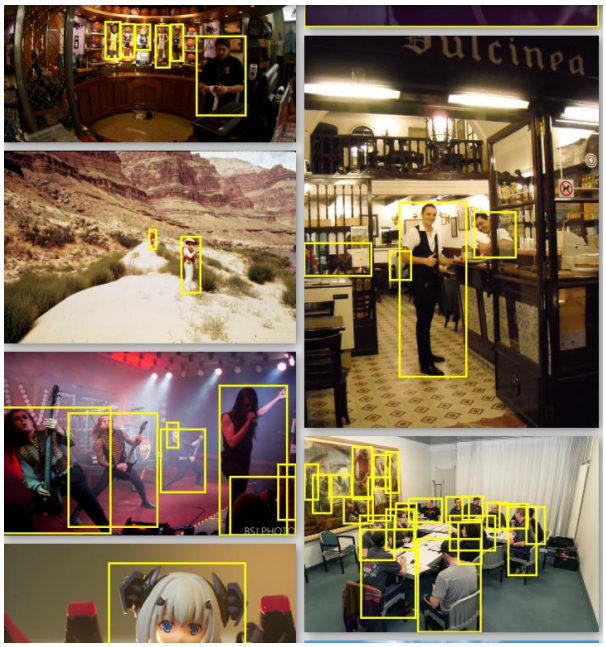


Fig. 1. Several Sample of Open Image Dataset for Training.

[1] https://opensource.google/projects/open-images-dataset.
[2] https://github.com/tzutalin/labelImg.

Fig. 2. Sample Image of Testing Data.

TABLE I. Testing Video Data Characteristics

No	Dataset	Duration	FPS	Scenario
1	MIPA 1	01:14	30	Indoor
2	MIPA 2	01:12	30	Indoor
3	MIPA 3	00:28	30	Indoor
4	Office	00:12	30	Indoor
5	Traffic	00:12	30	Outdoor
6	Kitchen	00:12	30	Indoor
7	School	00:13	30	Outdoor

```
frame,x1,y1,x2,y2
1.0,740,482,818,664
1.0,878,465,942,626
1.0,977,453,1047,612
1.0,814,451,877,612
1.0,393,269,414,336
1.0,365,277,384,331
1.0,518,316,546,413
2.0,740,485,822,666
```

Fig. 3. Saved Bounding Box for Testing Data.

C. Detection Strategy using the Dynamic Thresholding

As shown in Fig. 4, the proposed method starts with extracting human candidate regions using a deep learning model and is then followed by the validation stage using thresholding. The basic method we use in detecting humans is by using YOLO. YOLO network will generate candidate regions of objects with certain confidence values. If the confidence value exceeds the threshold, we will classify this object as human and vice versa. However, selecting the threshold of confidence is very challenging. Using a large threshold will cause human objects far from the camera not to be detected properly. Conversely, if we use a small threshold, it will result in a lot of false positives. To solve this problem, we propose using dynamic thresholding for verifying the human region based on the object's position in the vertical direction, as shown in Fig. 5. The closer the object's position to the top, the smaller the object's threshold score will be and vice versa. This strategy is applied with the assumption that objects close to the top are far from the camera. Objects far from the camera will have faint details, so we use a small threshold so that the object can still be detected. However, if this small threshold is applied to objects close to the camera, it can be false positive.

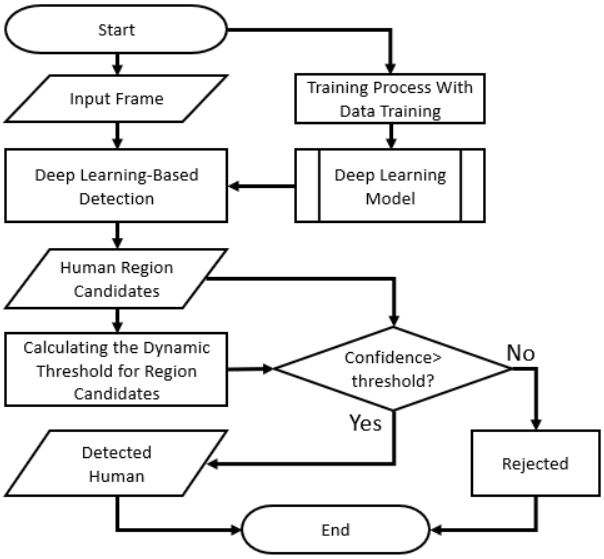


Fig. 4. The Flowchart of the Proposed Method.

Fig. 5. Ilustration of Calculating the Object Vertical Position.

We determine the threshold with the following strategy. First, we scale the object's vertical position by dividing the object's vertical center point by the image height (H). This value will be used to obtain a threshold on a scale of 0.5 to 1 using the following equation:

$$y' = \frac{y}{H} \quad (1)$$

$$t = y' \times (t_{max} - t_{min}) + t_{min} \quad (2)$$

III. Experiment and Results

This section presents the evaluation protocol used in the experiment and the result of the proposed method with a comparison to the YOLO method. In addition, a discussion of the effectiveness of the proposed method is also presented.

A. Evaluation Protocol

We will compare performance and processing time between the basic YOLO method and the YOLO + dynamic threshold in the testing stage. Thus, it can be seen whether the dynamic threshold process will significantly affect the accuracy and processing time. We use hardware with the following specifications: Processor AMD Ryzen 9 3900x, VGA Nvidia RTX 2080 Ti, RAM 64 Gb with Ubuntu 20.04 operating system.

There are seven videos from different CCTV as test objects. The evaluation process will calculate the IoU value between the detected bounding box and ground truth for each CCTV video frame. Intersection over Union (IoU) is the value based on the statistical similarity and diversity of the sample set whose purpose is to evaluate the area of overlap between the two bounding boxes, namely the predicted bounding box and the ground truth bounding box [15]. IoU can be found using the following equation:

$$IoU = \frac{A \cap B}{A \cup B} \quad (3)$$

We use 0.5 as a threshold for the IoU value. That means any detected object with an IoU value greater than the threshold will be considered true positives (TP). In contrast, if the IoU value is less than the threshold, it will be considered as false positive (FP), and any undetected object will be considered as false negative (FN). To evaluate the method's performance, we use precision, recall, and f-measure [16]. These metrics are used to evaluate the prediction of each frame. To evaluate the model, the precision, recall, and f-measure of all the frames are calculated.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$FMeasure = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

B. Comparison Results

Based on the testing results using seven different datasets, the following results were obtained in Table 2. It can be seen that in all datasets, the proposed strategy achieves 0.89, 0.95, and 0.91 in precision, recall, and f-measure, respectively, for detecting the human object on the video. These results are better than the YOLO method, which only archives 0.71, 0.94, and 0.80 in precision, recall, and f-measure, respectively. Furthermore, applying our strategy could increase the f-measure of human detection by around 0.11. Thus, it proved that the proposed approach is effective for increasing accuracy and reducing the false positive.

C. Discussion

In general, for each data test, it can be seen that the recall value tends to be higher than the precision value for both YOLO and the proposed method. This indicates that both methods are weak against false positives or often detect other objects as human objects. Some examples of test results for each datatest are shown in the image below. The blue box shows the results of YOLO detection, while the red boxes show the results of YOLO detection and the proposed method.

TABLE II. COMPARISON RESULTS THE PROPOSED METHOD AND YOLO

No	Datatest	Method					
		YOLO [13]			Proposed Method		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	MIPA 1	0.67	0.92	0.78	0.83	0.91	0.87
2	MIPA 2	0.73	0.97	0.83	0.95	0.99	0.97
3	MIPA 3	0.84	1.00	0.91	0.95	1.00	0.97
4	Office	0.56	0.92	0.70	0.81	0.88	0.84
5	Traffic	0.47	0.87	0.62	0.81	0.86	0.84
6	Kitchen	0.98	1.00	0.98	0.99	1.00	0.99
7	School	0.72	0.92	0.81	0.87	0.93	0.90
AVERAGE		**0.71**	**0.94**	**0.80**	**0.89**	**0.95**	**0.91**

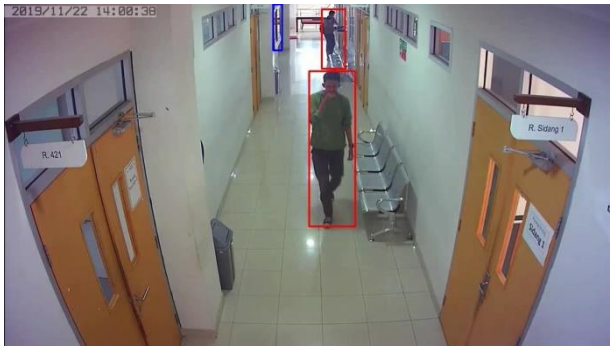
In video 1, the proposed method almost always succeeds in detecting human objects without errors. Still, there are false positives in the YOLO method by detecting the announcement box as a human object, as shown in Fig. 6(a). According to the YOLO network, the announcement box has a confidence value of 0.3, as it is close to the top area. The proposed dynamic thresholding process has eliminated the announcement box because it has a confidence value that does not meet the threshold. Same as in the first video, in video 2, the YOLO method is still wrong in detecting the announcement box as a human object, as shown in Fig. 6(b). However, it can be seen if the human object in front of the announcement box can be detected correctly by both methods.

In video 3, although the camera angle is the same as videos 1 and 2, there is a false positive for another object, namely the trash box, when detected using the YOLO method, as shown in Fig. 6(c). While in the proposed method, there are no errors. In video 4, there are false negatives if detected using the proposed method, as shown in Fig. 6(d). False negatives occur when there is an occlusion of a human object. Occlusion makes the object's confidence value low due to the lack of features that can be extracted due to some features covered by other objects. There were many false positives when testing on video 5 using the YOLO method by detecting road cones as humans, as shown in Fig. 6(e)-(g). At the same time, the proposed method is still able to detect well every human object.

Both YOLO and the proposed method can detect human objects well because the distance of the camera to the object is still quite ideal so that the details of the object are quite clear and there is no occlusion on the object. Therefore, it can be concluded that the proposed method has better performance than the YOLO method, as shown in Table 2. In the example of the detection results, it can also be seen that there are errors in the YOLO method in detecting objects. Overall, both YOLO and the proposed method can almost always detect human objects but still often detect other objects as human objects. This is indicated by the recall value, which is higher than the precision value.

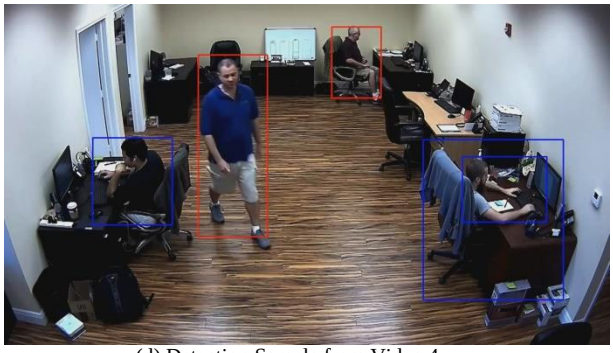
Nevertheless, the proposed method fails to detect objects that are very far from the camera because the details of the object are not clear, so that there are not many features that can be extracted. It makes the object's confidence value below the threshold. To solve this problem, we may improve the selection of dynamic threshold by considering the distance between the candidate object and the vanishing point. In this case, we should integrate the vanishing point detection [17] [18]. Another solution is by utilizing the super-resolution method for very far objects [19]. However, this solution may require a long processing time.

(a) Detection Sample from Video 1.



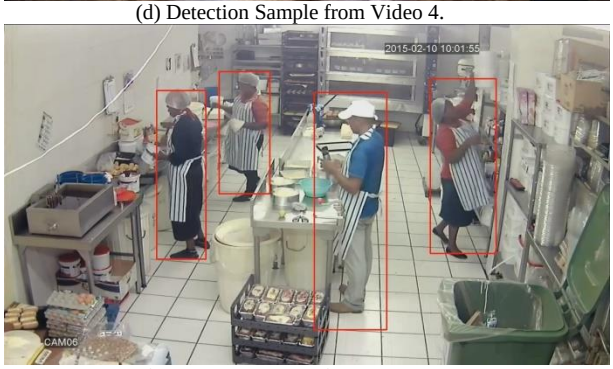
(b) Detection Sample from video 2.

(c) Detection Sample from Video 3.

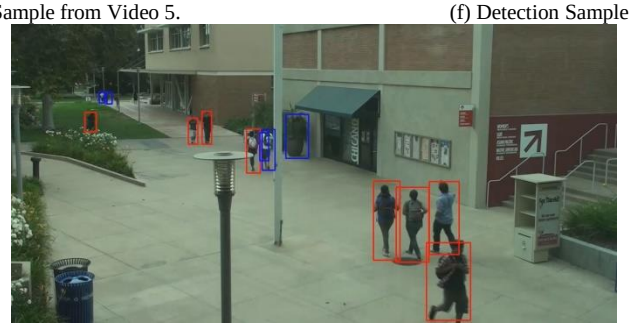


(d) Detection Sample from Video 4.

(e) Detection Sample from Video 5.



(f) Detection Sample from Video 6.



(g) Detection Sample from Video 7.

Fig. 6. Detected Sample for Video 1 until Video 7 with various conditions.

IV. Conclusion

The use of deep learning in human detection provides fairly good accuracy. However, this result is still influenced by selecting the threshold for the confidence value in the decision stage. The use of a static threshold is still not optimal in detecting objects that are far from the camera. This paper has succeeded in proposing the use of a dynamic threshold which is proven to provide a fairly good increase in f-measure, which is around 11% compared to the use of YOLO without a dynamic threshold. It should be noted that the use of dynamic thresholding can be used not only in YOLO but also in other deep-learning architectures. Even so, the dynamic threshold is still possible to be improved by considering the vanishing point in the image or super-resolution image [20].

V. Acknowledgment

This research was supported by 2021 Penelitian Dasar Unggulan Perguruan Tinggi-PDUPT (*College Excellence Basic Research*), funded by the Ministry of Education, Culture, Research and Technology, the Republic of Indonesia with Grant Number 6/E1/KP.PTNBH/2021 and 1691/UN1/DITLIT/DIT-LIT/PT/2021.

References

[1] Wahyono, A Filonenko, KH Jo, "Unattended Object Identification for Intelligent Surveillance Systems Using Sequence of Dual Background Difference", IEEE Transactions on Industrial Informatics vol. 12, no. 6, 2247-2255, 2016, doi: 10.1109/TII.2016.2605582.

[2] M. P. Kharisma and Wahyono, "A Novel Method for Handling Partial Occlusion on Person Re-Identification Using Partial Siamese Network", International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, pp. 313-321, 2021, doi: 10.14569/IJACSA.2021.0120735.

[3] E.U. Haq, H. Jianjun, K. Li, and H.U Haq, "Human detection and tracking with deep convolutional neural networks under the constrained of noise and occluded scenes", Multimedia Tools and Applications, vol. 79, pp. 30685–30708, 2020, doi: 10.1007/s11042-020-09579-x.

[4] M. Padmashini, R. Manjusha, L. Parameswaran, "Vision Based Algorithm for People Counting Using Deep Learning", International Journal of Engineering and Technology, vol. 7, no. 3, 2018, doi: 10.14419/ijet.v7i3.6.14942.

[5] N. A. Simanjuntak, J. Hendarto, and Wahyono, "The Effect of Image Preprocessing Techniques on Convolutional Neural Network-Based Human Action Recognition", Journal of Theoretical and Applied Information Technology, vol. 98, no. 16, pp. 3364-3374, 2020.

[6] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection”, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 20-25 Juni 2015, doi: 10.1109/CVPR.2005.177.

[7] X. Wang, T.X. Han, S. Yan, "An HOG-LBP human detector with partial occlusion handling", 2009 IEEE 12th International Conference on Computer Vision, doi: 10.1109/ICCV.2009.5459207.

[8] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient HOG human detection", Signal Processing, vol. 91, no. 4, April 2011, pp. 773-781, doi: 10.1016/j.sigpro.2010.08.010.

[9] H. Skibbe, M. Reisert, and H. Burkhardt, "SHOG - Spherical HOG Descriptors for Rotation Invariant 3D Object Detection", Mester R., Felsberg M. (eds) Pattern Recognition. DAGM 2011. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol 6835, pp. 142-151, 2011, doi: 10.1007/978-3-642-23123-0_15.

[10] K Liu, H Skibbe, T Schmidt, T Blein, K Palme, T Brox, O Ronneberger, "Rotation-invariant HOG descriptors using Fourier analysis in polar and spherical coordinates", International Journal of Computer Vision vol. 106, no. 3, pp. 342-364, 2014.

[11] E. Martinson, V. Yalla, "Real-time human detection for robots using CNN with a feature-based layered pre-filter", 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 26-31 Aug. 2016, doi:10.1109/ROMAN.2016.7745248.

[12] Y. Kim, T. Moon, "Human Detection and Activity Classification Based on Micro-Doppler Signatures Using Deep Convolutional Neural Networks", IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 1, pp. 8-12, doi: 10.1109/LGRS.2015.2491329.

[13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", Conference on Computer Vision and Pattern Recognition, 27-30 June 2016, doi: 10.1109/CVPR.2016.91.

[14] A. Kuznetsova, et al., "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale", International Journal of Computer Vision vol. 128, pp.1956–1981, 2020, 10.1007/s11263-020-01316-z.

[15] H. Rezatofighi, et al., "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), doi: 10.1109/CVPR.2019.00075.

[16] C. Goutte, E. Gaussier, Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Losada D.E., Fernández-Luna J.M. (eds) Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, vol 3408. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25.

[17] J. Kim, "Efficient Vanishing Point Detection for Driving Assistance Based on Visual Saliency Map and Image Segmentation from a Vehicle Black-Box Camera", Symmetry, vol. 11, no. 12, pp.1492, 2019; doi: 10.3390/sym11121492.

[18] A. Tai, J. Kittler, M. Petrou, and T. Windeatt, Vanishing point detection, Image and Vision Computing, vol. 11, no. 4, May 1993, pp.240-245, doi: 10.1016/0262-8856(93)90042-F.

[19] S.-J. Park, H. Son, S.Cho, K.-S. Hong, and S. Lee, "SRFeat: Single Image Super-Resolution with Feature Discrimination", The 2018 European Conference on Computer Vision, doi: 10.1007/978-3-030-01270-0_27.

[20] Z. Wang, J. Chen, and S.C.H. Hoi, Deep Learning for Image Super-Resolution: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, pp. 3365-3387, Oct 2021, doi: 10.1109/TPAMI.2020.2982166.