To Detect, and Beyond: Integrating Text-Guided Object Detection and Super-Resolution

Hyun Jun Yook, Hee Jae Chung, Yun Chang Choi, Su Yeon Kim, Tae Hyung Kim, and Youn Kyu Lee*
Department of Computer Engineering
Hongik University
Seoul, Republic of Korea
hyunjunyook@g.hongik.ac.kr, clover@g.hongik.ac.kr, bambini77@g.hongik.ac.kr, suyeonkim@g.hongik.ac.kr,
taehyung@hongik.ac.kr, younkyul@hongik.ac.kr

***Abstract*—Object detection in conventional surveillance systems is typically constrained to predefined objects. Furthermore, the low resolution of the video often impairs the identifiability of these detected objects. In this paper, we propose a novel framework that not only detects unseen objects but also enhances the identifiability of detected objects. The proposed framework introduces text guidance into object detection and super-resolution processes, enabling the detection of untrained objects and restoring them to high resolution while preserving their details. Moreover, text-guided object detection and super-resolution are synergistically integrated through the shared text encoding between the two processes. Our evaluation on a real world dataset demonstrates the effectiveness of the proposed framework in minimizing distortion of detected objects and enhancing identifiability by fusing information from different modalities of image and text during restoration.**

***Index Terms*—object detection, super-resolution, text-guidance, surveillance system, multi-modal learning**

I. Introduction

Object detection is a technique for recognizing and categorizing objects in images which is widely employed in surveillance systems (e.g., CCTV, IP Camera) [1]. However, since the object detection is generally limited to detecting predefined objects, the applicability to surveillance systems can be limited [2]. Moreover, additional manual inspection may be necessary in scenarios where, for example, a part of the object is occluded [3]. However, when the resolution of the image is low or interference (e.g., external lighting, reflections, etc.) occurs, recognizing the detected object can be challenging for humans, which reduces the efficiency of the surveillance systems [4]–[6].

Text-guided object detection (*TGOD*) has been proposed to detect untrained (Unseen) objects by providing information about the object through text prompts, but if the detected object is low resolution and too small, it is challenging for humans to recognize it accurately [5]–[7]. To address this challenge, text-guided super-resolution (*TGSR*) has recently been proposed to restore low resolution images to high resolution by providing details of the object as text [8]–[10]. However, it requires extra text encoder to extract text embeddings and utilizing generative models (i.e., diffusion-based model) may potentially lead to not realistic results or create artifacts [9], [10].

*Corresponding Author

In this paper, we propose a novel framework that automatically detects unseen objects by using text information and restores the detected objects in high resolution while preserving the details of the objects. The proposed framework detects objects using *TGOD*, and the detected objects are restored in high resolution using *TGSR*. However, given the applicability and efficiency of surveillance systems, since different text embeddings require different text encoders, simple integration of existing *TGOD* and *TGSR* is computationally inefficient, and the detected objects may be distorted during the super-resolution process. In the procedure, we introduce an *Adapter* comprising a few trainable parameters, enabling *TGSR* to leverage the text embeddings of pre-trained *TGOD* [11], [12]. *Adapter* effectively fuses the text embeddings utilized by *TGOD* for object detection with the visual feature maps that *TGSR* restores at high resolution. As a result, the detected objects can be restored to high resolution without distortion by preserving the details of the objects. We evaluated the effectiveness of the proposed framework on the COCO dataset [13], a representative benchmark for object detection, and confirmed both quantitatively and qualitatively the improvement in the quality of the restored images.

The contributions of this paper are as follows:

1) Proposing a novel framework that simultaneously provides object detection and super-resolution based on textual information.
2) Proposing a novel *Adapter* mechanism to effectively transfer text embeddings from *TGOD* to *TGSR*.
3) Validating the proposed framework through evaluation on a real world dataset.

This paper is organized as follows. Section II describes related work. Section III introduces the proposed method and Section IV presents the experimental settings and results. Finally, Section V provides conclusion including future work.

II. Related Work

A. Object Detection Methods

Object detection is a technique for detecting specific objects (e.g., people, animals) in an image or video [14], [15]. *Ren et al.* [16] proposed Faster R-CNN which improved detection speed and accuracy by introducing the Region

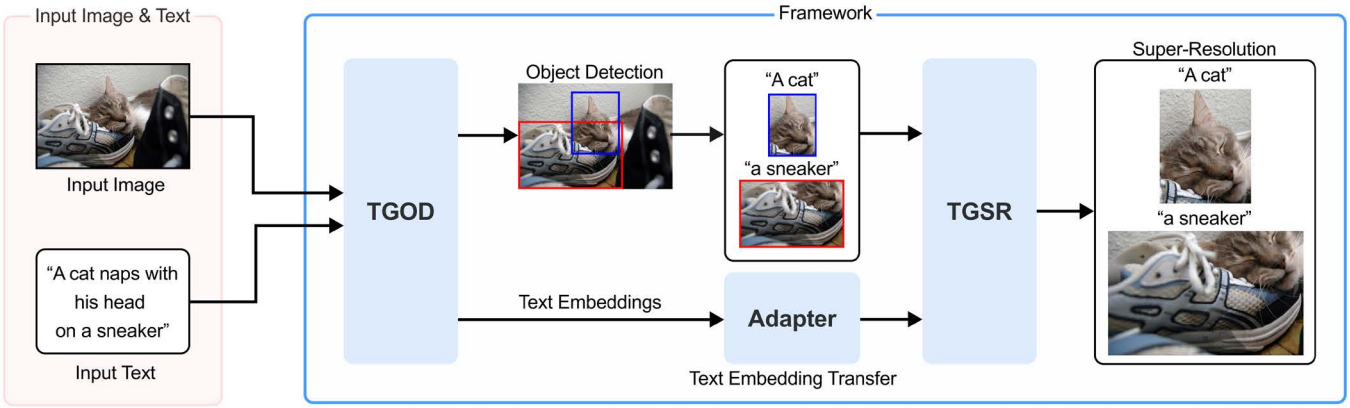

Fig. 1. Overview of the proposed framework.

Proposal Network to address the problem of slow region proposal generation. *Redmon et al.* [17] proposed YOLO which improved detection speed and accuracy by predicting the location and class probability of objects within each grid cell of an image. However, these methods had limitations in accurately detecting unseen objects and small objects [18], [19]. Recently, *TGOD* has been proposed to detect unseen objects in images by introducing text guidance [20], [21]. *Li et al.* [22] proposed GLIP which improved zero-shot detection accuracy by addressing the issue of requiring human annotations, based on the connectivity between detected objects and their textual representations. *Yao et al.* [23] proposed DetCLIP which improved the accuracy of zero-shot detection by minimizing unnecessary interactions between classes and noun phrases in captions by matching the class of an object, images, and a caption for that image. *Liu et al.* [24] proposed Grounding DINO which introduced cross-attention modules of image-to-text and text-to-image, and improved detection accuracy for unseen objects by learning sub-sentence level representations. However, in surveillance systems, when existing object detection methods are applied without additional image processing, factors such as small object size, low image resolution, or external interferences like lighting and reflections can make it difficult for humans to recognize the detected object [5]–[7].

B. Super-Resolution Methods

Super-resolution is a technique for restoring low resolution images to high resolution images [25]. *Dong et al.* [26] proposed SRCNN which is a lightweight structure to learn an end-to-end mapping between high resolution and low resolution images, improving processing speed with little pre/post processing. *Kim et al.* [27] proposed VDSR which addresses the vanishing gradient problem in very deep networks, enabling faster optimization by utilizing residual learning and extremely high learning rates. However, these methods are limited by the difficulty of restoring details and photorealistic visual quality at high upscaling factors (e.g., $\times 8$) [8]. Recently, *TGSR* has been proposed to preserve details of objects by introducing text guidance [20], [21]. *Gandikota et al.* [9] proposed zero-shot open domain image super-resolution which preserves image details by modifying the generation process of text-to-image diffusion models. *Li et al.* [10] proposed ESTGN which better preserves the semantic information of images even at large upscaling factors by automatically extracting task-relevant text [9], [10]. However, existing *TGSR* methods may require additional models to extract text embeddings [10]. In addition, restored images generated by diffusion models can potentially exhibit not realistic features and undesired artifacts [9].

III. Proposed Method

In this paper, to address the challenges of applicability and efficiency of surveillance systems, we propose a novel framework that provides the detection of unseen objects using *TGOD* and the restoration of the identifiability of detected objects through *TGSR*. As shown in Fig. 1, the proposed framework consists of *TGOD*, *Adapter*, and *TGSR*. *TGOD* can detect unseen objects in the target image through the provided input text. In this procedure, the text embeddings extracted through the text encoder of *TGOD* are effectively fused with the visual feature maps of *TGSR* for high resolution restoration of the detected objects through *Adapter* module. This fusion enables the restoration of the detected objects to high resolution without distortion while preserving detailed information about the objects. The detailed explanation of the framework is as follows.

A. Text-Guided Object Detection (TGOD)

TGOD is a multi-modal module that utilizes input images and texts to detect target objects in the input images. *TGOD* detects the objects in the input image that are most relevant to the input text, which is designed based on Grounding DINO [24], a state-of-the-art *TGOD* method trained through a cross-attention mechanism. *TGOD* detects unseen objects and forwards the extracted text embeddings to *Adapter*.

B. Adapter

Adapter is a module that transfers text embeddings from *TGOD* to *TGSR* [28]. *Adapter* consists of a few trainable

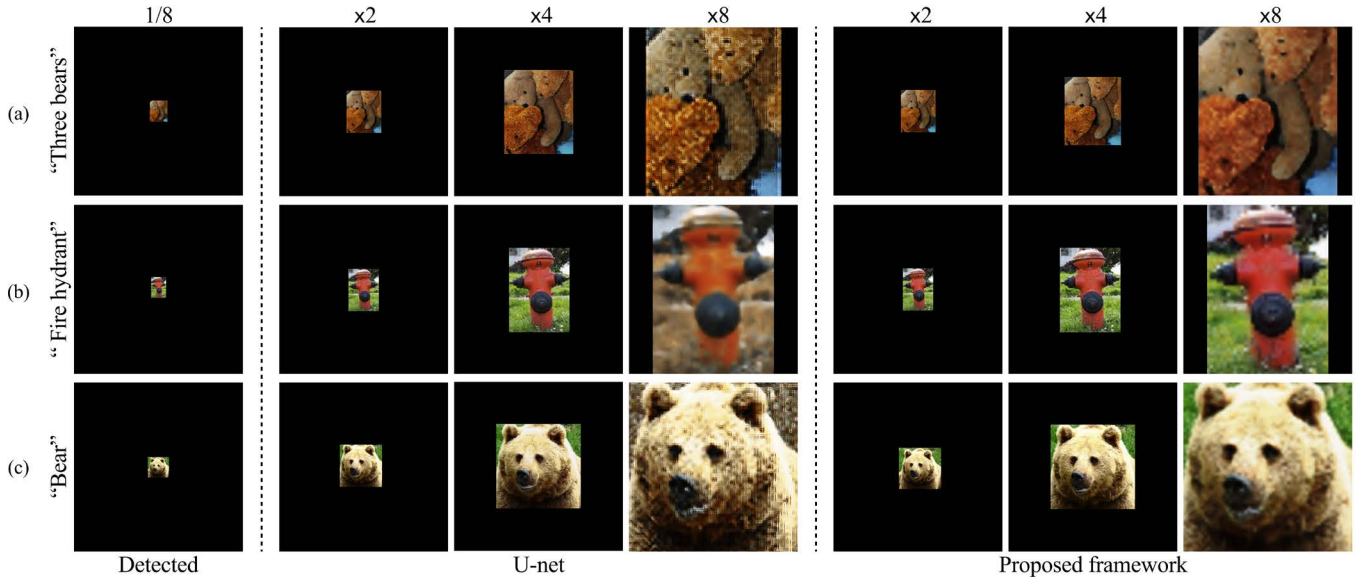

Fig. 2. Qualitative visual comparison of the proposed framework and U-net in super-resolution.

parameters that transform the one-dimensional text embeddings extracted to guide object detection into two-dimensional vectors for fusing the visual feature maps of *TGSR*. The text embeddings transformed into two-dimensional vectors can provide *TGSR* with text embeddings without requiring a separate text encoder, which enables *TGSR* to utilize detailed shared textual information about the object.

C. Text-Guided Super-Resolution (TGSR)

TGSR is a module that preserves the details of detected objects while restoring them to high resolution without distortion. *TGSR* is constructed based on a U-net architecture, comprising convolutional layers organized within an encoder-decoder structure, and skip connections. The encoder extracts visual feature maps from the detected objects, which the decoder subsequently restores to high resolution [29]. The skip connection transfers the visual feature maps extracted from each convolutional layer of the encoder to their corresponding decoder layers. Each convolutional layer of the encoder performs a 1/2 downsampling in the process of extracting the visual feature maps, whereas the decoder conversely upsamples by $\times 2$. Therefore, for super-resolution, the decoder requires more layers than the encoder, with each additional layer doubling the resolution relative to the detected object. However, given that detailed information may be lost during the downsampling process in the encoder, the proposed *TGSR* integrates with *Adapter*'s two-dimensional text embeddings to the encoder and each visual feature map of additional decoder layer [30]. Even decoder layers that are not directly integrated with the skip connections can receive detailed information about the object. Throughout this procedure, the proposed *TGSR* effectively preserves object details during the high resolution process.

IV. Evaluation

To evaluate the effectiveness of the proposed framework, we conducted evaluations based on the following research questions:

- **RQ#1**: Does the proposed framework provide identifiable levels of high resolution restoration for surveillance systems?
- **RQ#2**: Does *Adapter* in the proposed framework improve the performance of high resolution restoration?

A. Experimental Settings

To evaluate the effectiveness of the proposed framework, we selected the COCO 2017 dataset [13], which consists of a variety of images and their corresponding captions (Training images: 118,000 and Validation images: 5,000). Specifically, we selected 1,000 and 200 images from the training and validation images of the COCO 2017 dataset as the training and testing set of this experiment, respectively, to include small objects (i.e., 32×32 or smaller in size) that are typically detected in surveillance systems [31]. Since each image contains five captions, one of the five captions was randomly selected for each epoch in the training phase, and one caption was fixed for the evaluation phase to ensure a fair comparison.

The experiments in this paper were performed using an NVIDIA GeForce RTX 3090 GPU, Python 3.8.10, and PyTorch 1.12.1, with the following hyperparameters: optimizer=AdamW, learning rate=1e-4, epochs=100, and batch size=1.

B. Experimental Results

(RQ#1) Effectiveness of the proposed framework in super-resolution: To evaluate RQ#1, we conducted a qualitative comparison of restored images between the proposed

TABLE I
QUANTITATIVE EVALUATION OF THE SUPER-RESOLUTION TASK BETWEEN THE PROPOSED FRAMEWORK AND A METHOD WITHOUT *Adapter*.

Methods	x2		x4		x8	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Proposed framework (W/ *Adapter*)	30.08	**0.90**	**25.61**	**0.86**	**20.45**	**0.61**
W/O *Adapter*	**30.17**	**0.90**	23.31	0.69	18.48	0.48
Resize	26.25	0.82	20.52	0.62	17.87	0.44

framework and a conventional super-resolution method (U-net). In Fig. 2, each row represents the restored images to $\times 2$, $\times 4$, and $\times 8$ using U-net and the proposed framework, respectively. Each test image is of a detected unseen object using *TGOD* with text in the proposed framework [24]. All detected objects are small in size, which makes it difficult for humans to identify them. When the proposed framework and U-net restore to $\times 2$ and $\times 4$, respectively, the restored images are comparable in terms of quality. However, when restored to $\times 8$ which is the relatively high upscaling factor, the results are noticeable. Specifically, the restored image using the proposed framework preserves the original colors and details with less distortions. Conversely, the image restored using U-net is not only inconsistent with the original colors but also exhibits certain distortion that can interfere with identification. These results demonstrate that the proposed framework can preserve the details of detected objects to provide improved identification.

(RQ#2) Effectiveness of *Adapter* in the proposed framework: To evaluate RQ#2, we conducted a quantitative comparison between the proposed framework with *Adapter* module ("Proposed framework") and without *Adapter* module ("W/O *Adapter*"). As a baseline, we also present the performance of simple image resizing through bilinear interpolation. In Table I, each column represents the average PSNR and SSIM performance when restoring the detected objects at $\times 2$, $\times 4$, and $\times 8$ for each method. For $\times 2$, the PSNR of the proposed framework and W/O *Adapter* achieved 30.08 dB and 30.17 dB, respectively, which is slightly lower, while SSIM achieved the same performance of 0.90. For $\times 4$, the PSNR of the proposed framework and W/O *Adapter* are 25.61 dB and 23.31 dB, and the SSIM is 0.86 and 0.69, respectively, achieving 2.30 and 0.17 higher performance for the PSNR and SSIM. For $\times 8$, the PSNR of the proposed framework and W/O *Adapter* are 20.45 dB and 18.48 dB, and the SSIM is 0.61 and 0.48, respectively, achieving 1.97 and 0.13 higher performance for PSNR and SSIM. Moreover, compared to the baseline, the proposed framework achieves 2.58 to 5.09 dB higher performance for PSNR and 0.08 to 0.24 higher performance for SSIM. These results demonstrate that the proposed framework not only provides superior performance compared to the baseline, but also effectively transfers text embeddings to *TGSR* through *Adapter* to preserve the restoration performance especially for high upscaling factors.

V. CONCLUSION

In this paper, we propose a novel framework that provides unseen object detection and high resolution restoration of detected objects to enhance the applicability and efficiency of surveillance systems. The proposed framework detects the objects in the input images that are most relevant to the input texts through *TGOD*, and the extracted text embeddings are transferred through *Adapter*. The transformed text embeddings enable *TGSR* to restore the detected low resolution objects in high resolution while preserving their details. The experimental results demonstrate that the proposed framework provides unseen object detection and identifiable high resolution restoration. Moreover, by effectively fusing different modalities of images and texts, it can minimize the distortion of the restored images while significantly improving their quality. Our future work includes providing the improved identifiability of detected objects in environments of varying brightness, and investigating mechanisms to achieve higher upscaling factors (i.e., $\times 16$, $\times 32$). We also plan to evaluate the applicability of the proposed framework in real surveillance systems.

ACKNOWLEDGMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1F1A1074786).

REFERENCES

[1] C. Tang, Y. Feng, X. Yang, C. Zheng, and Y. Zhou, "The object detection based on deep learning," in *Proceedings of the International Conference on Information Science and Control Engineering*, pp. 723–728, 2017.

[2] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024.

[3] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[4] S. Fahad, S. U. Rahman, F. Alam, M. Yousaf, F. S. Alamri, N. Abbas, S. A. Bahaj, and A. R. Khan, "A modified singular value decomposition (msvd) approach for the enhancement of cctv low-quality images," *IEEE Access*, vol. 12, pp. 20138–20151, 2024.

[5] J. Shao, F. Chao, M. Luo, and J. C. Lin, "A super-resolution reconstruction algorithm for surveillance video," *Journal of Forensic Science and Medicine*, vol. 3, no. 1, pp. 26–30, 2017.

[6] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4792–4800, 2016.

[7] G. Wang, H. Ding, M. Duan, Y. Pu, Z. Yang, and H. Li, "Fighting against terrorism: A real-time cctv autonomous weapons detection based on improved yolo v4," *Digital Signal Processing*, vol. 132, p. 103790, 2023.

[8] C. Ma, B. Yan, Q. Lin, W. Tan, and S. Chen, "Rethinking super-resolution as text-guided details generation," in *Proceedings of the Association for Computing Machinery International Conference on Multimedia*, pp. 3461–3469, 2022.

[9] K. V. Gandikota and P. Chandramouli, "Text-guided explorable image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25900–25911, 2024.

[10] Q. Li, Z. Ying, D. Pan, Z. Fan, and P. Shi, "Estgn: Enhanced self-mined text guided super-resolution network for superior image super resolution," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3655–3659, 2024.

[11] J. Liu, Y. Wang, C. Ju, C. Ma, Y. Zhang, and W. Xie, "Annotation-free audio-visual segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5604–5614, 2024.
[12] Z. Yu, Y. Qiao, Y. Xie, and Q. Wu, "Multi-modal adapter for medical vision-and-language learning," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging*, pp. 393–402, 2024.
[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.
[14] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.
[15] H. Yoo, P. M. Hong, T. Kim, J. W. Yoon, and Y. K. Lee, "Defending against adversarial fingerprint attacks based on deep image prior," *IEEE Access*, vol. 11, pp. 78713–78725, 2023.
[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.
[18] S.-J. Ji, Q.-H. Ling, and F. Han, "An improved algorithm for small object detection based on yolo v4 and multi-scale contextual information," *Computers and Electrical Engineering*, vol. 105, p. 108490, 2023.
[19] N. Hayat, M. Hayat, S. Rahman, S. Khan, S. W. Zamir, and F. S. Khan, "Synthesizing the unseen for zero-shot object detection," in *Proceedings of the Asian Conference on Computer Vision*, pp. 155–170, 2020.
[20] J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee, "Small object detection using context and attention," in *Proceedings of the International Conference on Artificial Intelligence in Information and Communication*, pp. 181–186, 2021.
[21] R. Shen, N. Inoue, and K. Shinoda, "Text-guided object detector for multi-modal video question answering," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1032–1042, 2023.
[22] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.
[23] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. XU, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," in *Proceedings of the Neural Information Processing Systems*, vol. 35, pp. 9125–9138, 2022.
[24] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv*, 2023.
[25] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *Association for Computing Machinery Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.
[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.
[27] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.
[28] S. Park, S. B. Son, Y. K. Lee, S. Jung, and J. Kim, "Two-stage architectural fine-tuning for neural architecture search in efficient transfer learning," *Electronics Letters*, vol. 59, no. 24, p. e13066, 2023.
[29] S.-T. Tran, C.-H. Cheng, T.-T. Nguyen, M.-H. Le, and D.-G. Liu, "Tmd-unet: Triple-unet with multi-scale input features and dense skip connection for medical image segmentation," *Healthcare*, vol. 9, no. 1, p. 54, 2021.
[30] Y. Wang, L. Wang, H. Wang, and P. Li, "Information-compensated downsampling for image super-resolution," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 685–689, 2018.
[31] A. R. P. Gondosiswojo and G. P. Kusuma, "Low resolution face recognition on cctv images using a combination of super resolution and face recognition models," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 20, 2023.