Multimodal Deep Generative Models for Remote Medical Applications

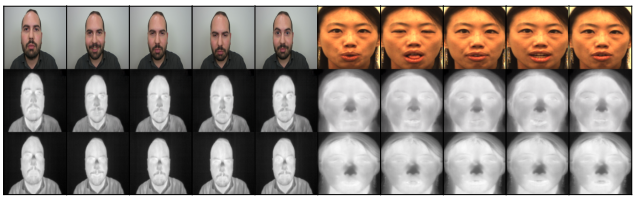
In the spirit of the 28th AAAI Doctoral Consortium theme, I am pleased to submit my thesis proposal which unifies several fields in AI - multimodal machine learning, Generative Adversarial Networks (GAN), affective computing, multispectral imagery, image registration and alignment, edge computing, and Federated Learning.

Introduction

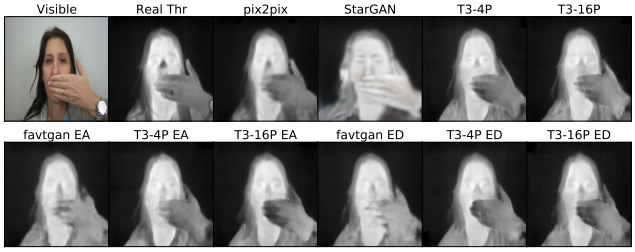
Telemedicine experienced a renaissance during the Covid pandemic. However, many patients and providers are still constrained to the facets of a simple web-meeting experience, offering few personalized analytics. We speculate that one reason for the lack of innovation is the restriction of onboard computer and smartphone sensors. In particular, the RGB camera is limited to the visible spectra. If telemedicine applications could take advantage of signals residing outside of this band, for example the long-wave infrared (LWIR) spectrum (8 - 14um), greater information about a patient's state of health could be obtained during a virtual consultation. Medical research in the well-studied field of thermal physiology provides this exact framework (Buddharaju et al. 2007; Pavlidis et al. 2007). Since LWIR detects heat emitted from the surface of the facial skin in complete darkness, signs of inflammation and anxiety can be visualized in a contact-free manner, all correlated to gold standard vital measures. Such information is hidden in the visible spectra, preventing physicians from assessing important clues about patient health. Unfortunately, the universal installation of LWIR sensors onto existing computers and smartphones is not feasible for numerous technical and economic reasons. **Motivated by these observations, my thesis contributes new algorithms such as conditional Generative Adversarial Networks (cGAN) and related deep learning methods, so that AI can stand in as a proxy for thermal hardware**.By taking an RGB image normally available on computer systems and translating it into a thermal image, these signs of patient stress and health can be visualized without needing a thermal sensor. My research is divided into three phases, described below: 1) Phase I - Visible-to-Thermal Facial GAN, 2) Phase II - Multimodal Data and Latent Factors, 3) Phase III - Multimodal Generative Capacity in FL.

Phase I - Visible-to-Thermal Facial GAN

Phase I asks, "What is the feasibility of translating visible faces into the thermal modality, and what are the associated technical challenges?" Before starting, I completed an overview of thermal AI limitations in facial emotion recognition (FER), in order to study the ethical impacts and biases in existing thermal datasets and studies (Ordun et al. 2020b). This study armed me with greater knowledge towards ethical impacts as this research proceeds into Phase III. Although thermal-to-visible (TV) translation has been applied successfully for person re-identification in law enforcement (Mallat et al. 2019; Zhang et al. 2019, 2018), translating in the opposite direction of visible-to-thermal (VT) is more challenging since it requires mapping high frequency edges



(a) Set of thermal faces generated by TFC-GAN.



(b) TFC-GAN vs. sample baseline methods (T: TFC-GAN)

Figure 1: **Sample thermal faces generated with my approach, TFC-GAN.**

to lower frequency, smooth thermal textures. Further, the subject identity can be lost, in addition to poor geometric alignment of the face. To address these challenges, I completed the development of the Facial-Visible-to-Thermal GAN (favtGAN) (Ordun et al. 2021), proving that auxiliary sensor labels can improve generated thermal face quality, by conditioning the generator. I extended favtGAN into the Thermal Face Contrastive GAN (TFC-GAN) in order to improve thermal image quality, resolution, and perceptual clarity. TFC-GAN incorporates contrastive losses for regional patch and temperature differentiation, with a relativistic loss and anti-aliasing to promote shift invariance to the input visible image. It achieves up to an 81.15% FID improvement on highly diverse, challenging visible-thermal facial datasets. I am currently incorporating TFC-GAN in an end-to-end pipeline to generate well-aligned thermal faces. The goal is to resolve severely misaligned facial pairs, a problem that plagues all VT/TV datasets since pre-paired sets are not available. Three contributions include: i) thermal-visible image alignment with affine rotation, facial landmarks, and feature descriptors, ii) registration via spatial transformation networks (Jaderberg, et al. 2015) to automatically learn a deformation grid, and iii) TFC-GAN with a Fourier domain loss module. Lastly, I am exploring denoising diffusion implicit models (DDIMs) (Song et al. 2020), popularized by the recent text-to-image Imagen and Stable Diffusion models, as an alternative thermal generation strategy as opposed to GAN. Phase I is 75% complete, with the aforementioned two works (i.e. pipeline, diffusion).

Phase II - Multimodal Data and Latent Factors

Phase II asks, "When applied to a real disease condition, such as cancer chronic pain, how does auxiliary information in the form of multimodal inputs such as facial landmarks, audio, and pain scores impact the translation of thermal faces?" Existing pain instruments fail to accurately re-

port chronic pain experienced by cancer patients, making it difficult for physicians to manage pain during the course of treatment. To this extent, I developed deep learning models for chronic cancer pain detection, using data from an ongoing clinical trial at the National Institutes of Health (NIH) entitled Intelligent Sight & Sound (ISS) across 29 patients (Ordun, Cha et al. 2022). This dataset is the first of its kind and consists of multimodal extracts drawn from patient narrative videos - facial landmarks, audio statistics, audio spectrograms, text transcripts, and self-reported pain scores. It varies markedly from existing pain datasets (Lucey et al. 2011) since chronic pain is not acute (e.g. stimulated from muscular contractions) and thereby patients display subdued and hard-to-detect facial emotions. As mentioned in the Introduction, thermal imagery offers valuable insights that are invisible on RGB images such as signs of inflammation and stress - common symptoms of cancer chronic pain patients. To this extent, the clinical trial has been collecting thermal video during in-clinic patient visits. In Phase I with the favt-GAN, I used auxiliary inputs in the form of thermal sensor class labels to improve the translation of thermal faces. Similarly, I intend to generate thermal faces using auxiliary inputs of facial landmarks, audio, text, and pain scores to not only improve the perceptual resolution but also to identify which features are most impactful in the generation of thermal specific regions. This is important since temperature on the eyes, nose, cheeks, and forehead each carry different physiological meaning. Phase II is 50% complete, where future work mentioned above (i.e. TFC-GAN, latent factors) will incorporate data from over 50 subjects.

Phase III - Multimodal Generative Capacity in FL

Phase III asks "What are the challenges in generative capacity in a Federated Learning (FL) model when deploying our multimodal translation GAN, across homo/heterogeneous devices?" A loose federation of client devices that can collaboratively learn a central model while preserving the privacy of local patient health data, makes FL (Kairouz et al. 2021) a promising option to explore for telemedicine. Recently, a handful of works have explored how to federate vanilla and image-translation GANs (Li et al. 2022; Xie et al. 2022). Generative capacity, or the ability to output high resolution and diverse images, is a shared problem that is so far being investigated through different weighting mechanisms and distribution architectures for the generator and discriminator. However, these works use datasets like MNIST which are far simpler than local distributions emanating from human faces of varying diversity, pose, and angle. This, coupled with the task of translating across optical spectra will undoubtedly reveal weaknesses in the existing approaches to generative capacity. As a result, Phase III intends to offer new innovations for these sets of challenges and will develop experiments across homogeneous (e.g. three NVIDIA Xaviers), and heterogenous edge devices (i.e. NVIDIA Nano, TX2, Xavier), serving as the local clients. Phase III is the least developed and will require the most research effort.

Table 1: **Progress Summary as of Sep. 2, 2022.** As first author of all completed works, I was responsible for writing the code, math, and paper, to include the entire AI pipeline (proc, algo, train, eval). Advisors provided strategy and recommendations. Co-authors of the NIH paper (Ordun, Cha et al. 2022) provided clinical protocol, IRB approval, ethical, and medical subject-matter expertise †(Ordun et al. 2020a,b, 2021) + TFC-GAN in review. ‡(Ordun, Cha et al. 2022) .

Phase	Works Completed	Works In-Progress	Est. % Remaining
I - Visible-to-Thermal Facial GAN	$4^{\dagger}$	2	25%
II - Multimodal Data & Latent Factors	$1^{\ddagger}$	2	50%
III - Multimodal Generative Capacity in FL	0	0	100%

References

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

Jaderberg, M.; ; et al. 2015. Spatial transformer networks. *NeurIPS*, 28.

Kairouz, P.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Li, W.; et al. 2022. IFL-GAN: Improved Federated Learning Generative Adversarial Network With Maximum Mean Discrepancy Model Aggregation. *IEEE TNNLS*.

Lucey, P.; et al. 2011. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *2011 IEEE ICAFGR*, 57–64. IEEE.

Mallat, K.; et al. 2019. Cross-spectrum thermal to visible face recognition based on cascaded image synthesis. In *ICB*, 1–8. IEEE.

Ordun, C.; Cha, A. N.; et al. 2022. Intelligent Sight and Sound: A Chronic Cancer Pain Dataset. *NeurIPS Datasets and Benchmarks*.

Ordun, C.; et al. 2020a. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *epiDAMIK 2020*.

Ordun, C.; et al. 2020b. The Use of AI for Thermal Emotion Recognition: A Review of Problems and Limitations in Standard Design and Data. *AAAI*.

Ordun, C.; et al. 2021. Generating Thermal Human Faces for Physiological Assessment Using Thermal Sensor Auxiliary Labels. *arXiv preprint arXiv:2106.08091*.

Pavlidis, I.; et al. 2007. Interacting with human physiology. 108(1-2): 150–170.

Song, J.; et al. 2020. Denoising diffusion implicit models. *ICLR*.

Xie, G.; et al. 2022. FedMed-GAN: Federated Multi-Modal Unsupervised Brain Image Synthesis. *arXiv preprint arXiv:2201.08953*.

Zhang, H.; et al. 2019. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *IJCV*, 127(6-7): 845–862.

Zhang, T.; et al. 2018. TV-gan: Generative adversarial network based thermal to visible face recognition. In *ICB*.