2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

TVFusionGAN: Thermal-Visible Image Fusion Based on Multi-level Adversarial Network Strategy

Guoyu Lu

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611671

Abstract— **Thermal imaging is effective in low-light or nighttime conditions due to its ability to capture thermal radiation differences, but lacks texture compared to visible images. Conversely, visible images retain more texture information, particularly during the daytime, but perform poorly at night. To address the limitations of both modalities, recent methods have utilized fusion techniques to generate images that combine thermal and visible properties. This paper presents an end-to-end fusion network leveraging generative adversarial networks (GANs) to fuse salient components from both modalities. Our network includes a generator and two discriminators. The generator produces fusion images with salient objects using a specially designed CIoU loss, while the discriminators ensure that the fused images are salient at both holistic and local scales. One discriminator encourages the fused images to resemble visible images overall, while the other ensures that targeted objects in the fused images are as salient as in thermal images. Our method effectively preserves thermal radiation of salient objects in infrared images while incorporating the textures of visible images.**

I. Introduction

Fusion aims to combine salient features from different modalities' images, resulting in a single image that retains the strengths of both. Thermal images, also known as long-wave infrared images, capture objects via thermal radiation, while visible images provide texture and intensity information. The fused result offers a comprehensive and clear depiction by leveraging the complementarity of the modalities. The key challenge in fusion is to extract effective salient features from different image types and merge them into a single image. Various fusion methods have been developed, including multi-scale transform [34], [20], [21], non-multi-scale transform [16], [49], [3], [15], sparse representation [22], [52], [45], and saliency-based methods [53], [11]. These methods focus on manual feature extraction and fusion rules for improved performance. However, as fusion quality requirements advance, the complexity of fusion rules and feature extraction methods increases, posing limitations in terms of computational cost and implementation difficulty.

Early learning-based approaches focused on feature extraction, while traditional fusion rules were used for the fusion process [24], [19]. Manual fusion rules may overlook salient features, degrading the quality of the fused image. To address this, fusion methods based on regular GANs and their variants have been proposed to overcome the lack of ground truth. However, these GAN-based methods tend to make the fused image resemble one source image, leading to the loss of critical information from the other source images.

Guoyu Lu is with the Intelligent Vision and Sensing Lab at the University of Georgia, USA. guoyu.lu@uga.edu

Motivated by recent advancements in GAN-based image fusion techniques [30], [29], we introduce a novel approach for fusing thermal and visible images using a dual-discriminator least-squares generative adversarial network (GAN) [54]. Our method aims to seamlessly combine the thermal pixel intensities of target objects with the holistic visible appearance and textures. The generator of our network is tasked with producing fused images that capture both the thermal radiation characteristics of objects and the detailed textures present in visible images, leveraging the Complete Intersection over Union (CIoU) constraint [54]. Additionally, our approach employs two discriminators, each focusing on different aspects of the fusion process: one emphasizing texture clues from visible images and the other ensuring the preservation of salient intensity clues from thermal images. By adopting this architecture, we eliminate the need for manual fusion rule design.

In summary, our contributions are threefold: (1) We propose an end-to-end TVFusionGAN framework for infrared and visible image fusion, offering a seamless fusion process without the need for manual intervention. (2) Through the utilization of two discriminators with distinct emphases, our model effectively preserves both holistic texture information from visible images and local salient object information from thermal images. (3) Our proposed structure optimally leverages the benefits of enriched image information obtained from the two discriminators, while maintaining high fusion efficiency with a single generator.

II. Related work

Conventional Image Fusion Methods: Multi-scale transform is a widely adopted approach in image fusion, involving decomposing infrared and visible images into different scales and fusing them using specific fusion rules. Pyramidal transforms [43], [4], [26], wavelet transform [34], [25], and curvelet transform [5], [6], [7] are among the most classical methods, along with their variants [20].

Non-multi-scale transform methods encompass various techniques not solely reliant on multi-scale transform. These include non-linear methods, pixel-level weighted averaging, estimation-based methods, and color composite fusion. Therrien et al. [41] proposed a spatially adaptive enhancement approach followed by fusion to combine low-light visible and infrared images, demonstrating effective performance [8], [9], [10]. Principal component analysis (PCA) [39] and adaptive weighted averaging [17] are two representative pixel-level methods. Estimation-based approaches, such as maximum a posteriori (MAP) theory, utilize prior and image formation models. For instance, Shen et al. [40] introduced a

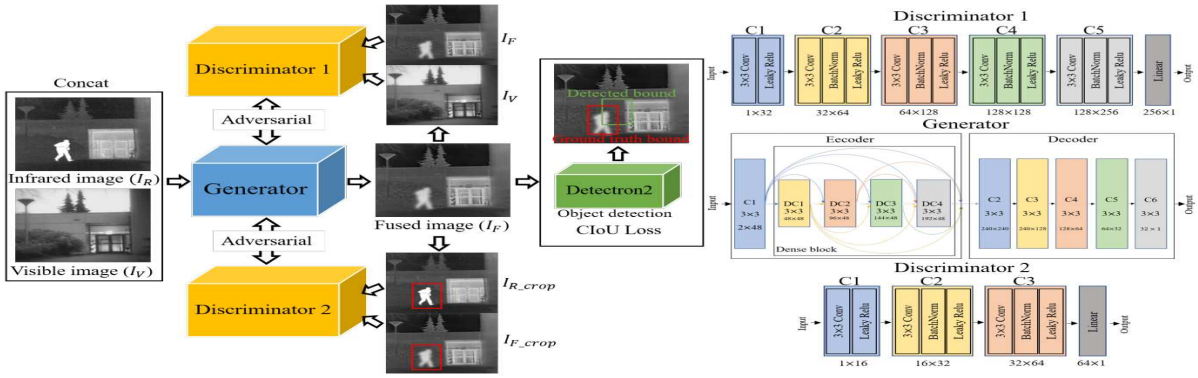

Fig. 1. Illustration of the proposed image fusion pipeline. The generator synthesizes a fused image from input thermal and visible images. Discriminator 1 emphasizes preserving holistic visual characteristics, while Discriminator 2 ensures the retention of salient objects from the thermal imagery. The network architecture on the right illustrates the components and their connections.

hierarchical multivariate conditional Gaussian random field model based on physiological findings for local contrast detection probability. Additionally, color composite fusion methods have been explored [1], [50].

Sparse representation methods utilize an over-complete dictionary to represent images with sparse coefficients, facilitating efficient representation of salient features [51], [35], [22]. These methods are less susceptible to mis-registration and leverage fixed-basis functions. Yang and Li proposed a multi-focus image fusion technique employing sparse representation [51]. Pati et al. introduced an orthogonal matching pursuit (OMP) algorithm for obtaining sparse coefficients [35]. Additionally, Li et al. proposed a Dictionary Learning method with Group Sparsity and Graph Regularization (DLGSGR), ensuring group sparsity and preserving local group geometrical structure [22].

Saliency-based techniques improve the visual quality of fused images by highlighting salient objects and pixel intensities. Zhang et al. [53] introduced a hybrid method that integrates multi-scale decomposition and saliency detection to retain global salient edges, local salient objects, and object contrast. However, these approaches frequently incorporate similar salient features, such as edges and lines, and depend on manual feature extraction and fusion rules, resulting in complex implementations to achieve enhanced performance.

Learning-based image fusion: In addition to the aforementioned methods, deep learning-based approaches [37], [19], [18], [30], [29], [28], [14], [23] have garnered significant attention for their ability to extract salient features from various image types. These methods leverage convolutional neural networks (CNNs) for feature extraction combined with manual fusion rules [37], [19], [18]. For instance, Prabhakar et al. introduced Deepfuse [37], an unsupervised CNN fusion architecture operating in the Y channel of two YCbCr images for multi-exposure fusion. The generated luminance channel (Y_{fused}) is then merged with Cb_{fused} and Cr_{fused} using different fusion strategies [36], [42], [46]. Liu et al. [19] proposed a fusion model based on a deep CNN to generate an accurate score map, followed by their fusion scheme for obtaining the final fusion results. Li and Wu presented DenseFuse [18], a learning-based network consisting of an encoder for feature extraction and a decoder for fusion image reconstruction. While DenseFuse employs offline fusion strategies during testing, these strategies are applied between the encoder and decoder networks. Notably, the CNN-based fusion models mentioned above rely on manual methods to obtain the final fusion result.

Subsequently, GAN-based fusion methods automate the fusion process by establishing an adversarial game between a generator and a discriminator [30], [28], [29]. The generator synthesizes a fusion image sample from infrared and visible images to deceive the discriminator, which distinguishes real and fake data. Various improved GAN-based fusion methods have been proposed, including least-squares GANs and conditional GANs. These methods enhance the fusion process by incorporating additional loss functions in the generator. For instance, Ma et al. [29] introduced a dual-discriminator conditional GAN with an improved generator loss, where the discriminators specialize in analyzing the infrared and visible properties within the fused image.

III. TVFusionGAN Adversarial Network

The proposed model leverages an adversarial game between a dual-discriminator and a generator to reconstruct fused images, incorporating a CIoU constraint and a salient target-based constraint. The salient target-based constraint ensures that the pixel intensities of the fused image closely match those of the corresponding objects in the thermal images, while preserving the holistic intensity texture information from the visible images.

A. Multi-level Adversarial Learning Network

Figure 1 illustrates the dual-discriminator generative adversarial fusion framework. This framework jointly learns an image that captures the object's thermal radiation from the infrared image and appearance texture from the visible image. The framework consists of three parts.

In Fig. 1, the blue-marked network on the left side represents the generative network. It employs an encoder-decoder structure, depicted in the middle of the right side, with the concatenated infrared and visible grayscale images as input. The encoder network consists of a convolutional layer (C1) and a dense block for salient feature extraction. C1 utilizes a 3×3 filter and batch normalization [13] for preliminary feature extraction. The dense block comprises four convolutional layers, each with a 3×3 filter and

batch normalization, where the inputs to each layer are the concatenation of outputs from the previous layers. The dense block aims to preserve more salient features for subsequent fusion. The decoder network includes five convolutional layers, each with a 3×3 filter, for further interpreting the feature representations. The final fused image is generated using the tanh activation function.

The second component, referred to as Discriminator 1 in Fig. 1, is responsible for discerning whether the input image is real or fake, with the objective of ensuring that the fused images resemble visible images. It comprises five convolutional layers followed by a linear layer for classification. Each convolutional layer, excluding the first one, employs a 3×3 filter and utilizes the leaky ReLU activation function [32]. Batch normalization is applied to the second through fifth layers, while the final layer serves as a linear classifier.

The third component, labeled as Discriminator 2, shares a similar architecture and functionality with Discriminator 1. However, its objective is to align specific target objects in the fused images with their corresponding objects in the infrared images. Objects with salient temperature in infrared images exhibit clearer and brighter pixel intensities compared to visible images. Discriminator 2 is also composed of five convolutional layers followed by a linear layer. The first layer does not include a batch normalization layer, while the subsequent convolutional layers mirror those of Discriminator 1 from the second to fifth layers.

B. Fusion loss constraints

We adopt Least Squares Generative Adversarial Networks (LSGANs) proposed by Mao et al. [33] to tackle the thermal and visible image fusion challenge. LSGANs have been shown to enhance the quality of generated images compared to other GAN variants and offer improved training stability over Wasserstein GANs (WGANs) [2] and standard GANs. WGANs tend to have slow convergence speeds, while standard GANs may suffer from the gradient vanishing problem during training. The objective function of TVFusionGAN in our task is as follows:

$$\begin{aligned} \min_D L(D) &= \frac{1}{2} E_{x \sim p_x} (D(x)b)^2 + \frac{1}{2} E_{z \sim p_z} (D(G(z)) - a)^2 \\ \min_G L(G) &= \frac{1}{2} E_{z \sim p_z} (D(G(z)) - c)^2 \end{aligned} \tag{1}$$

where D and G represent the discriminator and generator, respectively. The coding mechanism of TVFusionGAN is reflected in the labels a, b, and c. a and b represent fake and real images, respectively, while c signifies that D treats the data generated by G as real data. There are generally two methods to determine the values of a, b, and c in Eq. 1. One method sets $b - c = 1$ and $b - a = 2$, making the objective function equivalent to Pearson χ^2 divergence. Typically, $a = -1$, $b = 1$, and $c = 0$ in this case. Another method sets $b = c$. These two methods usually yield similar performance.

Building upon LSGANs, we propose the loss function for our fusion network, comprising three components: the loss function for the generator (G), discriminator 1 $(D1)$, and discriminator 2 $(D2)$. The generator G constraint includes adversarial loss, infrared intensity loss, gradient loss, and CIoU loss, formulated as:

$$L_G = L_{adv} + \lambda_1 L_{IR} + \lambda_2 L_{gra} + \lambda_3 L_{CIoU} \tag{2}$$

where L_G represents the total loss of G. λ_1, λ_2, and λ_3 are weights for the three losses. L_{adv} is the adversarial loss between G and the two discriminators $D1$ and $D2$ as:

$$\begin{aligned} L_{adv} = E\left(D_1\left(G\left(\text{Concat}\left(I_R, I_V\right)\right)\right) - c\right)^2 + \\ \alpha E\left(D_2\left(\text{Crop}\left(G\left(\text{Concat}\left(I_R, I_V\right)\right)\right)\right) - c\right)^2 \end{aligned} \tag{3}$$

where I_R and I_V represent the infrared and visible images, respectively. The "Crop" operation denotes cropping the fused images using known bounding box coordinates to ensure the fused component closer to real thermal data. c is the value that generator aims for the discriminator to believe for real data, and we set $c = 0$. α is the weight for $D2$.

L_{IR} and L_{gra} represent the pixel intensity loss of infrared images and the gradient loss of visible images, respectively. Thermal radiation features in infrared images are represented by their pixel intensities, while appearance texture features in visible images are expressed through their gradients [27]. Therefore, these two losses individually aim to enforce the fused images I_F to have similar pixel intensities with I_R at the instance object level and similar gradients with I_V at the entire image level. The two losses are defined as:

$$L_{IR} + L_{gra} = \left\| I_F - I_R \right\|_F^2 + \left\| gra\left(I_F\right) - gra\left(I_V\right) \right\|_F^2 \tag{4}$$

where $\| \cdot \|_F$ represents the matrix Frobenius norm and gra denotes the gradient operation. The components of L_{CIoU} utilize an existing object detection model (Detectron2) to detect objects in fused images in real-time. Then, the CIoU is calculated using the known bounding box coordinates to enhance fused images with more object details and improve object detection performance. The loss equation is as:

$$L_{CIoU} = 1 - \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|} + \frac{\rho^2\left(B, B^{gt}\right)}{c^2} + \alpha\nu \tag{5}$$

where B and B^{gt} represent the detected bounding boxes from Detectron2 and the ground truth bounding boxes, respectively. $\rho(\cdot)$ denotes the Euclidean distance, and c is the diagonal length of the smallest box enclosing the two bounding boxes. α is the weight function, and ν measures the similarity of aspect ratios. The CIoU loss is chosen because it better describes the regression of rectangular boxes.

Discriminators D_1 and D_2 engage in an adversarial game with the generator to discern whether the generated data is real or fake in each discriminator. The discriminator loss reflects the distribution of input data. The loss of D_1 is defined as follows:

$$L_{D1} = E\left(D_1\left(I_V\right) - b\right)^2 + E\left(D_1\left(I_F\right) - a\right)^2 \tag{6}$$

Different from the D_1 loss function, D_2 crops the corresponding objects in the infrared images I_R and the fused images I_F using known bounding boxes. Subsequently, it reshapes the cropped images $(I_{R_{\text{crop}}}, I_{F_{\text{crop}}})$ to match the resolution of the input of D_2. This process aims to ensure that the objects in the fused images closely resemble those in original infrared images. D_2 loss can be formulated as:

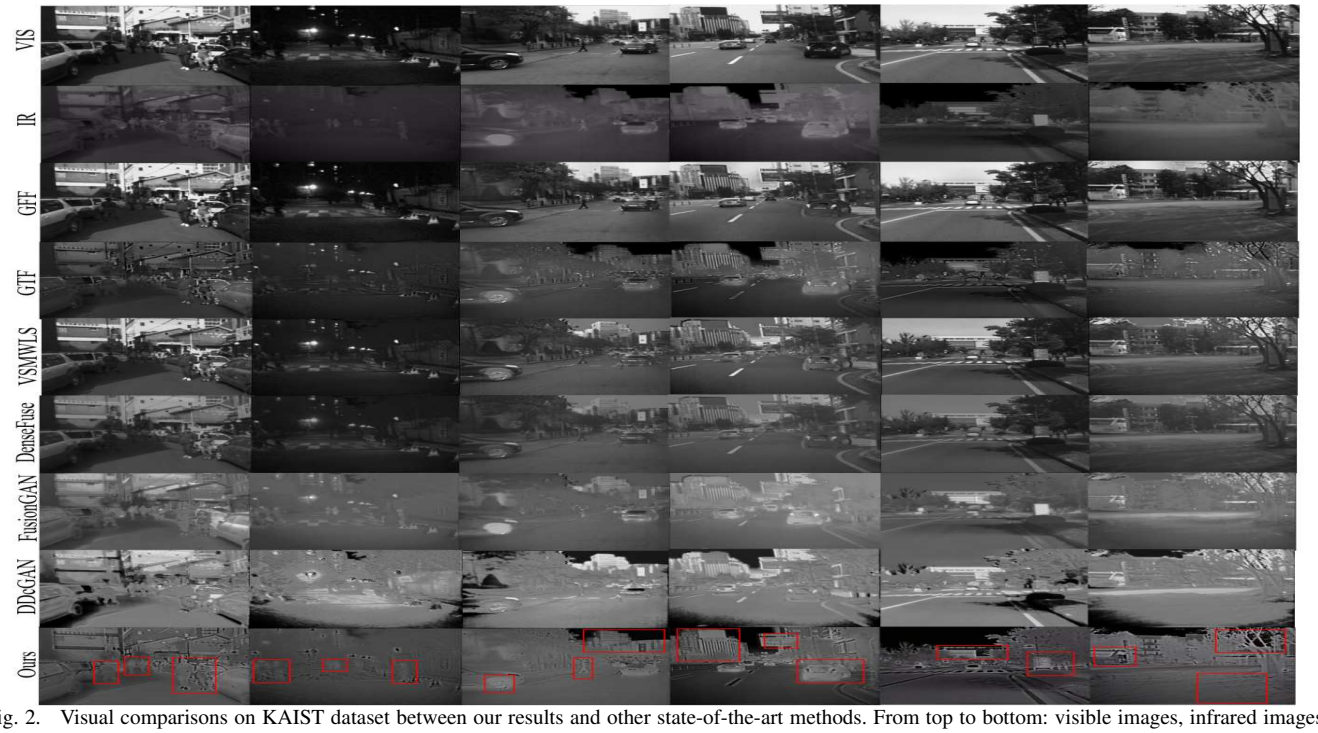

Fig. 2. Visual comparisons on KAIST dataset between our results and other state-of-the-art methods. From top to bottom: visible images, infrared images, and fusion results of GFF [20], GTF [27], VSMWLS [31], DenseFuse [18], FusionGAN [30], DDcGAN [29], and our proposed method.

$$L_{D2} = E\left(D_2\left(I_{R\_\text{crop}}\right) - b\right)^2 + E\left(D_2\left(I_{F\_\text{crop}}\right) - a\right)^2 \tag{7}$$

where a and b in both L_{D1} and L_{D2} are set to -1 and 1 respectively in our paper.

IV. Experiment Results and Analysis

A. Datasets

We evaluate our model on both KAIST [12] and TNO [44] datasets. For the TNO dataset, the model is trained on the KAIST dataset and tested on the TNO dataset. The KAIST dataset is a multi-spectral pedestrian detection benchmark dataset comprising approximately 95,000 color-thermal image pairs, each with dimensions of 640×480. The dataset annotations satisfy our loss requirements, eliminating the need for post-processing. During training, we randomly selected 8,500 images containing pedestrians, while 2,000 images were reserved for testing. The TNO dataset, widely used for infrared and visible image fusion, features diverse military-related scenes. It comprises 60 pairs of infrared and visible images, distributed across three sequences with 19, 23, and 18 image pairs each. We set the loss weights as $\alpha = 3$, $\lambda_1 = 1.9$, $\lambda_2 = 2$, and $\lambda_3 = 1$.

B. Training Details

During training, we utilize a batch size of n for the input pairs of infrared and visible images to train the two discriminators. The discriminators are optimized once using the RMSprop optimizer with a learning rate of 1×10^{-4}. Subsequently, the generator is trained using the same optimizer with a learning rate of 2×10^{-4}. We set $n = 16$, and the loss weights as $\alpha = 3$, $\lambda_1 = 1.9$, $\lambda_2 = 2$, and $\lambda_3 = 1$. Training is conducted on an NVIDIA Tesla P40 GPU with 24GB GPU memory, employing PyTorch 1.5.0 with CUDA 10.1. In testing phase, we evaluate the performance of our model on both KAIST and TNO datasets.

C. Comparison with State-of-the-art Methods

1) Qualitative results and analysis: We compare our proposed method visually with recent state-of-the-art fusion methods, including GFF [20], GTF [27], VSMWLS [31], DenseFuse [18], FusionGAN [30], and DDcGAN [29], on the KAIST and TNO datasets. Our approach preserves more details from both infrared and visible images, as illustrated in Fig. 2 and Fig. 3. The boundaries of objects in our fused images are clearer, with higher intensity contrast, enabling better distinction of temperature features, particularly for pedestrians, from the background. Moreover, our fused images maintain extensive texture information from visible images, such as clouds, grass, and pavilion structures.

We present visual results from our model and other state-of-the-art methods on the KAIST dataset (Fig. 2). Our network generates fusion images with enhanced details from both infrared and visible images. The proposed CIoU loss and D_2 loss effectively preserve local thermal radiation in infrared images, particularly for pedestrians. The designated gradient loss and D_1 loss successfully retain holistic appearance textures from visible images.

In the first column, pedestrians exhibit enhanced texture and clearer appearance, retaining edge contour information and intensities of pedestrian's infrared radiation. The second column preserves visible image textures, such as sidewalk and wall textures, along with prominent surface temperature information of pedestrians in infrared images. The third column retains both thermal radiation (mid-road pedestrians) and clear textures (vehicle wheels and billboard text). Similar performance is observed in the fourth and fifth columns, highlighting pedestrians in different scenes. Even in the absence of pedestrians in source images (last column), our

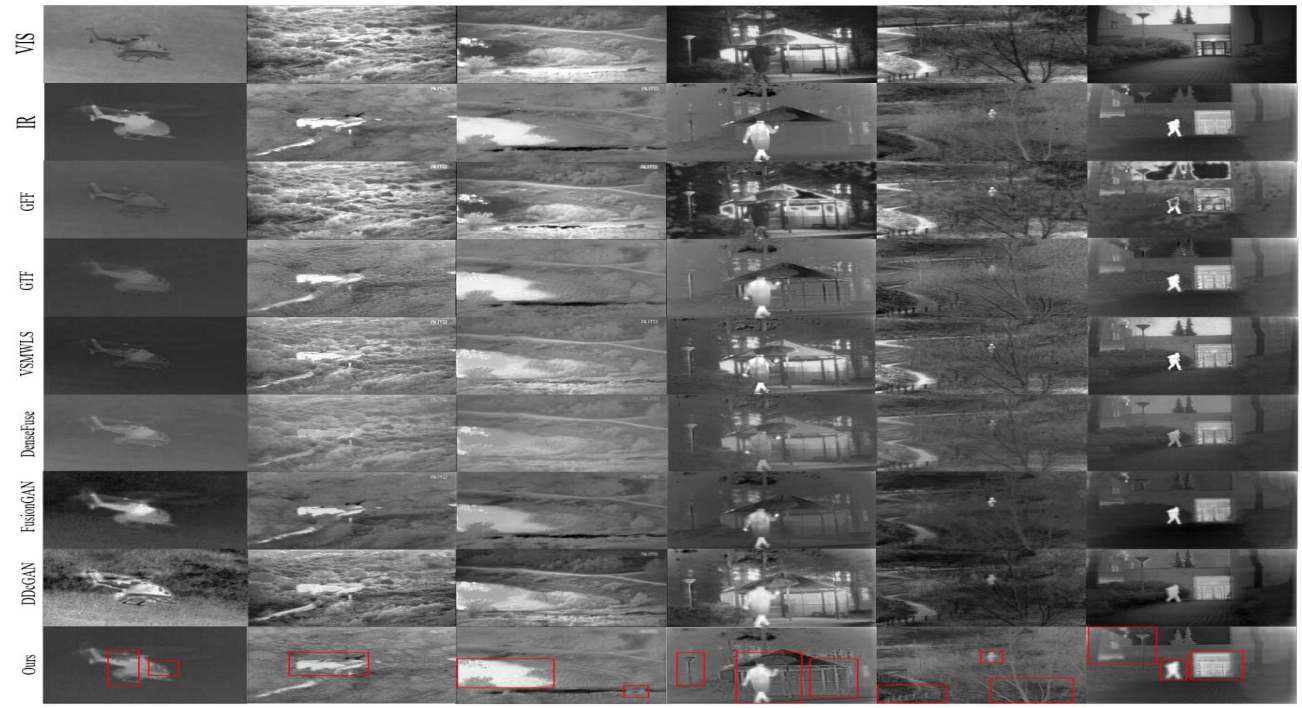

Fig. 3. Visual comparisons on the TNO dataset between our method and state-of-the-art techniques. From top to bottom: visible images, infrared images, and fusion outcomes of GFF [20], GTF [27], VSMWLS [31], DenseFuse [18], FusionGAN [30], DDcGAN [29], and our proposed method.

fused image exhibits extensive details, including fine and scattered tree shadows on roads.

We also present visual results from our model and other state-of-the-art approaches on the TNO dataset (Fig. 3). Our fusion results comprehensively capture the main features of both infrared and visual images. The thermal radiation of pedestrians in the source infrared image is effectively preserved in the last three columns. Extensive texture information from the source visual image is clearly retained in our fused results, such as the windows of the helicopter (first column), building contour (second column), streetlight, and pavilion (fourth column), as well as railing and pavement (fifth column).

The fused result in the second column exhibits higher contrast and more prominent textures compared to other methods, showcasing the combination of thermal radiation from the source infrared image and edge contour information from the source visible image, as seen in the examples.

2) Quantitative Results and Analysis: To provide a quantitative evaluation, we present results obtained from our model and other methods using five evaluation metrics: entropy (EN) [38], spatial frequency (SF), standard deviation (SD), structural similarity index measure (SSIM) [47], and correlation coefficient (CC). Each metric is calculated between the fusion image and both the visual and infrared images, with the average of the two metrics taken into account. The summarized quantitative analysis results are presented in the left side of Table I.

Entropy (EN) indicates the richness of information in the image, with higher values suggesting more information content. Spatial frequency (SF) reflects the presence of edges and appearance textures, with larger values indicating richer details. Standard deviation (SD) measures the contrast of the image, with higher values indicating greater contrast. Structural similarity index measure (SSIM) and correlation coefficient (CC) both assess the similarity between the fused image and the source images.

While our model shows slightly lower EN and SD metric results compared to DDcGAN, the visual inspection reveals that our fusion results contain more details and exhibit more obvious contrast on both datasets. Despite having an SSIM result equal to FusionGAN on the TNO dataset, our CC metric result is superior to FusionGAN, reflecting similar performance as SSIM. Moreover, our numerical results outperform other methods under most evaluation metrics. These findings suggest that our model effectively preserves salient features on both holistic and local scales.

3) Object Detection Verification: To further assess the effectiveness of our fused images, we conduct object detection comparisons across various fusion methods and the two source images. Fig. 4 illustrates the results, showcasing the stability and robustness of our fused images in object detection tasks, despite some bounding boxes in our results not being complete compared to ground truth.

On the KAIST dataset, in the daytime scene depicted in the first row, our detection results exhibit complete and accurate consistency with the ground truth. Notably, our method outperforms other results, except for the source visible image and the GFF method's fused image. Although the source visible image and the GFF method's fused image achieve identical object detection results as the ground truth, they lack crucial thermal radiation information from the infrared image. This limitation is reflected in the similarity between fused images of the GFF method and source visible images.

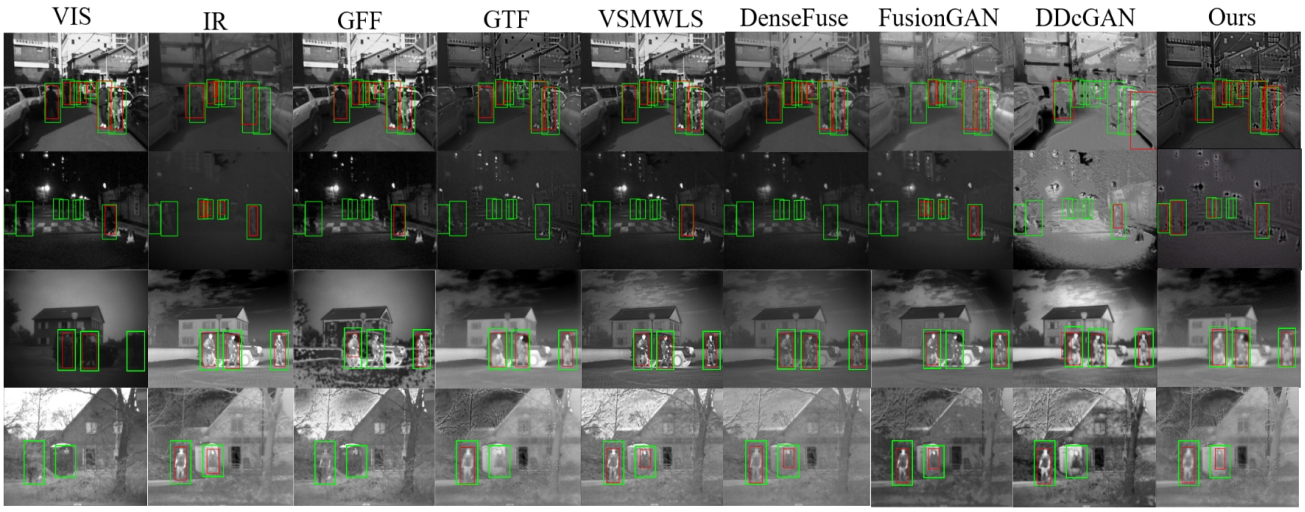

Fig. 4. Object detection results for fused images in different methods. Detection performance on both KAIST and TNO datasets are evaluated using Detectron2 [48]. Red and green bounding boxes respectively represent detected boxes and ground truth boxes.

Datasets	Metrics	GFF [20]	GTF [27]	VSMWLS [31]	DenseFuse [18]	FusionGAN [30]	DDcGAN [29]	Ours
KAIST	EN [38]	5.96	6.03	5.48	6.34	6.38	**6.67**	6.51
	SF	6.94	6.18	5.64	5.85	5.83	10.10	**10.25**
	SD	22.90	24.08	20.08	27.72	25.45	**40.43**	39.58
	SSIM[47]	0.48	0.39	0.51	0.44	0.53	0.50	**0.58**
	CC	0.46	0.41	0.50	0.51	0.44	0.55	**0.57**
TNO	EN [38]	6.38	6.01	5.89	6.43	6.44	**7.42**	6.64
	SF	6.99	6.28	5.74	5.95	5.25	11.72	**11.83**
	SD	20.97	23.46	23.82	24.72	28.43	**45.58**	40.02
	SSIM [47]	0.40	0.50	0.45	0.47	**0.57**	0.51	**0.57**
	CC	0.47	0.48	0.51	0.53	0.44	0.52	**0.56**

Metrics	VIS	IR	GFF [20]	GTF [27]	VSMWLS [31]	DenseFuse [18]	FusionGAN [30]	DDcGAN [29]	Ours
Precision	50.00%	53.84%	42.85%	43.75%	**64.29 %**	53.33%	60.00%	33.33%	54.83%
Recall	37.04%	25.93%	33.33%	25.93%	33.33%	29.63%	22.22%	17.41%	**40.74 %**
mAP	29.50%	20.56%	25.84%	22.42%	29.53%	25.84%	16.85%	14.94%	**31.00 %**
Precision	66.67%	71.43%	75.00%	66.67%	71.43%	85.71%	57.14%	42.86%	**85.71 %**
Recall	25.00%	62.50%	37.50%	50.00%	62.50%	**75.00 %**	50.00%	37.50%	**75.00 %**
mAP	23.33%	24.29%	25.00%	23.33%	24.29%	**28.14 %**	21.43%	18.57%	**28.14 %**

TABLE I

THE OVERALL EVALUATIONS OF IMAGE FUSION EFFECT AND ITS APPLICATIONS ON OBJECT DETECTION ARE PRESENTED. ON THE LEFT, THE FUSED RESULTS OF QUANTITATIVE ANALYSIS ON BOTH KAIST AND TNO DATASETS ARE SHOWN WITH FIVE EVALUATION METRICS: ENTROPY (EN), SPATIAL FREQUENCY (SF), STANDARD DEVIATION (SD), STRUCTURAL SIMILARITY INDEX MEASURE (SSIM), AND CORRELATION COEFFICIENT (CC). ON THE RIGHT, THE OBJECT DETECTION RESULTS OF QUANTITATIVE ANALYSIS ON BOTH KAIST AND TNO DATASETS ARE DISPLAYED WITH THREE EVALUATION METRICS: PRECISION, RECALL, AND MEAN AVERAGE PRECISION (MAP). BOLD FONT INDICATES THE BEST RESULT.

In the night scene depicted in the second row, our fused result significantly outperforms both source images and other methods, enhancing the intensity of pedestrians. Consequently, some pedestrians undetectable in other methods become detectable in ours, such as the two pedestrians on the left of the image. Moving to the TNO dataset, as shown in the last two rows of Figure 4, our fused image's detection results surpass those of other methods and both source images, aligning well with the ground truth. While visual detection results in several other fusion methods may appear similar to ours, the comprehensive quantitative results demonstrate the superiority of our fusion approach.

Overall, our fused images not only achieve better object detection but also contain richer scene information compared to any single source image. The success of object detection further validates the effectiveness of our proposed fusion model. We also provide quantitative results for object detection performance, comparing our approach with others. Three evaluation metrics, precision, recall, and mean average precision (mAP), are employed to assess the quality of our fused images. These metrics are defined as follows: precision $= \frac{TPs}{TPs+FPs}$, recall $= \frac{TPs}{P}$, and $mAP = \sum \text{Average Precision} / N(\text{Classes})$. True Positives (TP) and False Positives (FP) are determined using Intersection over Union (IoU) with a threshold of 0.5. A detected bounding box is labeled TP if its IoU is greater than or equal to 0.5; otherwise, it is labeled FP. P in the recall equation represents the total number of ground truth bounding boxes. Higher values for these metrics indicate better quality of fused images.

The results presented on the right side of Table I for both datasets show that our method outperforms both source images and other fusion methods in terms of recall and mAP on the KAIST dataset, except for slightly lower mAP value compared to VSMWLS. Our recall results indicate that our method can detect the most correct items compared to others. On the TNO dataset, our quantitative results surpass both source images and other fusion methods.

V. CONCLUSION

This paper presents a novel fusion approach based on generative adversarial networks. We introduce TVFusionGAN, which incorporates two discriminators and a generator in its training scheme to generate fused images with rich details. To establish a reliable image fusion model, we introduce the CIoU loss in the generator to enhance the quality of the fused image, aiming to facilitate subsequent object detection. One discriminator is dedicated to preserving the holistic textures of visible images, while the other focuses on enhancing the instance-level salience of the fused image, particularly for pedestrians. Despite being trained on the KAIST dataset, our fusion model demonstrates exceptional performance on the TNO dataset as well.

Acknowledgement: This publication is based upon work supported by NSF under Awards No. 2334246 and 2334690.

REFERENCES

[1] M. Aguilar and J. R. New. Fusion of multi-modality volumetric medical imagery. In *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002.(IEEE Cat. No. 02EX5997)*, volume 2, pages 1206–1212. IEEE, 2002.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[3] D. P. Bavirisetti, G. Xiao, and G. Liu. Multi-sensor image fusion based on fourth order partial differential equations. In *2017 20th International conference on information fusion (Fusion)*, pages 1–9. IEEE, 2017.

[4] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.

[5] E. Candes, L. Demanet, D. Donoho, and L. Ying. Fast discrete curvelet transforms. *multiscale modeling & simulation*, 5(3):861–899, 2006.

[6] E. J. Candes and D. L. Donoho. Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Technical report, Stanford Univ Ca Dept of Statistics, 2000.

[7] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(2):219–266, 2004.

[8] A. Dogra, S. Agrawal, B. Goyal, N. Khandelwal, and C. K. Ahuja. Color and grey scale fusion of osseous and vascular information. *Journal of computational science*, 17:103–114, 2016.

[9] A. Dogra, B. Goyal, and S. Agrawal. Bone vessel image fusion via generalized reisz wavelet transform using averaging fusion rule. *Journal of computational science*, 21:371–378, 2017.

[10] A. Dogra, B. Goyal, S. Agrawal, and C. K. Ahuja. Efficient fusion of osseous and vascular details in wavelet domain. *Pattern Recognition Letters*, 94:189–193, 2017.

[11] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman. Pedestrian detection in thermal images using saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[14] M. X. H. Z. Jiayi Ma, Linfeng Tang and G. Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[15] W. Kong, Y. Lei, and H. Zhao. Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. *Infrared Physics & Technology*, 67:161–172, 2014.

[16] W. Kong, L. Zhang, and Y. Lei. Novel fusion method for visible light and infrared images based on nsst–sf–pcnn. *Infrared Physics & Technology*, 65:103–112, 2014.

[17] E. Lallier and M. Farooq. A real time pixel-level based image fusion via adaptive weight averaging. In *Proceedings of the third international conference on information fusion*, volume 2, pages WEC3–3. IEEE, 2000.

[18] H. Li and X.-J. Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

[19] H. Li, X.-J. Wu, and J. Kittler. Infrared and visible image fusion using a deep learning framework. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2705–2710. IEEE, 2018.

[20] S. Li, X. Kang, and J. Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013.

[21] S. Li, B. Yang, and J. Hu. Performance comparison of different multi-resolution transforms for image fusion. *Information Fusion*, 12(2):74–84, 2011.

[22] S. Li, H. Yin, and L. Fang. Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Transactions on biomedical engineering*, 59(12):3450–3459, 2012.

[23] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[24] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang. Image fusion with convolutional sparse representation. *IEEE signal processing letters*, 23(12):1882–1886, 2016.

[25] Y. Liu, J. Jin, Q. Wang, Y. Shen, and X. Dong. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Processing*, 97:9–30, 2014.

[26] Z. Liu, K. Tsukada, K. Hanasaki, Y.-K. Ho, and Y. Dai. Image fusion by using steerable pyramid. *Pattern Recognition Letters*, 22(9):929–939, 2001.

[27] J. Ma, C. Chen, C. Li, and J. Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016.

[28] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.

[29] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[30] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26, 2019.

[31] J. Ma, Z. Zhou, B. Wang, and H. Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82:8–17, 2017.

[32] A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA, 2013.

[33] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[34] G. Pajares and J. M. De La Cruz. A wavelet-based image fusion tutorial. *Pattern recognition*, 37(9):1855–1872, 2004.

[35] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

[36] K. R. Prabhakar and R. V. Babu. Ghosting-free multi-exposure image fusion in gradient domain. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1766–1770. IEEE, 2016.

[37] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE international conference on computer vision*, pages 4714–4722, 2017.

[38] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

[39] O. Rockinger and T. Fechner. Pixel-level image fusion: the case of image sequences. In *Signal processing, sensor fusion, and target recognition VII*, volume 3374, pages 378–388. SPIE, 1998.

[40] R. Shen, I. Cheng, and A. Basu. Qoe-based multi-exposure fusion in hierarchical multivariate gaussian crf. *IEEE Transactions on Image Processing*, 22(6):2469–2478, 2012.

[41] C. W. Therrien, J. W. Scrofani, and W. Kreb. An adaptive technique for the enhanced fusion of low-light visible with uncooled thermal infrared imagery. In *Proceedings of International Conference on Image Processing*, volume 1, pages 405–408. IEEE, 1997.

[42] M. Tico and K. Pulli. Image enhancement method via blur and noisy image fusion. In *2009 16th IEEE international conference on image processing (ICIP)*, pages 1521–1524. IEEE, 2009.

[43] A. Toet. Image fusion by a ratio of low-pass pyramid. *Pattern recognition letters*, 9(4):245–253, 1989.

[44] A. Toet et al. TNO image fusion dataset. https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029/1.

[45] J. Wang, J. Peng, X. Feng, G. He, and J. Fan. Fusion method for infrared and visible images by using non-negative sparse representation. *Infrared Physics & Technology*, 67:477–489, 2014.

[46] J. Wang, D. Xu, and B. Li. Exposure fusion based on steerable pyramid for displaying high dynamic range scenes. *Optical Engineering*, 48(11):117003–117003, 2009.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[48] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.
[49] T. Xiang, L. Yan, and R. Gao. A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain. *Infrared Physics & Technology*, 69:53–61, 2015.
[50] Z. Xue and R. S. Blum. Concealed weapon detection using color image fusion. In *Proceedings of the 6th international conference on information fusion*, volume 1, pages 622–627. Citeseer, 2003.
[51] B. Yang and S. Li. Multifocus image fusion and restoration with sparse representation. *IEEE transactions on Instrumentation and Measurement*, 59(4):884–892, 2009.
[52] B. Yang and S. Li. Visual attention guided image fusion with sparse representation. *Optik*, 125(17):4881–4888, 2014.
[53] X. Zhang, Y. Ma, F. Fan, Y. Zhang, and J. Huang. Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. *JOSA A*, 34(8):1400–1410, 2017.
[54] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.