ET-CycleGAN: Generating Thermal Images from Images in the Visible Spectrum for Facial Emotion Recognition

Gerard Pons
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Gerard.Pons@cwi.nl

Abdallah El Ali
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Abdallah.El.Ali@cwi.nl

Pablo Cesar
Centrum Wiskunde & Informatica
Delft University of Technology
Amsterdam, The Netherlands
P.S.Cesar@cwi.nl

ABSTRACT

Facial thermal imaging has in recent years shown to be an efficient modality for facial emotion recognition. However, the use of deep learning in this field is still not fully exploited given the small number and size of the current datasets. The goal of this work is to improve the performance of the existing deep networks in thermal facial emotion recognition by generating new synthesized thermal images from images in the visual spectrum (RGB). To address this challenging problem, we propose an emotion-guided thermal CycleGAN (ET-CycleGAN). This Generative Adversarial Network (GAN) regularizes the training with facial and emotion priors by extracting features from Convolutional Neural Networks (CNNs) trained for face recognition and facial emotion recognition, respectively. To assess this approach, we generated synthesized images from the training set of the USTC-NVIE dataset, and included the new data to the training set as a data augmentation strategy. By including images generated using the ET-CycleGAN, the accuracy for emotion recognition increased by 10.9%. Our initial findings highlight the importance of adding priors related to training set image attributes (in our case face and emotion priors), to ensure such attributes are maintained in the generated images.

CCS CONCEPTS

• **Computing methodologies → Reconstruction**.

KEYWORDS

emotion recognition; thermal imaging; generative adversarial networks

ACM Reference Format:
Gerard Pons, Abdallah El Ali, and Pablo Cesar. 2020. ET-CycleGAN: Generating Thermal Images from Images in the Visible Spectrum for Facial Emotion Recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425258

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425258

1 INTRODUCTION

In the recent years, automatic facial emotion recognition has attracted more attention due to the wide range of applications that benefit from it [1, 22, 26]. Most of the work on facial emotion recognition is carried out using images from the visible spectrum, but these images are sensitive to illumination changes, which can influence the performance of the emotion recognition approaches [18]. Thermal imaging records facial temperatures and is thus unaffected to illumination changes. This property indicates that emotion recognition from thermal face images may be more feasible for certain applications and situations, such as recordings during the night or under poor illumination during the day [10].

Although facial emotion recognition in the thermal spectrum has attracted more and more attention in the recent years [5, 15, 24], it is still a challenging task. One of the main reasons is the small amount of data available. Nowadays, very few datasets of facial images in the thermal spectrum are annotated with emotions. In contrast, in the last few years, several large datasets of visible images for emotion recognition were made available, such as FER2013 [8], EmotiW [6] and EmotioNet [2]. The amount of public data combined with the recent breakthroughs with Convolutional Neural Networks (CNN) [21], boosted the performance of emotion recognition from visible images [17]. Hence, we hypothesize that increasing the amount of data for emotion recognition from thermal face images may lead to an improvement of the performance.

The aim of this early work is to develop a novel Generative Adversarial Network (GAN) that, given a face image from the visible spectrum, transforms the image to the thermal spectrum preserving the present emotion and the facial traits. We present an emotion-guided thermal CycleGAN (ET-CycleGAN) to regularize GAN training with emotion priors in order to generate synthesized thermal images for facial emotion recognition. Specifically, the emotion priors are extracted from a network trained for facial emotion recognition in thermal imaging, and used to compute an emotion features loss function during the training of the GAN. The original CycleGAN [29] has demonstrated state-of-the-art results in facial image synthesis without the constraint of using paired aligned training data. This is particularly suitable in our case because the visual and thermal images in the available thermal datasets are recorded from different points of view and are, as a consequence, unaligned. In summary, the main contributions of this work are:

- We propose the ET-CycleGAN, a GAN that uses a facial and emotion priors that regularize the training in the mapping from visual to thermal face images.
- We demonstrate that by including the thermal images generated by the proposed GAN in the training process of a facial

emotion recognition network, results improve. We investigate the impact of the different losses of the proposed GAN in the training.

2 RELATED WORK

In the last few years, thermal imaging has attracted the attention of the community in affective computing due to its potential use in certain conditions [7]. Due to the small size of thermal datasets, most of the works in the literature used hand-crafted features to train a classifier. Latif et al. [14] presented a method for thermal image feature extraction using the Gray Level Co-occurrence Matrix, with a classification accuracy of 99.1%. Nguyen et al. [19] used features extracted from regions of the face, with a performance of 89.9%. Kopaczka et al. [13] achieved their best results (75.5% accuracy) using SVM together with dense SIFT. Fewer works have focused on deep approaches. Wang et al. [23] proposed to use the deep Boltzmann machine to learn features from thermal facial images, achieving an accuracy of 62.9%. Lee et al. [16] proposed a CNN for detecting emotion to identify aggressive driving using input images of the driver's face, obtained using near-infrared light and thermal camera sensors, showing a performance of 99.9%. Kamath et al. [12] proposed a customized CNN network that uses the weights obtained from the VGGFace[20] model and is fine-tuned using thermal images. The results show a performance of 96.2 %. Direct comparison among these works is not possible since they validated their methods using different (and in most cases private) thermal datasets. Only the works of He et al. [9] and Wang et al. [28] evaluated their methods using the USTC-NVIE [25] dataset, which is the dataset used in our work. However, [9] used only images for disgust, fear, and happy. In [28] they classified the 6 emotions, but only from 22 randomly selected subjects. To the best of our knowledge, none of the methods in the literature deal with the generation of synthesized thermal face images for facial emotion recognition. Our proposal is inspired by [27] and [4], in the way they include networks to compute losses during the training of the GAN.

3 PROPOSED APPROACH

In this work, the proposed ET-CycleGAN is based on an existing GAN model, known as CycleGAN. Our main interest is to generate synthesized thermal face images for emotion recognition purposes. We add constraint to the CycleGAN model by including two loss functions: facial and emotion features loss. The facial features loss function aims to guide the learning to generate visual face images that preserve the facial composition. The emotion features loss function encourages the mapping of the thermal face images to ensure that it is consistent in terms of emotion. The proposed method is depicted in Figure 1.

The loss functions involved in the proposed ET-CycleGAN are defined as follows:

Adversarial loss: The adversarial loss function $\mathcal{L}_{GAN}$ is defined as in [29]:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D) = &\min_D \max_G \{\mathbb{E}_y[\log D(y)]\} \\ &+ \mathbb{E}_x[\log(1 - D(G(x)))]\end{aligned} \tag{1}$$

where G is the generator and D the discriminator, which are trained following a minimax game strategy. $x \in X$ is the visual image and $y \in Y$ is the target thermal image. The goal of G is to synthesize thermal images from visible images, while D aims to distinguish the target thermal images from the synthesized ones. In CycleGAN, X and Y are two different image representations, and the network learns the translation $X \rightarrow Y$ and $Y \rightarrow X$ simultaneously.

Cycle consistency loss: The cycle consistency loss function $\mathcal{L}_C$ is defined as in [29]:

$$\begin{aligned}\mathcal{L}_C(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = &\parallel G_{Y\rightarrow X}(G_{X\rightarrow Y}(x)) - x \parallel_1 \\ &+ \parallel G_{X\rightarrow Y}(G_{Y\rightarrow X}(y)) - y \parallel_1\end{aligned} \tag{2}$$

Identity loss: The identity loss function $\mathcal{L}_I$ is defined as in [29]:

$$\begin{aligned}\mathcal{L}_I(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = &\mathbb{E}_y[\parallel G_{X\rightarrow Y}(y) - y \parallel_1] \\ &+ \mathbb{E}_x[\parallel G_{Y\rightarrow X}(x) - x \parallel_1]\end{aligned} \tag{3}$$

Facial features loss: The facial features loss function $\mathcal{L}_F$ is defined as follows:

$$\mathcal{L}_F(G_{Y\rightarrow X}) = \mathbb{E}_{x,y} \parallel \phi_F(G_{Y\rightarrow X}(y)) - \phi_F(x) \parallel_1 \tag{4}$$

where ϕ_F denotes the features extracted from multiple layers of the VGG-19 network pre-trained on the VGGFace2 [3] as used in [4]. $\mathcal{L}_F$ ensures that the synthesized visual image in the cycle training contains facial features that are similar to the ground-truth image.

Emotion features loss: The emotion features loss function $\mathcal{L}_E$ is defined as follows:

$$\mathcal{L}_E(G_{X\rightarrow Y}) = \mathbb{E}_{x,y} \parallel \phi_E(G_{X\rightarrow Y}(x)) - \phi_E(y) \parallel_1 \tag{5}$$

where ϕ_E denotes the features extracted from the ResNet-50 network pre-trained with thermal images of the USTC-NVIE [25] dataset for facial emotion recognition. $\mathcal{L}_E$ ensures that the synthesized thermal image contains emotion-related features that are similar to the thermal ground-truth image.

Full objective: The loss function of the ET-CycleGAN is defied as:

$$\begin{aligned}\mathcal{L}(G_{X\rightarrow Y}, G_{Y\rightarrow X}, D_X, D_Y) = &\mathcal{L}_{GAN}(G_{X\rightarrow Y}, D_Y) \\ &+ \mathcal{L}_{GAN}(G_{Y\rightarrow X}, D_X) \\ &+ \mathcal{L}_C(G_{X\rightarrow Y}, G_{Y\rightarrow X}) \\ &+ \mathcal{L}_I(G_{X\rightarrow Y}, G_{Y\rightarrow X}) \\ &+ \mathcal{L}_F(G_{Y\rightarrow X}) + \mathcal{L}_E(G_{X\rightarrow Y})\end{aligned} \tag{6}$$

4 EXPERIMENTAL RESULTS

To assess the effectiveness of the proposed GAN, we conducted experiments using the posed expressions partition of the USTC-NVIE [25] dataset. This dataset contains both spontaneous and posed expressions of 105 subjects, which were recorded simultaneously by a visible and an infrared thermal camera. The images are labelled with 6 discrete emotions: angry, disgust, fear, happy, sad, and surprise. In this work, it is used to generate the synthesized data and evaluate their influence in the facial emotion recognition task. Since our goal is to generate emotion-related thermal images of faces, we discovered that the spontaneous expressions were too subtle, leading to a generation of neutral faces in most of the cases. Given

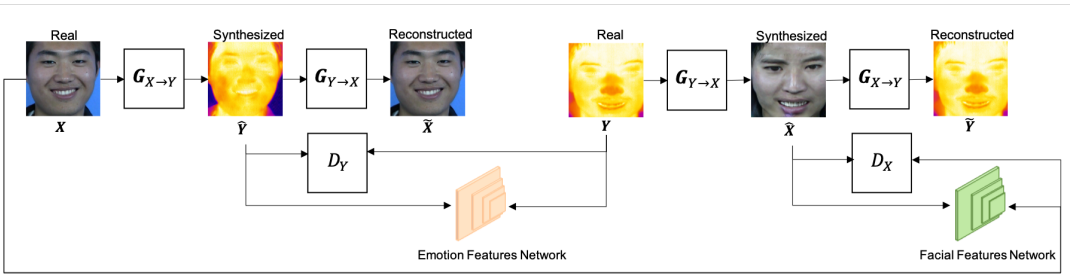


Figure 1: Flowchart illustrating the training of the proposed ET-CycleGAN. In contrast with the original CycleGAN, we included two new losses: facial and emotion. The facial network accepts the synthesized and the real image in the visible spectrum to compute the facial features loss, while the emotion network takes the synthesized and real thermal image to compute the emotion features loss.

the low number of subjects, a 10-fold cross-validation strategy was followed in all the experiments granting that images from the same subject were not present in both validation and training sets.

4.1 Implementation details

The proposed GAN builds upon the Pytorch implementation of the CycleGAN. For the Generator we used 9 ResNet blocks, and for the Discriminator we adopted the PatchGAN [11] architecture with 3 layers. For the facial network we used a VGG-19 pre-trained with the VGGFace2 dataset. As in [4], the extracted features are obtained from the *relu1-1*, *relu2-1*, *relu3-1*, *relu4-1*, and *relu5-1* layers. For the emotion network we used a ResNet-50 pre-trained with the VGGFace2, the FER2013 [8], and the USTC-NVIE-posed. For this network, the features are extracted from the last layer of the network. The GANs were trained during 200 epochs, with a learning rate of 0.0002 that was linearly decayed after 100 epochs. Since training a CycleGAN involves training four CNNs at the same time (two generators and two discriminators) and we included the forward passing of two more CNNs (VGG-19 and ResNET-50), only one image per batch could be fitted in an Nvidia GeForce RTX 2080 Ti during training. The training time per experiment is approximately 10 hours.

4.2 Quantitative evaluation

The aim of this work is to generate new synthesized thermal face images to improve the training of deep learning models in emotion recognition. To verify the effectiveness of the generated images, we included them in the training process and evaluated their impact on the learning of the network. Following a 10-fold-cross-validation strategy, for each fold we include in the training the generated images by the GAN for the other folds. For instance, for fold 0, the generated images for fold 1 to 9 were added to the training data (~ 415 images), for fold 1 the images generated during fold 0 and from 2 to 9, and so on. This way, no similar images were simultaneously used in the training and test set.

To assess the performance in facial emotion recognition, we used the same architecture as used in extracting features for the emotion loss in ET-CycleGAN. The network was trained using a training batch of 32 images, rescaling the images to a size of 224 × 224. Given the difference between domains in the pre-trained network

Method	Accuracy (mean ± 95% CI)	F1 score (mean ± 95% CI)
Baseline	0.401 ± 0.038	0.376 ± 0.040
CycleGAN	0.489 ± 0.046	0.477 ± 0.047
ET-CycleGAN	**0.510 ± 0.039**	**0.487 ± 0.041**

Table 1: Results of emotion classification after including the synthesized images generated for each method. The baseline method refers to the results obtained without including synthesized data. Results reported in average for the 10-fold-cross-validation with the 95% Confidence Intertval (CI).

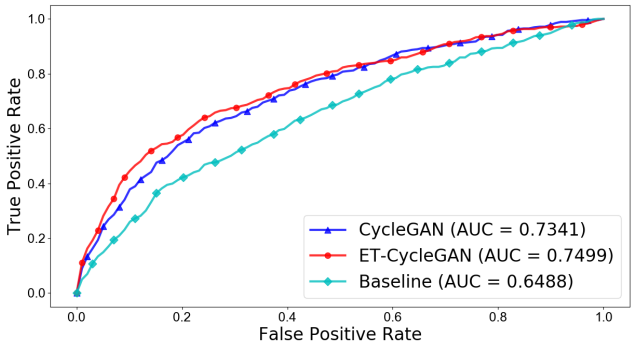


Figure 2: ROC curves for emotion classification after adding synthesized images. Baseline denotes results without including synthesized data.

and the data (visual to thermal), a learning rate of 0.01 was used along the entire training process. Batch normalization was used to prevent overfitting. An early-stopping approach was used with a maximum of 200 epochs. Table 1 summarizes the obtained results including images generated by the proposed ET-CycleGAN and the CycleGAN, and compares them with the results obtained when no synthetic images are added in the training. In a similar manner, Figure 2 shows the Receiver Operating Characteristic (ROC) curves and their corresponding area (AUC), and Figure 3 the confusion matrices. The quantitative results show that including new

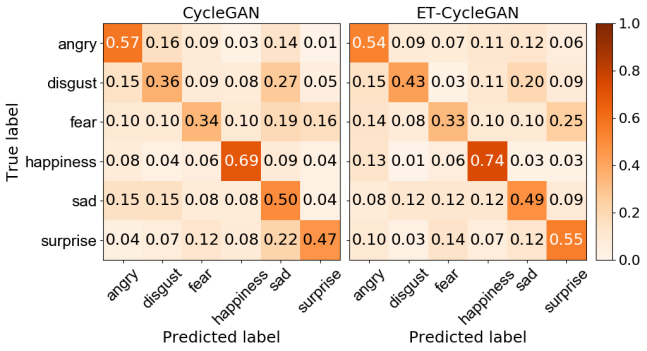


Figure 3: Confusion matrices for emotion classification after adding synthesized images.

Method	Accuracy (mean ± 95% CI)	F1 score (mean ± 95% CI)
ET-CycleGAN $\ominus\mathcal{L}_F$	0.490 ± 0.044	0.485 ± 0.065
ET-CycleGAN $\ominus\mathcal{L}_E$	0.502 ± 0.037	0.483 ± 0.033
ET-CycleGAN	**0.510 ± 0.039**	**0.487 ± 0.041**

Table 2: Ablation study results. Results of emotion classification after including the synthesized images generated for each method. $\ominus$ denotes the removal of the corresponding loss function. Results reported in average for the 10-fold-cross-validation with the 95% Confidence Intertval (CI).

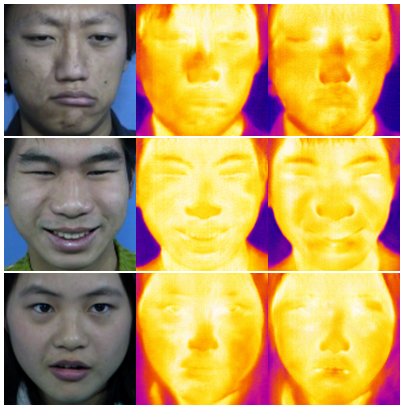
synthesized images in the training process improves the results of the facial emotion recognition network up to 10.9%. The proposed ET-CycleGAN generated the best images to improve the performance of the emotion recognition network. The inclusion of the facial and emotion losses contribute to a generation of images that preserve the facial composition and is consistent in terms of the present emotion. The confusion matrices show that ET-CycleGAN improves for disgust, happiness, and surprise, while for the other emotions results are almost the same as CycleGAN.

4.3 Ablation study

To demonstrate the effectiveness of the different loss functions included in the proposed ET-CycleGAN, an ablation study was conducted. Following the experimental procedure detailed in the previous section, we evaluated the performance of an emotion recognition network including the images generated by the ET-CycleGAN without each of the loss functions. The results of this study are shown in Table 2. ET-CycleGAN $\ominus\mathcal{L}_F$ corresponds to the proposed method without the $\mathcal{L}_F$ loss function, and ET-CycleGAN $\ominus\mathcal{L}_E$ without the $\mathcal{L}_E$ loss function. As it can be seen in Table 2, the emotion loss function $\mathcal{L}_E$ contributes slightly to the improvement, while the addition of the face $\mathcal{L}_F$ loss function results in a more pronounced improvement in the performance.

4.4 Qualitative evaluation

In the previous section we showed that including images generated with ET-CycleGAN improves the results for disgust, happiness, and surprise. In this section we show an example of the visual differences of these three emotions. Figure 4 presents examples of three images of different subjects, and their corresponding synthesized images using CycleGAN and ET-CycleGAN. The differences are subtle, but for the first image we can see that ET-CycleGAN was able to preserve the expression of the mouth. For the second example, ET-CycleGAN generated an image that shows the rise of the cheeks during the smile. Finally, in the example for surprise, our method depicted the opened mouth with a darker color.



(a) (b) (c)

Figure 4: Comparison of the generated images. First row shows an example of disgust, second row happiness, and third surprise. (a) Shows the the input RGB images, (b) the results for CycleGAN, and (c) the results for ET-CycleGAN.

5 CONCLUSIONS AND FUTURE WORK

This work intends to tackle an unresolved challenging problem, which is the generation of synthesized images to improve the performance of facial emotion recognition using thermal face images. The main limitation of this work is intrinsically related to the fact that we face a chicken and egg kind of problem: we want to generate images because the available datasets are too small to train deep networks, but in order to generate these images we need to train deep networks. Therefore, the obtained results are constrained by the dataset used in the experiments. We have presented a novel synthesis-based method for generating thermal images from images in the visible spectrum using a GAN-based approach. The proposed ET-CycleGAN method includes facial and emotion features loss functions that regularize the training. Initial results showed that this contribution improved the quality of the generated synthesized thermal images. The generated images were included as a data augmentation approach in the training of an emotion recognition network for thermal face images. The images generated by the proposed approach led to a better performance of the network. These findings highlight that adding priors related to the image attributes (in our case face and emotion priors) helps to ensure such attributes are maintained in the generated images. Future work would focus on exploring new modifications in the proposed GAN to allow us to use data from different thermal datasets and, as a consequence, improve the generalization capacity of the network.

REFERENCES

[1] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. 2003. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03.*, Vol. 5. IEEE, Madison, Wisconsin, USA, 53–53.

[2] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. 2017. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210* (2017).

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

[4] Cunjian Chen and Arun Ross. 2019. Matching Thermal to Visible Face Images Using a Semantic-Guided Generative Adversarial Network. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

[5] Irving A. Cruz-Albarran, Juan P. Benitez-Rangel, Roque A. Osornio-Rios, and Luis A. Morales-Hernandez. 2017. Human emotions detection based on a smart-thermal system of thermographic images. *Infrared Physics and Technology* 81 (2017), 250 – 261. https://doi.org/10.1016/j.infrared.2017.01.002

[6] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 653–656.

[7] Chiara Filippini, David Perpetuini, Daniela Cardone, Antonio Maria Chiarelli, and Arcangelo Merla. 2020. Thermal Infrared Imaging-Based Affective Computing and Its Application to Facilitate Human Robot Interaction: A Review. *Applied Sciences* 10, 8 (2020), 2924.

[8] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In *Neural Information Processing*, Minho Lee, Akira Hirose, Zeng-Guang Hou, and Rhee Man Kil (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 117–124.

[9] Shan He, Shangfei Wang, Wuwei Lan, Huan Fu, and Qiang Ji. 2013. Facial expression recognition using deep Boltzmann machine from thermal infrared images. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 239–244.

[10] Shuowen Hu, Nathaniel Short, Kristan Gurton, and Benjamin Riggan. 2018. Overview of polarimetric thermal imaging for biometrics. In *Polarization: Measurement, Analysis, and Remote Sensing XIII*, David B. Chenault and Dennis H. Goldstein (Eds.), Vol. 10655. International Society for Optics and Photonics, SPIE, 1 – 8.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*.

[12] Shreyas Kamath KM, Rahul Rajendran, Qianwen Wan, Karen Panetta, and Sos S Agaian. 2019. TERNet: A deep learning approach for thermal face emotion recognition. In *Mobile Multimedia/Image Processing, Security, and Applications 2019*, Vol. 10993. International Society for Optics and Photonics, 1099309.

[13] Marcin Kopaczka, Raphael Kolk, and Dorit Merhof. 2018. A fully annotated thermal face database and its application for thermal facial expression recognition. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 1–6.

[14] MH Latif, SNI Sidek, N Rusli, S Fatai, et al. 2016. Emotion detection from thermal facial imprint based on GLCM features. *ARPN J. Eng. Appl. Sci* 11 (2016), 345–350.

[15] Kwan Lee, Hyo Yoon, Jong Song, and Kang Park. 2018. Convolutional Neural Network-Based Classification of Driver's Emotion during Aggressive and Smooth Driving Using Multi-Modal Camera Sensors. *Sensors* 18, 4 (Mar 2018), 957. https://doi.org/10.3390/s18040957

[16] Kwan Woo Lee, Hyo Sik Yoon, Jong Min Song, and Kang Ryoung Park. 2018. Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors. *Sensors* 18, 4 (2018), 957.

[17] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020).

[18] Hung Nguyen, Kazunori Kotani, Fan Chen, and Bac Le. 2014. A Thermal Facial Emotion Database and Its Analysis. In *Image and Video Technology*, Reinhard Klette, Mariano Rivera, and Shin'ichi Satoh (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 397–408.

[19] Thu Nguyen, Khang Tran, and Hung Nguyen. 2018. Towards thermal region of interest for human emotion estimation. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 152–157.

[20] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

[21] Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017).

[22] Mohammad Soleymani and Maja Pantic. 2013. Emotionally Aware TV. In *Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI*. Paris, France.

[23] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang Ji. 2014. Emotion recognition from thermal infrared images using deep Boltzmann machine. *Frontiers of Computer Science* 8, 4 (2014), 609–618.

[24] Shangfei Wang and Shan He. 2013. Spontaneous Facial Expression Recognition by Fusing Thermal Infrared and Visible Images. In *Intelligent Autonomous Systems 12*, Sukhan Lee, Hyungsuck Cho, Kwang-Joon Yoon, and Jangmyung Lee (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 263–272.

[25] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang. 2010. A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. *IEEE Transactions on Multimedia* 12, 7 (Nov 2010), 682–691. https://doi.org/10.1109/TMM.2010.2060716

[26] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[27] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-Attention Generative Adversarial Networks. arXiv:stat.ML/1805.08318

[28] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3438–3446.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.