GENERATING THERMAL HUMAN FACES FOR PHYSIOLOGICAL ASSESSMENT USING THERMAL SENSOR AUXILIARY LABELS

Catherine Ordun[†]   *Edward Raff*[⋆,†]   *Sanjay Purushotham*[†]

[†]University of Maryland, Baltimore County
[⋆]Booz Allen Hamilton

ABSTRACT

Thermal images reveal medically important physiological information about human stress, signs of inflammation, and emotional mood that cannot be seen on visible images. Providing a method to generate thermal faces from visible images would be highly valuable for the telemedicine community in order to show this medical information. To the best of our knowledge, there are limited works on visible-to-thermal (VT) face translation, and many current works go the opposite direction to generate visible faces from thermal surveillance images (TV) for law enforcement applications. As a result, we introduce **favtGAN**, a VT GAN which uses the pix2pix image translation model with an auxiliary sensor label prediction network for generating thermal faces from visible images. Since most TV methods are trained on only one data source drawn from one thermal sensor, we combine datasets from faces and cityscapes. These combined data are captured from similar sensors in order to bootstrap the training and transfer learning task, especially valuable because visible-thermal face datasets are limited. Experiments on these combined datasets show that favtGAN demonstrates an increase in SSIM and PSNR scores of generated thermal faces, compared to training on a single face dataset alone.

Index Terms— generative adversarial networks, thermal images, image translation

1. INTRODUCTION

Thermal imaging can support medical AI as a diagnostic tool since decades of physiological research have shown that temperatures in facial thermograms are highly correlated to vital measures and reveal signs of stress and inflammation, otherwise hidden on visible images [1, 2, 3, 4, 5, 6, 7]. However, due to time-consuming data collection for human subjects using thermal cameras, there is usually insufficient thermal medical data to train any AI system for thermal imaging. To the best of our knowledge, we introduce the first $visible \rightarrow thermal$ (VT) facial translation method, which can be used to bootstrap or augment thermal training data. The ability to generate a thermal face image from a visible face may be challenging due to the fact that there are limited, paired thermal-visible face datasets [8]. Therefore, the ability to bootstrap the image-to-image translation task with additional training data from similar thermal sensors that share the same optical properties would improve training, even if the data are from different domains (e.g. faces and cityscapes). We train on combined visible-thermal datasets in our method, called **favtGAN** (facial-visible-thermal-GAN or "favorite GAN"). favtGAN is a simple Generative Adversarial Network (GAN) architecture built on the popular pix2pix framework [9]. Unlike pix2pix, favtGAN uses a multi-task discriminator network to perform two prediction tasks: an auxiliary network for predicting the thermal sensor label, and an adversarial network for predicting whether the generated thermal image is real or fake. We focus on the thermal sensor class prediction as an auxiliary task in an attempt to gain information about optical differences in images resulting from various thermal sensors such as sensitivity and resolution (e.g. uncooled microbolometer VOx sensor and BST ferroelectric sensor). favtGAN shows markedly better image quality results over pix2pix when combining data from face and cityscapes captured from the same thermal sensor. Our contributions are:

- The first work to study VT translation of human faces, by developing a pix2pix-based favtGAN model.
- We study the image quality of generated thermal face images which is important for medical applications.
- We bootstrap training of image translation with additional data from different domains but similar thermal sensors to improve thermal image generation.

1.1. Related Work

Several TV image-to-image translation GANs operate in the $thermal \rightarrow visible$ (TV) direction. These are not used for medical applications, but rather law enforcement and person re-identification focused on reconstructing the visible identity of persons from thermal surveillance images and thermal face recognition [10,
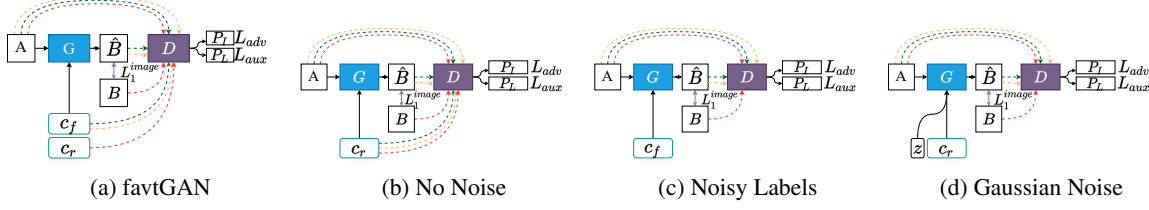
ICIP 2021

Fig. 1: **favtGAN and Three Variations for Using the Thermal Side Labels as Noise and/or Real Inputs to the Discriminator**. (a) favtGAN Baseline model where G is conditioned on "noisy labels", c_f, in the form of discrete integers (0, 1, 2), and D is conditioned on real true sensor labels, c_r in the same set of integers; (b) "No Noise" variation conditions G and D only on c_r; (c) "Noisy Labels" variation conditions G, only, on c_f; (d) "'Gaussian" variation conditions G, only, on both Gaussian noise, z, and c_r. Green, red, and yellow lines indicate three adversarial losses and three auxiliary losses - one for the G (green), real D (red), and fake D (yellow). P_I refers to the predicted image (1, 16, 16) tensor (fake or real) and P_L refers to the predicted sensor label. Full symbol notation is in the Approach.

11, 12, 13, 14, 15, 16]. Further, the TV GAN approaches we identified are trained on one single dataset collected by a single thermal sensor [1, 2, 3, 4, 5, 6, 7, 11, 12]. ThermalGAN [17] is a pipeline where the first generator is BicycleGAN [18] with some slight modifications to U-NET [19], that outputs multi-modal images of segmented masks, followed by a second generator using pix2pix [9]. ThermalGAN is trained and tested on a single dataset from one sensor, and is focused not on faces but the entire body for surveillance applications. "TV-GAN" [10] also uses pix2pix without modification, and implements a closed set recognition identity task using log loss. Side information has been used in TV GANs to explicitly regularize networks, where Zhang et al. [10] uses one-hot encoding of subject identifiers, Kniaz et al. [17] uses temperature vectors, and Chen and Ross [15] use facial labels (i.e left eye, right eye, right brow, etc.). In another approach, Zhang et al [11] explicitly feature-engineer their design for polarimetric facial TV translation. Their approach involves preservation of texture and geometric features by leveraging multiple thermal polarimetric modalities, through feature-level fusion of Stokes images. Li et al. generates thermal pedestrian landscapes from visible images as a data augmentation tool for a downstream object detection task [20]. They apply CycleGAN [21] for the landscape generation followed by an intensity inversion transformation [20]. The TV methods are different from our approach as they customize their networks with perceptual loss to improve the resolution quality, specific for data generated from one thermal sensor type. Their primary goals are to guarantee the identity from synthesized visible faces. Whereas, in our approach we seek to combine data from multiple thermal sensors in order to reconstruct thermal facial features. While the work by Li, et al. is similar, their focus is on the entire body and pedestrian landscapes versus close-up thermal face translation which makes our approach different.

2. APPROACH

Our approach, favtGAN, modifies the pix2pix image-to-image translation framework by conditioning the generator and discriminator on thermal sensor labels shown in Fig.1a. We refer to the favtGAN model as the "favtGAN Baseline" for the rest of the paper. The discriminator is a multi-task network that outputs fake or real thermal patch probabilities used to calculate an adversarial loss, and probabilities to calculate an auxiliary loss that predicts the thermal sensor label. We design three variations of favtGAN shown in Fig.1b-d, to test the impact of using fake sensor labels, c_f, or c_r, real sensor labels to condition the network. Notation for Fig. 1 and formulas that follow are: G is the generator, D is the discriminator, c_f are noisy labels which are random discrete integers in the set of thermal sensor labels (i.e. 0, 1, 2), c_r are the real thermal sensor labels, A is the real visible image, B is the real thermal image, and $\hat{B}$ is the generated thermal image. Our code is available at https://github.com/umbc-sanjaylab/favtGAN.

2.1. Generator

We use U-NET [19] as a single generator for favtGAN. The generator uses both the noisy class label c_f and visible image A to generate images $\hat{B} = G(A, c_f)$. The total generator loss in Eq.4 is a composition of the adversarial loss in Eq.1., auxiliary (cross entropy) loss Eq.2, and the L_1^{image} pixel reconstruction loss shown in Eq.3. Note that in Eq.1, we adopt the least-squares GAN approach (LSGAN) [22] as the generator adversarial loss between the discriminator output of $D(A, \hat{B}, c_f)$ and 1, the real adversarial label. We applied one-sided label smoothing where 1 is replaced with 0.9 per [23].

$$L_{Adv}(G) = \frac{1}{2} \mathop{\mathbb{E}}_{A \sim p_{vis}, c_f \sim U\{0,1\}, \hat{B} \sim p_G} [(D(A, \hat{B}, c_f) - 1)^2]$$
(1)

$$L_{aux}(G) = \mathop{\mathbb{E}}_{A \sim p_{vis}, c_f \sim U\{0,1\}, \hat{B} \sim p_G} [\log C(A, \hat{B}, c_f)] \quad (2)$$

1320

$$L_1^{image}(G) = \mathop{\mathbb{E}}_{B \sim p_{thr}, \hat{B} \sim p_G} \|B - \hat{B})\|_1 \quad (3)$$

$$L_G = L_{Adv}(G) + L_{aux}(G) + \lambda L_1^{image}(G) \quad (4)$$

2.2. Discriminator

We use the same architecture as the pix2pix PatchGAN discriminator but modify it by adding a second layer that performs sensor class label prediction using softmax activation. As a result, the discriminator takes as input the visible image, thermal image, and the sensor label. Using this input, our discriminator conducts the adversarial task of predicting real versus fake on a 16 x 16 patch across a 256 x 256 input and second, a classification task to predict the sensor class labels c for the input. The adversarial real and fake discriminator losses are shown in Eq.5 and Eq.6, respectively, using the LSGAN approach [22]. Auxiliary losses are shown in Eqs. 7 and 8. The total discriminator loss is shown in Eq.9. The full objective function for favtGAN is given by Eq.10, and solved using the Adam optimizer.

$$L_{Adv_{D_{real}}} = \frac{1}{2} \mathop{\mathbb{E}}_{A \sim p_{vis}, B \sim p_{thr}, c_r \sim p_{thr}^l} \left[(D(A, B, c_r) - 1)^2 \right]$$
$$(5)$$

$$L_{Adv_{D_{fake}}} = \frac{1}{2} \mathop{\mathbb{E}}_{A \sim p_{vis}, c_f \sim U\{0,1\}, \hat{B} \sim p_G} \left[(D(A, \hat{B}, c_f) - 0)^2 \right]$$
$$(6)$$

$$L_{aux_{D_{real}}} = \mathop{\mathbb{E}}_{A \sim p_{vis}, B \sim p_{thr}, c_r \sim p_{thr}^l} \left[\log C(A, B, c_r) \right] \quad (7)$$

$$L_{aux_{D_{fake}}} = \mathop{\mathbb{E}}_{A \sim p_{vis}, c_f \sim U\{0,1\}, \hat{B} \sim p_G} \left[\log C(A, \hat{B}, c_f) \right] \quad (8)$$

$$L_D = \frac{1}{2} \left[(L_{Adv_{D_{real}}} + L_{aux_{D_{real}}}) + (L_{Adv_{D_{fake}}} + L_{aux_{D_{fake}}}) \right]$$
$$(9)$$

$$G^* = \min_G L_G + \min_D L_D \quad (10)$$

3. EXPERIMENTS

3.1. Datasets

We use four paired thermal-visible image datasets: Eurecom [24], FLIR ADAS [25], Iris [26], and Oklahoma State University (OSU) Color-Thermal [27]. From 2100 images, the Eurecom dataset was converted into a paired dataset of 1050 images (945 train images and 45 train subjects, 105 test images and 5 test subjects). The Iris dataset consists of faces with expressions (mouth open, smiling, mouth closed), and five illumination rotations. We only use expressions leading to 944 paired images (846 train images and 26 train subjects, 98 test images and 3 test subjects). We used the Advanced Driver Assistance Systems (ADAS) made by Flir [25] and randomly sampled images to make 940 pairs (842 train images, 98 test images) to balance against Iris and Eurecom. We also used the OSU dataset [27] consisting of 1054 pairs (843 train images, 211 test images) taken from

"Sequence 1". Eurecom and ADAS thermal images are captured by Flir cameras which use microbolometer sensors. Iris and OSU used a Raytheon Palm-IR-Pro and Raytheon 300D thermal camera, respectively, which are BST ferroelectric sensors.

3.2. Experimental Design

There are no VT GAN translation architectures designed for thermal face translation, to our knowledge. As a result, we compared our approach to pix2pix, which has been used in both "TV GAN" [10] and ThermalGAN [17], in addition to three variations of the favtGAN architecture shown in Fig.2. Prior to selecting pix2pix as a comparative model, we also trained our datasets on CycleGAN [21] and StarGAN [28] which resulted in poor generated thermal images. We implement seven total experiments to generate Eurecom and Iris thermal faces. First, we train pix2pix without modification on only the single face dataset of either Eurecom or Iris to establish baseline SSIM and PSNR scores. We then train five experiments on the combined face datasets of Eurecom + Iris (EI) using favtGAN and pix2pix unmodified. The seventh experiment, indicated in bold in Table 1, trains the best performing favtGAN on different cityscape and human faces: ADAS + Eurecom (EA), and OSU + Iris (IO). We modified the implementation of the pix2pix GAN developed by Erik Lindernoren [29]. Models were trained using PyTorch on a RTX 8000 GPU, trained to 2000 epochs. Each experiment is set to a batch size of 12, Adam optimizer with a learning rate of 0.00002, b1 of 0.5, b2 of 0.999, decay of 100, 3 channels, and a lambda hyperparameter for the pixel reconstruction loss of 100. We use SSIM and PSNR to evaluate the image quality of the generated thermal faces since these metrics have been used to judge the quality of multispectral and near IR images from remote sensing, polarimetric face recognition, and biomedical applications [11, 30, 31, 32, 33].

4. RESULTS

4.1. Quantitative Results

Results are shown in Table 1. We obtain the best results when training the favtGAN Noisy Labels variation on faces and cityscapes with similar thermal properties. The ADAS + Eurecom experiment trained on favtGAN Noisy Labels show 2.69% increase of SSIM and 3.29% increase of PSNR, over pix2pix on Eurecom alone. The same domain transfer effect occurs with Iris. Training the favtGAN Noisy Label implementation on OSU + Iris led to a 0.72% increase of SSIM and 2.58% increase of PSNR scores, over pix2pix trained on Iris, alone. But, we observe that training favtGAN Baseline on combined faces (Eurecom + Iris), even from different sensors, improves scores over training on Eurecom, alone, with a
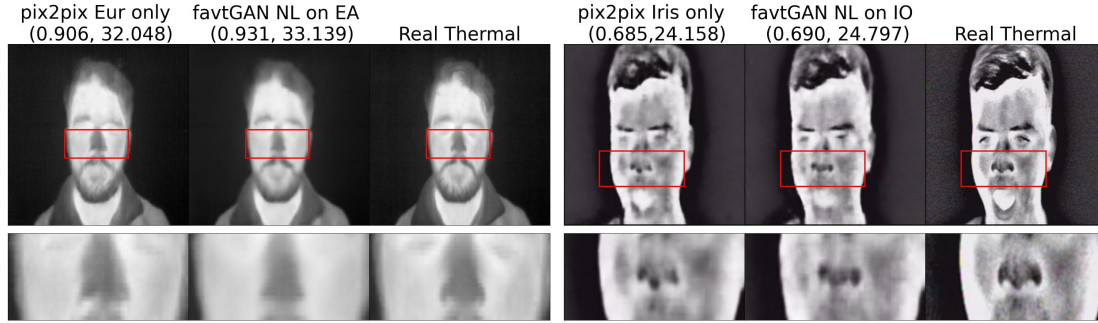
Fig. 2: **Generated Thermal Images Translated from the Visible Test Set.** Samples are shown from the best performing favtGAN experiment trained on combined face and cityscape datasets, compared to training pix2pix on a single dataset. Average SSIM and PSNR scores are provided. Red boxes show regions of interest, which are magnified in the second row. "NL": Fig.1c Noisy Labels variation.

2.09% improvement in SSIM and 0.98% improvement in PSNR. However, best results are seen when combining data from similar thermal sensors. Even if the domains are different, this may be a practical way to bootstrap training and transfer learning for VT face translation. These results are significant because improved image quality represents physiological accuracy of temperature patterns. The favtGAN variations that do not condition the generator on noise, but instead use the true sensor labels, c_r, show the lowest scores.

4.2. Qualitative Results

In Fig.2 we show samples comparing pix2pix trained on a single dataset and the best performing favtGAN architecture trained on the combined face and cityscape dataset. Pix2pix images trained on a single dataset alone, generate coarser textures and darker pixels than favtGAN. The red box shows the nose region which we observed had the greatest variability when qualitatively inspected. These are magnified in the bottom row. Pix2pix trained only on Eurecom shows structural deformities and smudging with less articulation than favtGAN. Iris images are more challenging where the differences are nuanced. The thermal image generated from pix2pix has slightly coarser texture on the right cheek compared to the favtGAN image. The nose regions are different where pix2pix generates lighter pixels and irregular structure than favtGAN. The placement of pixels, their contrast, hue, and distribution is important for thermal physiology because they are representative of facial temperatures. These temperatures are linked to different physiological responses and conditions [1]. Therefore, failed reconstruction can lead to poor medical interpretation.

5. CONCLUSION

We have demonstrated the ability to implement visible-to-thermal image translation of human faces using favtGAN, a modified pix2pix network conditioned on ther-

Table 1: **Image Quality Metrics using Mean SSIM and Mean PSNR for Generated Thermal Images, Translated from the Visible Test Set.** SSIM % and PSNR % show the relative change compared to pix2pix trained only a single face dataset. FG: favtGAN

		Eurecom			
Dataset	**Experiment**	**SSIM**	**PSNR**	**SSIM %**	**PSNR %**
Eurecom	pix2pix	0.906	32.048	-	-
EI	pix2pix	0.924	32.133	1.98%	0.26%
EI	FG-Baseline	0.925	32.366	2.09%	0.98%
EI	FG-No Noise	0.914	29.230	0.85%	-9.64%
EI	FG-Noisy Labels	0.925	31.835	2.02%	-0.67%
EI	FG-Gauss. Noise	0.909	28.242	0.36%	-13.48%
EA	FG-Baseline	0.931	33.104	2.69%	3.19%
EA	**FG-Noisy Labels**	**0.931**	**33.139**	**2.69%**	**3.29%**

		Iris			
Dataset	**Experiment**	**SSIM**	**PSNR**	**SSIM %**	**PSNR %**
Iris	pix2pix	0.685	24.158	-	-
EI	pix2pix	0.681	23.946	-0.54%	-0.89%
EI	FG-Baseline	0.682	24.060	-0.37%	-0.41%
EI	FG-No Noise	0.653	22.000	-4.91%	-9.81%
EI	FG-Noisy Labels	0.682	23.990	-0.42%	-0.70%
EI	FG-Gauss. Noise	0.652	22.083	-5.07%	-9.40%
IO	FG-Baseline	0.686	24.474	0.15%	1.29%
IO	**FG-Noisy Labels**	**0.690**	**24.797**	**0.72%**	**2.58%**

mal sensor class labels. Empirical results show that training favtGAN on combined faces and cityscapes data improves image quality if they share similar thermal sensor types. These results are preliminary and open up many new research questions to include explicit application of domain adaptation of various thermal domains, and modeling optical long-wave infrared properties.

References

[1] S. Ioannou *et al.*, "Thermal infrared imaging in psychophysiology: potentialities and limits," *Psychophysiology*, vol. 51, no. 10, pp. 951–963, 2014.

[2] A. Selinger *et al.*, "Appearance-based facial recognition using visible and thermal imagery: a comparative study," Equinox Corp., Tech. Rep., 2006.

[3] P. Buddharaju *et al.*, "Physiology-based face recognition in the thermal infrared spectrum," *IEEE TPAMI*, 2007.

[4] I. Pavlidis *et al.*, "The imaging issue in an automatic face/disguise detection system," in *IEEE Computer Vision Beyond the Visible Spectrum*, 2000, pp. 15–24.

[5] J. Wilder *et al.*, "Comparison of visible and infra-red imagery for face recognition," in *FG 96*. IEEE, 1996.

[6] C. Goulart *et al.*, "Emotion analysis in children through facial emissivity of infrared thermal imaging," *PloS one*, vol. 14, no. 3, 2019.

[7] C. Puri *et al.*, "Stresscam: non-contact measurement of users' emotional states through thermal imaging," in *CHI'05*, 2005.

[8] C. Ordun *et al.*, "The use of AI for thermal emotion recognition: A review of problems and limitations in standard design and data," *AAAI*, 2020.

[9] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[10] T. Zhang *et al.*, "TV-gan: Generative adversarial network based thermal to visible face recognition," in *ICB*, 2018.

[11] H. Zhang *et al.*, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *IJCV*, vol. 127, no. 6-7, pp. 845–862, 2019.

[12] K. K. Babu *et al.*, "PCSGAN: Perceptual cyclic-synthesized generative adversarial networks for thermal and nir to visible image transformation," *arXiv preprint arXiv:2002.07082*, 2020.

[13] K. Lai and S. N. Yanushkevich, "Multi-metric evaluation of thermal-to-visual face recognition," in *EST*, 2019.

[14] Z. Wang *et al.*, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 8, 2018.

[15] C. Chen and A. Ross, "Matching thermal to visible face images using a semantic-guided generative adversarial network," in *FG*, 2019.

[16] W.-T. Chu *et al.*, "A parametric study of deep perceptual model on visible to thermal face recognition," in *VCIP*. IEEE, 2018, pp. 1–4.

[17] V. V. Kniaz *et al.*, "ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *ECCV*, 2018.

[18] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *NeurIPS*, 2017, pp. 465–476.

[19] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[20] W. Li *et al.*, "Unsupervised image-generation enhanced adaptation for object detection in thermal images," *arXiv preprint arXiv:2002.06770*, 2020.

[21] J.-Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[22] X. Mao *et al.*, "Least squares generative adversarial networks," in *ICCV*, 2017, pp. 2794–2802.

[23] Dec. [Online]. Available: https://github.com/soumith/ganhacks

[24] K. Mallat *et al.*, "A benchmark database of visible and thermal paired face images across multiple variations," in *BIOSIG*. IEEE, 2018, pp. 1–5.

[25] FLIR, "Thermal dataset for algorithm training," 2019. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[26] OTCBVS benchmark dataset collection. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/

[27] J. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery - ieee otcbvs ws series bench," *CVIU*, vol. 106, no. 2–3, pp. 162–182, 2007.

[28] Y. Choi *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.

[29] Apr 2018. [Online]. Available: https://github.com/eriklindernoren/PyTorch-GAN

[30] X. Kuang *et al.*, "Thermal infrared colorization via conditional generative adversarial network," *Infrared Physics & Technology*, 2020.

[31] J. Zhang *et al.*, "Feature-guided SAR-to-optical image translation," *IEEE Access*, 2020.

[32] Z. Xu *et al.*, "Adversarial uni-and multi-modal stream networks for multimodal image registration," in *MICCAI*, 2020, pp. 222–232.

[33] P. Perera *et al.*, "In2i: Unsupervised multi-image-to-image translation using generative adversarial networks," in *ICPR*, 2018, pp. 140–146.