VISTA-MORPH: UNSUPERVISED IMAGE REGISTRATION OF VISIBLE-THERMAL FACIAL PAIRS

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

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

ABSTRACT

For a variety of biometric cross-spectral tasks, Visible-Thermal (VT) facial pairs are used. However, due to a lack of calibration in the lab, photographic capture between two different sensors leads to severely misaligned pairs that can lead to poor results for person re-identification and generative AI. To solve this problem, we introduce our approach for VT image registration called **Vista Morph**. Unlike existing VT facial registration that requires manual, hand-crafted features for pixel matching and/or a supervised thermal reference, Vista Morph is completely unsupervised without the need for a reference. By learning the affine matrix through a Vision Transformer (ViT)-based Spatial Transformer Network (STN) and Generative Adversarial Networks (GAN), Vista Morph successfully aligns facial and non-facial VT images. Our approach learns warps in Hard, No, and Low-light visual settings and is robust to geometric perturbations and erasure at test time. We conduct a downstream generative AI task to show that registering training data with Vista Morph improves subject identity of generated thermal faces when performing V2T image translation.

1. INTRODUCTION

Multiple Visible-Thermal (VT) facial datasets are available for biometric tasks like emotion recognition, thermal face recognition, and person re-identification [1]. Unfortunately, misalignment is introduced at the time of data capture when two sensors (a thermal and visible camera) are positioned at different angles and distances. Given increasing interest in generative AI, this inherent misalignment between cross-spectral faces can weaken image quality in generative tasks [2] such as Visible-to-Thermal (V2T) image translation, due to shift invariance [3]. Manual scaling, cropping, and alignment by hand is infeasible when dealing with thousands of images. Further, existing VT alignment methods rely on supervised feature matching [4, 5, 6, 7]. As a result, to rapidly register VT faces on multiple VT facial datasets of varying scale and distortion, we offer **Vis**ible **T**hermal Facial Morph (*Vista Morph*). Our model is the first unsupervised approach to register VT faces, to our knowledge, and does not rely on feature matching or a target reference. Vista Morph combines two Generative Adversarial Networks (GAN) [8] and a Spatial Transformer Network (STN) [9] that uses Vision Transformer (ViT) [10] for the first time, as the localization network. This contrasts from similar cross-spectral/multi-modal works [11, 12, 13, 14] that rely on a traditional CNN

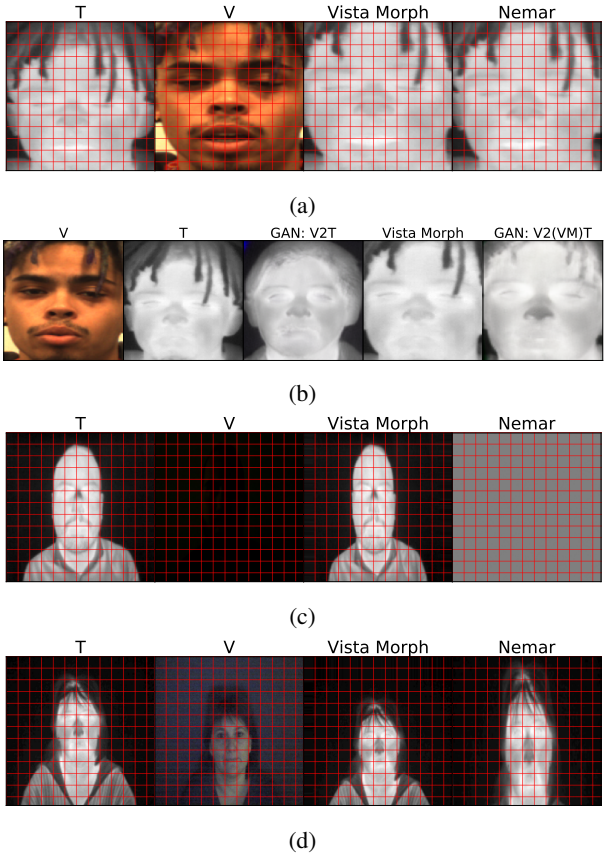


Fig. 1: **Vista Morph Registered Samples.** {a} Vista Morph aligns **T**hermal faces relative to the scale of the **V**isible face more accurately than the existing state-of-the-art (Nemar). {b} Registration improves generated thermal identity when using a GAN. {c,d} Vista Morph registers thermal faces in No- and Low-Light visible settings better than Nemar.

or U-NET [15] localization network for the STN. We select ViT because it applies self-attention across embedded image patches, making the spatial information fixed and preserved across layers of the network, whereas CNNs are less spatially discriminative [16, 17, 18]. Unlike traditional image registration methods, no similarity metric such as mean squared difference, normalized cross-correlation, or mutual information is optimized during training [19]. Only common GAN-based losses are learned. Further, Vista Morph integrates a Fourier Loss to learn how to align thermal images relative to No- and Low-light visible pairs by relying on the signal domain - so far unexplored in VT image registration but critical since Long-Wave Infrared (LWIR) sensors capture visible faces without the need for a light source.

We evaluate three VT facial datasets to align thermal faces relative to the visible face's geometry (T$\sim$V) and vice-versa (V$\sim$T). To examine generative image quality, we register each dataset with Vista Morph and train a conditional GAN [20] for the downstream task of Visible-to-Thermal (V2T) image translation. V2T image translation is increasingly researched for its value in person re-identification, thermal face recognition, and thermal physiology [21, 22, 23, 24, 25, 26, 27, 28]. We also train a Diffusion Model for the T2V generative task [29, 30]. We then use diagrams of the underlying facial vasculature, a thermal biometric asserted by [31], to analyze similarity between real and generated thermal identities. Our paper ends with a series of ablation studies on architectural settings and robustness. Our contributions are the following:

- The first unsupervised VT facial image registration called Vista Morph that uses ViT, for the first time, as a localization network in the STN framework.
- Registering pairs in challenging No- and Low-Light settings, a common scenario when using thermal sensors, by integrating a Fourier Loss in the Vista Morph model.
- Analyzing the identity of *generated* thermal faces from GANs by extracting vessel maps that visualize underlying thermal vasculature.
- Generalizability beyond faces with Vista Morph application to automated driving datasets and proven robustness against geometric transformations and erasure.

2. RELATED WORKS

Existing VT facial registration relies on feature-based matching such as edge maps, corner detection, intensity histograms, and SIFT features, or a supervised target, where these methods only evaluate on a single VT face dataset [4, 5, 6, 7]. Multimodal medical image registration methods such as DLIR [19] and Voxelmorph [32] are applied for CT and MRI imagery. However, while the images vary in density, they are still captured in the same optical spectra. This differs from our challenge where two images are obtained in different electromagnetic spectra altogether; the visible band (350 - 740 nm) and the LWIR band (8 - 15 μm). The task of unsupervised cross-spectral image translation is new. The most similar work to ours is the sentinel Nemar algorithm by Arar et al., [33] that first demonstrated unsupervised VT image registration on non-facial images. Since then, several similar Nemar-like approaches to CT/MRI images, remote sensing, and VT street scenes have been developed using varying translation and registration flows, new loss functions and/or fusion [34, 2, 35, 36]. No existing works tackle the challenge of non-rigid cross-spectral facial images since they contain abrupt changes and sudden deformations.

3. VISTA MORPH

We describe our approach, Vista Morph, by describing the training flows, loss functions to include the Fourier Loss for handling No- and Low-light settings, as well as using ViT as a novel localization network with a custom Multilayer Perceptron (MLP) for the regressor network of the STN [9] framework.

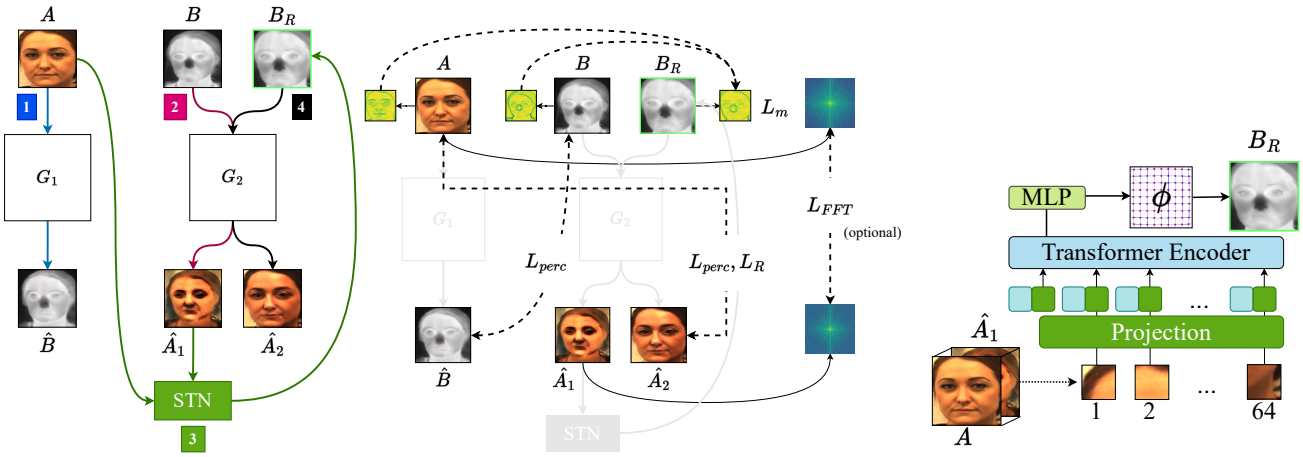
3.1. Generative Flows

Shown in Figure 2a, registration is trained using four flows, in an end-to-end fashion. First, the ground-truth visible image, A, is passed to the first Generator, G_1, that outputs the fake thermal image, $\hat{B}$. Second, the original thermal image, B, is passed to the second Generator, G_2, that outputs the first fake visible image, $\hat{A}_1$. Third, both A and $\hat{A}_1$ are used as inputs to the STN, in order to output the registered thermal image, B_R. In the fourth flow, B_R is passed back to G_2 in order to output $\hat{A}_2$. These flows force the STN to use a visible mapping of $\hat{B}$ translated into its thermal counterpart, $\hat{A}_1$, that preserves the geometry of the original thermal image as indicated in Figure 2c. By translating the visible to the thermal spectrum, the STN can now learn the scale between both modalities. Two GANs are used, where the generators, $G1$ and $G2$, are identical U-NETs with 5 encoder and 4 decoder modules with added BlurPool layers [3]. The discriminators, $D1$ and $D2$ are identical and comprise a traditional PatchGAN [37] architecture with a 16x16 patch, also incorporating BlurPool layers.

GAN Losses. The perceptual quality of the VT images is controlled using an LPIPS [38] perceptual loss, L_{perc}. Per Eq. 1, ϕ is the VGG-16 network and τ transforms network embeddings.

$$L_{perc} = \sum_n \tau^n(\phi^n(\hat{B}) - \phi^n(B)) + \sum_n \tau^n(\phi^n(\hat{A}) - \phi^n(\hat{A}_2)) \tag{1}$$

Recall that $\hat{A}_2$ is the visible image outputted by B_R after being warped by the STN. To enforce the alignment of



(a) Flows (b) Losses (c) STN with ViT

Fig. 2: **Training Overview.** Four flows trained in an end-to-end manner pass through a STN designed using a ViT. This process outputs the affine matrix (θ) used to warp the real thermal image (B) to its registered form (B_R) based on the intermediate fake visible image ($\hat{A}_2$). By translating $B \rightarrow \hat{A}_1$ using G_2, the STN can learn the scale between the visible and thermal spectra, using $\hat{A}_1$ as a proxy for B, all the while ensuring that B_R generates a visible face $\hat{A}_2$ similar to the original A.

B_R against the desired visible geometry of A, we set an L1 reconstruction loss shown in Eq. 2.

$$L_R = \|A - \hat{A}_2\|_1 \tag{2}$$

To control for structural similarity between B_R and A, we calculate morphological gradients for B_R, B, and A, and apply a triplet loss in Eq. 3.

$$L_m = \frac{1}{K}\sum_{k=1}^{K} \max\{d(B_R, A) - d(B_R, B) + 1, 0\} \tag{3}$$

Finally, for datasets with Low- or No-Light visible imagery, we add a **Fourier Loss** to learn the signal domain, as opposed to only spatial domain. The L_{FFT} shown in Eq. 6 is the L1 loss of the amplitude in Eq. 4 and phase in Eq. 5 of A and $\hat{A}1$.

$$L_{amp} = \|(|\mathcal{F}\{A\}_{u,v}|) - (|\mathcal{F}\{\hat{A}_1\}_{u,v}|)\|_1 \tag{4}$$

$$L_{pha} = \|(\angle\mathcal{F}\{A\}_{u,v}) - (\angle\mathcal{F}\{\hat{A}_1\}_{u,v})\|_1 \tag{5}$$

$$L_{FFT} = L_{amp}(A, \hat{A}_1) + L_{pha}(A, \hat{A}_1) \tag{6}$$

We train the GAN (L_{GAN}) using a relativistic adversarial loss [39] leading to the total Generator Loss, L_G shown in Equation 7. The total Discriminator loss L_D, is an average of the real and fake discriminator losses which are both relativistic. The total training objective is shown in Equation 8.

$$L_G = L_{GAN} + L_{perc} + L_R + L_m \tag{7}$$

$$G^* = \arg\min_G \max_D L_G + L_D \tag{8}$$

3.2. Registration Network

The registration network is a STN [9]. STN is not a model in itself, but rather, a framework where any differentiable function (e.g. neural network) can be used as the localization network. As a result, we use a 12-Layer ViT [10] as the localization network to extract features between the visible (aligned) and thermal (non-aligned) image, and add a 4-Layer MLP as the regressor network. Shown in Figure 2c, the concatenated input of $(A, \hat{A}_1)$ are passed to the ViT using a patch size of 64. The MLP consists of the following architecture: Linear (17*768,1024)-ReLU-Linear (1024,512)-Relu-Linear (512,256)-Sigmoid-Linear (256,6), outputting an affine matrix (ϕ) of size 6. Using a sigmoid activation function is important for the regressor since it outputs values between [-1,1] in the range of ϕ [9]. The ϕ is calculated for each (A, A_1) pair which represents the 2D flow field (sampling grid), given a batch of affine matrices. Using an affine transformation, the STN computes the registered output, B_R sampling the pixel locations from the field. The STN is trained jointly with $G1$ and $G2$.

4. EXPERIMENTS

4.1. Datasets

In this section, we evaluate our approach on three VT paired facial datasets: Carl [40], Army Research Lab (ARL) Devcom Dataset [41], and Eurecom [42]. To test our approach on a non-facial domain, we use the FLIR Advanced Driver Assistance Systems (ADAS) dataset [43]. For each dataset, we conduct a minimal amount of preprocessing. For example, with the Devcom dataset, we use the forward-facing

"baseline" and "expression" protocols and ignore the thermal bounding box metadata supplied for alignment. Instead, we use a FaceNet MTCNN [44, 45] to detect and crop visible faces, and apply a series of binary thresholding operations to crop the thermal face image away from its background. The results lead to a misaligned set of VT facial pairs with varying degrees of warp. We select a random 5% sample from our misaligned Devcom dataset (56,205 training pairs) due to compute limitations. We use the entire Eurecom, Carl, and FLIR ADAS datasets, with details in Table 1.

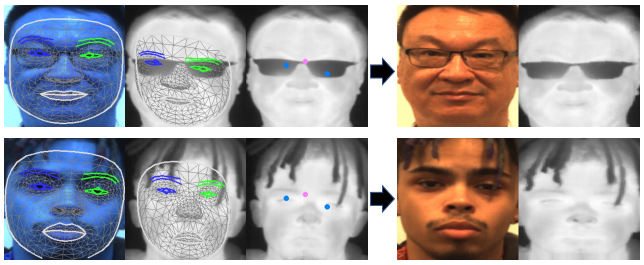


Fig. 3: **Manual Alignment.** Existing facial keypoint models work for visible spectra, but fail to consistently detect thermal facial landmarks. If available, manually estimating the affine matrix leads to misaligned pairs.

Dataset	Direction	Train Subj	Test Subj	Train Pairs
Devcom 5%	T ∼V	376	74	2810
Carl	T ∼V	32	9	1920
Eurecom (Vis)	V ∼T	45	5	945
ADAS	T ∼V	N/A	N/A	7,521
Dataset	**Direction**	**Test Pairs**	**Positions**	**Lighting**
Devcom 5%	T ∼V	247	F	H
Carl	T ∼V	540	F	H,L
Eurecom (Vis)	V ∼T	105	F, U, D, S	H, L, N
ADAS	T ∼V	1,159	N/A	H, L, N

Table 1: **Datasets**. Positions: **F**rontal, **U**p, **D**own, **S**ideways. Lighting: **H**ard, **L**ow, **N**one.

4.2. Baselines

First, we explore manual approaches using existing facial landmark algorithms in order to capture right and left eye coordinates implicit for affine matrix estimation. One algorithm is the Google MediaPipe [44] Face Mesh Active Appearance Model (AAM) based on 3D Morphable Models [46] that implements a Multi-task Cascaded Convolutional Network (MTCNN) model [45] using an InceptionResnetV1 model pre-trained on VGGFace2. Each detected face returns an array of 468 points for three coordinates used as keypoints. For the Devcom dataset, the AAM fails to detect landmarks for 40% of the thermal faces, thereby leaving only 60% of facial pairs usable. For the remaining pairs, the scale of the desired

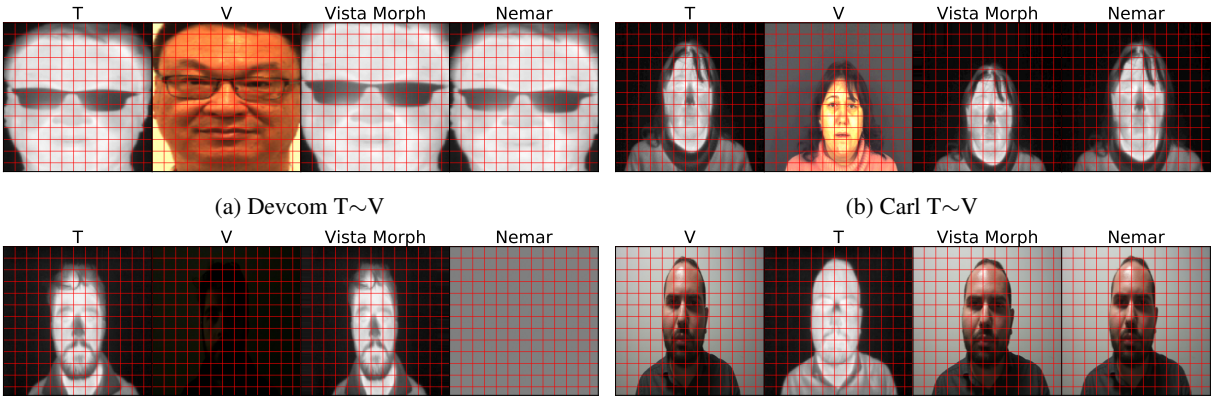
			SSIM Edges ↑		
Dataset	**Alig.**	**No Reg.**	**VM**	**Ne**	**Diff.**
Devcom	T ∼V	0.878	**0.898**	0.894	**0.5%**
Carl	T ∼V	0.804	**0.861**	0.820	**5.1%**
Eurecom (Thr)	T ∼V	0.833	**0.837**	0.830	**0.8%**
			NCC Edges ↑		
Dataset	**Alig.**	**No Reg.**	**VM**	**Ne**	**Diff.**
Devcom	T ∼V	0.013	**0.147**	0.079	**86.6%**
Carl	T ∼V	0.191	**0.411**	0.201	**105.0%**
Eurecom (Thr)	T ∼V	0.285	**0.329**	0.320	**2.9%**
			Mutual Information (MI) ↑		
Dataset	**Alig.**	**No Reg.**	**VM**	**Ne**	**Diff.**
Devcom	T ∼V	0.246	0.269	**0.270**	-0.2%
Carl	T ∼V	0.473	**0.649**	0.512	**26.8%**
Eurecom (Thr)	T ∼V	0.493	**0.526**	0.474	**11.0%**

Table 2: **Registration Results.** Best scores are in **bold**. VM: Vista Morph, Ne: Nemar, Diff: Relative percentage change of VM over Ne. Bold Diff. scores indicate VM improvement over Ne.

registered image must be determined by estimating ratios between eye distances among the current and target image. This enables retrieval of geometric parameters to compute the affine matrix. No singular set of parameters can address all variations of warp, and as a result, they must be calculated manually for every VT pair. Samples results shown in Figure 3 demonstrate the imperfection of this manual approach. Since the manual pipeline is not feasible for multiple VT datasets that contain a variable level of warp, scale, and abrupt changes, we compare our approach against the Nemar model [33] as the closest to Vista Morph's unsupervised approach.

4.3. Experimental Protocol

First, we train Vista Morph and Nemar in two directions to: (1) align the thermal face relative to the visible face (T∼V), and (2) for the Eurecom dataset, align visible relative to thermal (V∼T). As a result, we perform a total of four VT facial registration experiments: (1) Devcom (T∼V), (2) Carl (T∼V), (3) Eurecom (V∼T), (4) Eurecom (T∼V). For Carl and Eurecom, we trained the T∼V alignment, using L_{FFT} loss due to No- and Low-Light images. *Second*, we register the entire dataset (train, test) using Vista Morph and Nemar. *Third*, we conduct a downstream generative task, for image-to-image translation using the VTF-GAN [20], a conditional GAN specifically designed for Visible-to-Thermal (V2T) facial image-to-image translation. We generate both visible and thermal faces using both the unregistered original data and the Vista Morph registered pairs.



(a) Devcom T∼V

(b) Carl T∼V

(c) Eurecom T∼V

(d) V∼T

Fig. 4: **Registration Samples**. Row 1: Devcom T∼V, Row 2: Carl T∼V, Row 3: Eurecom T∼V, Row 4: Eurecom V∼T.

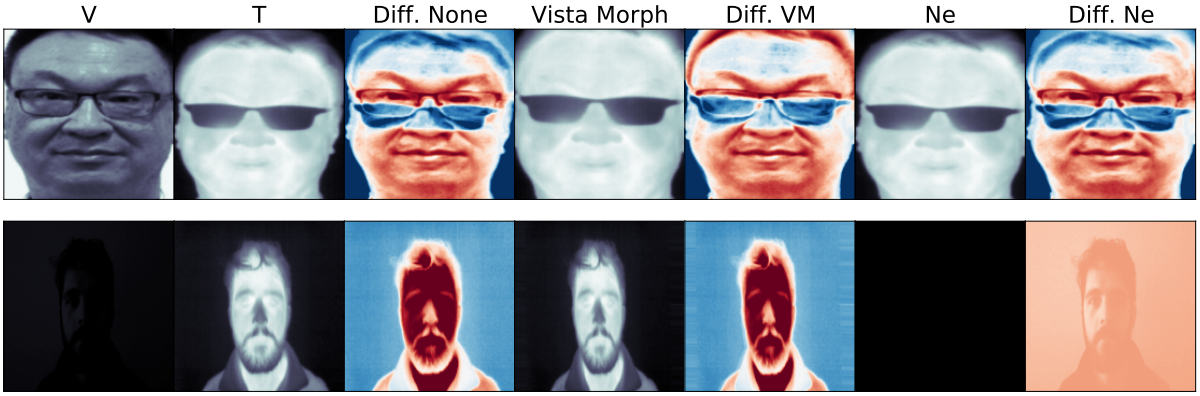


Fig. 5: **Difference Maps Before and After Registration**. The "Diff.VM" heatmap displays the most consistent overlap between red (visible) and blue (thermal) color maps. Notice the precise placement of sunglasses, compared to "Diff.Ne". Vista Morph registers the thermal image even when the visible image is No-Light, making apparent the "Diff.VM" heatmap. The "Diff.Ne" heatmap only shows the visible image in red because no thermal image was found/registered. V: Visible, T: Thermal, Diff. None (T-V), Diff. VM (Vista Morph - V), Diff. Ne (Ne - V)

4.4. Evaluation

To score registration results, we use Structural Similarity Index Measure (SSIM) and Normalized Cross Correlation (NCC) of the edge maps (e.g. morphological gradients of the visible and thermal images), in addition to Mutual Information (MI) [47] between both spectra. For generative results, we score with Frechet Inception distance (FID) [48] and LPIPS [38]. Lastly, we analyze a sample of Devcom generated thermal faces for retention of identity through facial vasculature maps.

4.5. Implementation

For registration experiments, we train our model and the baseline using PyTorch, to 50 epochs with a batch size of 32. For generative experiments, we train the VTF-GAN and VTF-Diff from scratch, to 200 epochs with a batch size of 32. For all experiments, we use automatic mixed precision and parallel training on two RTX-8000 GPUs. Training Vista Morph is fast, where registration is learned approximately 1 hr. on the Devcom dataset.

5. RESULTS

5.1. Registration

In Table 2, Vista Morph outperforms the Nemar baseline for T∼V alignment on all datasets. For Mutual Information, Vista Morph is comparable to Nemar with only a marginal difference (0.260, 0.279). Vista Morph shows the greatest registration gains with the Carl dataset for all three metrics (5.1%, 105.0%, 26.8%), which exhibits several Low-Light settings. Similarly, the T∼V alignment for Eurecom improves using Vista Morph, as this dataset includes Low- and

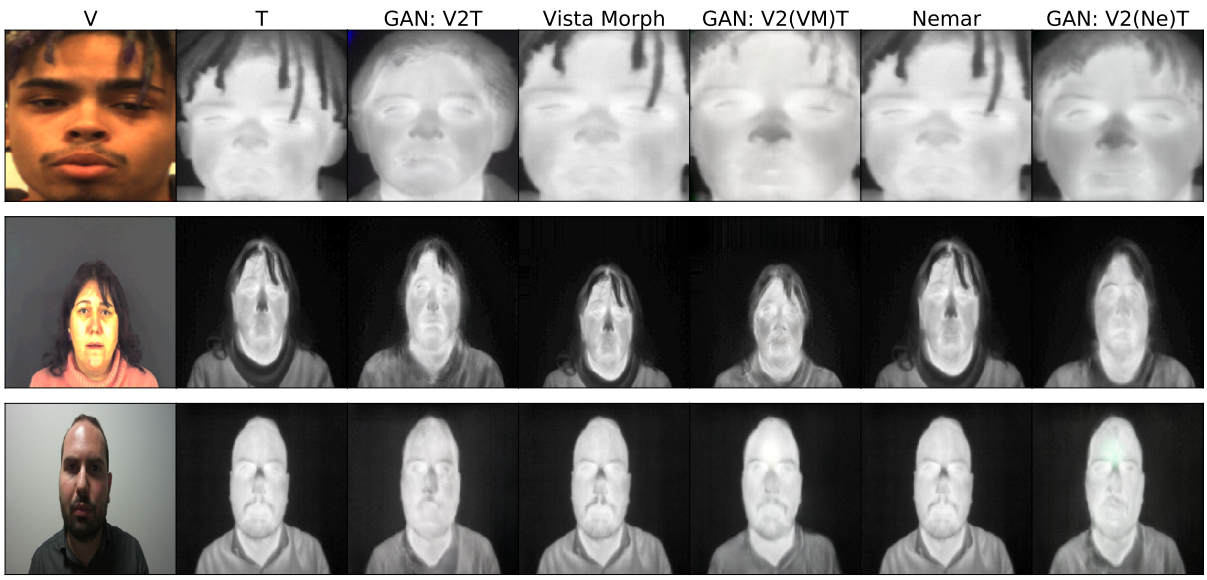


Fig. 6: **Generated Thermal Faces for V2T Image Translation.** The first two cols. are the original V and T faces. "GAN: V2T" is the generated thermal face from VTF-GAN using unregistered pairs. "Vista Morph" (VM) is the registered thermal face. "GAN: V2(VM)T" is the generated thermal face when VTF-GAN is trained on VM-registered pairs. "Nemar" is the registered thermal face using the Nemar baseline. "GAN: V2(Ne)T" is the generated thermal face when VTF-GAN is trained on Nemar-registered pairs. Notice that for all datasets, the quality and identity of the "GAN: V2(VM)T" has less distortion, artifacts, and more qualitative similarity to the registered thermal face.

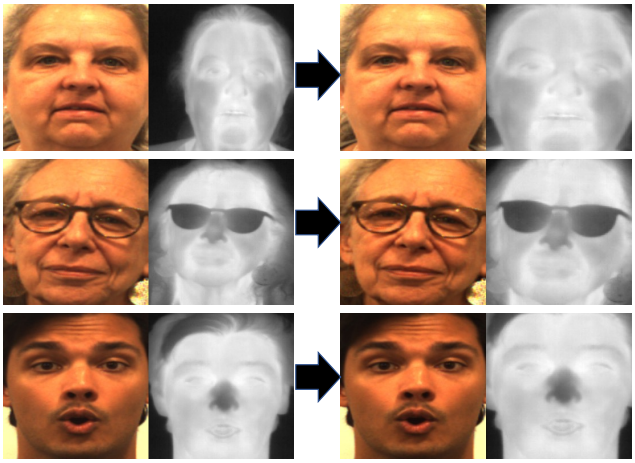


Fig. 7: **Vista Morph Results under Varying Warps and Scale.** Even when thermal faces are grossly misaligned (i.e. off-center, tilted with glasses), Vista Morph can learn the affine matrices to register $T \sim V$.

No-Light visual images. We show in Figure 4c that Nemar fails to register T∼V, since the visible face is captured in a No-Light setting. The T∼V results in Figure 4a and 4b, show the precise alignment of Vista Morph despite hair texture and differences in scale and height.

An intuitive view are difference maps shown in Figure 5. These plots visualize the shift of pixels between the registered and original images. For example, the top row of Figure 5 shows the difference without registration where the thermal glasses (blue) are not aligned to the visible eyes (red). After registration with Vista Morph, the glasses are superimposed on the eyes, whereas the baseline still demonstrates misalignment. Most noticeable is the Eurecom T∼V on the second to last row which shows no difference, only a light orange Nemar plot. This is because no thermal image was registered. Carl plots show blue ringing effects and shadows around the Nemar difference map indicating the imperfect alignment to the visible face's scale.

5.2. Generation

Table 3 shows results for the generative Visible-to-Thermal (V2T) image-to-image translation tasks using registered and unregistered training data. In all cases, scores improve significantly after VT pairs are registered. For FID scores, Nemar achieves slightly lower scores than Vista Morph yet LPIPS scores are 6.8% lower for Vista Morph. For the Carl and Eurecom V2T translations, their thermal faces show a -13.9% and -12.4% decrease in FID, and, -13.1% and -12.4% decrease in LPIPS, respectively. Sample generated images for the V2T direction are provided in Figure 6. Upon qualitative inspection, when using unregistered data or Nemar-registered pairs, the generated thermal faces ("GAN:V2T", "GAN:V2(Ne)T") introduce more artifacts, less texture and

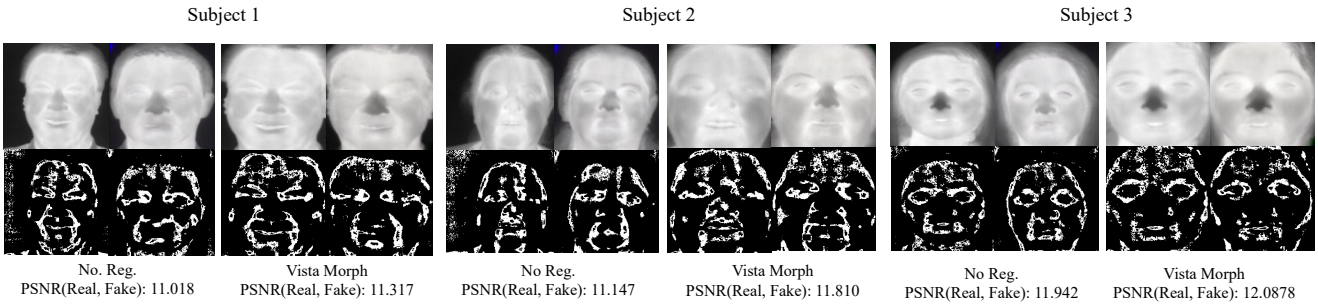
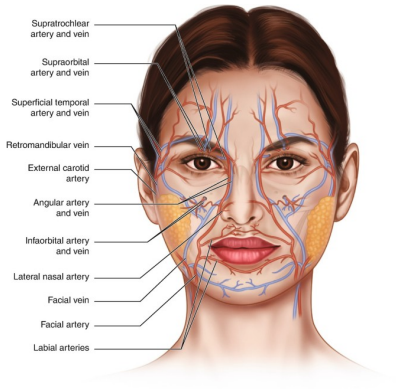
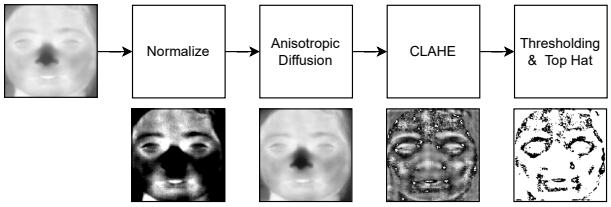


Fig. 8: **Thermal Vessels Before and After Vista Morph Registration.** Identity is more similar when VTF-GAN is trained on Vista Morph registered data.



(a) Facial anatomy of the face (Google Images).



(b) Vessel Extraction Processing

Fig. 9: **Facial anatomy and Preprocessing for Vessel Maps.**

consistency, and similarity to their ground truth thermal faces. The "GAN:V2(VM)T" faces retain distribution of pixel color (e.g. thereby thermal temperature), perceptual clarity, and hair texture which is important for maintaining minority and female identities. Additional results are in the Supplementary Materials.

5.2.1. Generated Identity Analysis

To visualize how Vista Morph registered data improves the generation of subject identity, we turn to facial vascular network extraction as defined by [31]. Building on biometric work stemming from retrieval of venous structure in palms and wrists [49, 50], Buddharaju, et al. [31] propose thermal

		FID ↓			
Dataset	**Translation**	**No Reg.**	**VM**	**Ne**	**Diff.**
Devcom	V2T	60.357	50.740	**50.338**	0.8%
Carl	V2T	52.865	**44.765**	51.972	**-13.9%**
Eurecom (Thr)	V2T	70.221	**69.893**	79.810	**-12.4%**
Dataset	**Translation**		**LPIPS ↓**		
Devcom	V2T	0.279	**0.218**	0.234	**-6.8%**
Carl	V2T	0.190	**0.168**	0.193	**-13.1%**
Eurecom (Thr)	V2T	0.157	**0.144**	0.165	**-12.4%**

Table 3: **Generation Results for V2T Image Translation.** Best FID and LPIPS scores in **bold**. VM (Vista Morph), Ne (Nemar), Diff: Relative perc. change of VM over Ne.

.

vasculature as a unique biometric that can be extracted from thermal faces through basic image processing shown in Figure 9b: anisotropic diffusion [51] to remove noise and enhance sigmoid edges, followed by CLAHE (Contrast Limited Adaptive Histogram Equalization) [52], and then finally top hat segmentation. We use samples from the Devcom dataset as a test bed since faces are close-up with minimal apparel. To measure similarity between the generated and real identity, we calculate the Peak Signal to Noise Ratio (PSNR) [53] between vessel diagrams. Subjects in Figure 8 show the facial vein, labial arteries (mouth and nose), angular artery and vein (eyes), superficial temporal artery and vein (edge of face), and supraorbital artery and vein (forehead). PSNR between the GAN is trained on registered pairs. For example, Subject 1 shows a PSNR of 11.018 of vessel maps before registration. Notice how "G", the generated image trained on unregistered data shows an identity dissimilar to its respective ground truth, "T". The PSNR of their respective vessel maps is 11.018. When Vista Morph registers T∼V in "RT", the PSNR between vessel maps increases to 11.317, indicating that the generated subject's identity is more similar to the ground truth, when registered. Similar evidence can be seen in Subjects 2 and 3 that display extreme variance in scale as well as head tilt between VT pairs.

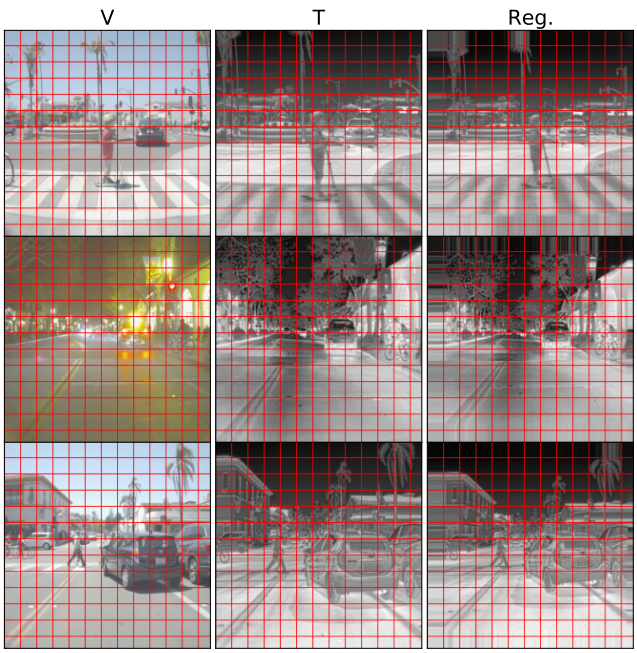


Fig. 10: **Vista Morph registers Non-Facial Data - ADAS Street Scenes**. For scenes of pedestrians in hard light and low-light, with different scales of objects, Vista Morph aligns $T \sim V$), shown in the last column.

Ablation Exp.	SSIM Edges	NCC Edges	MI
U-NET, Patch=64	0.880	-0.006	0.218
Patch = 16	0.894	0.068	0.230
Patch = 32	N/A	-0.054	0.075
Baseline	**0.898**	**0.147**	**0.269**
Patch = 128	0.897	0.140	0.262

Table 4: **Ablation Study for Devcom T∼V**. Results are compared to the Vista Morph Baseline (ViT, Patch=64).

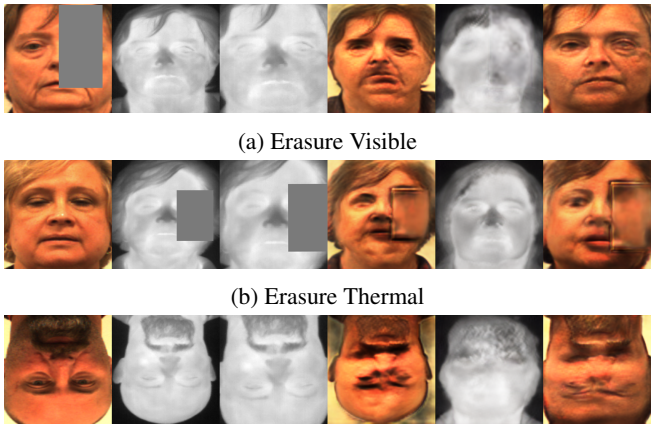
6. ABLATION STUDIES

6.1. Architecture and Patch Size

We conducted a brief ablation study using the Devcom dataset (T∼V). When using the traditional U-NET for the STN, all scores decrease significantly when compared to our implementation (SSIM: -2.02%, NCC: -104.15%, MI: -19.03%). Patch=32 size led to the worst results where the affine matrix could not be estimated. For facial images, more patches (e.g. increase of patch size) which preserve positional information, empirically leads to better registration results.

6.2. Non-Facial Domains

Vista Morph successfully registers non facial pairs, namely the FLIR ADAS street scenes dataset. We train a "deeper" STN regressor by adding two more linear layers to the baseline. We believe this enables finer focus of object features such as persons, stop signs, pedestrian crossings, and vehicles. Further, we incorporate Fourier Loss since ADAS includes several No- and Low-Light settings. Further, we find that 32 patches and removing the morphological loss also improves image quality. Registration scores for Vista Morph are 0.672 SSIM, 0.318 NCC, 0.607 MI compared to no registration at 0.627 SSIM, 0.250 NCC, 0.516 MI.



(a) Erasure Visible

(b) Erasure Thermal

(c) Vertical Flip

Fig. 11: **Vista Morph robustness against random geometric transformations and erasure**. Columns left to right: Real Visible, Real Thermal, Registered Thermal, A1, Fake Thermal, A2

6.3. Robustness

To illustrate Vista Morph's robustness, we test registration with geometric transformations and random erasure at inference time on the input pairs. Despite the permutations, Vista Morph successfully registers the T∼V. In Figure 11a, the thermal image registers to the scale of the visible input regardless of erasure. Figure 11b shows the expected behavior, that the entire thermal image is registered. Here, the generated $\hat{A}_1$ and $\hat{A}_2$ images demonstrate the translation of the erasure pixels. Figure 11c shows that when vertically flipped, Vista Morph will register the thermal image accordingly with respect to the visible geometry.

7. LIMITATIONS

Unfortunately, for $V \sim T$ registration, Vista Morph underperforms. To explore how registration effects a different generative model, we trained a conditional Denoising Diffusion Probabilistic Model (DDPM) [29, 30] called VTF-Diff [20] on registered and non-registered images in the T2V direction. GAN results improve with registration (Nemar), but registration does not improve the Diffusion results. Although the

	SSIM ↑	NCC ↑	MI ↑	VTF-GAN ↓	VTF-Diff ↓
No Reg.	0.833	0.285	0.493	111.142	**67.078**
Nemar	**0.843**	0.217	**0.556**	**83.346**	81.398
Vista Morph	0.832	**0.281**	0.530	87.887	86.102

Table 5: **Limitations of $V \sim T$ Registration and T2V Generation.** Vista Morph underperforms for visible face registration and the T2V translation when using the Vista Morph registered data. FID scores are provided for the VTF-GAN and -Diff model results. SSIM, NCC, MI are the same as Table 2, abbreviated for space.

generated diffusion results look geometrically and perceptually similar, there are differences in skin color, eye detail, and clothing. Further tests are needed beyond the Eurecom dataset.

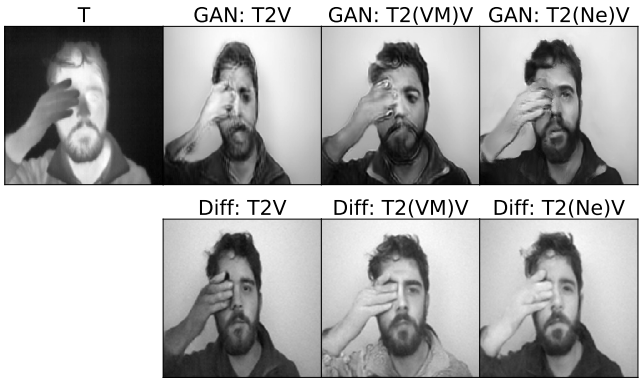


Fig. 12: **Generated T2V Image Translation Sample.** Training Vista Morph $V \sim T$ aligned images does not improve GAN results. Diffusion results quantitatively perform better without registration, despite looking qualitatively similar.

8. CONCLUSION

We present Vista Morph, the first unsupervised VT facial registration model that aligns facial pairs without a reference or feature matching. We evaluate three VT facial datasets leading to significantly improved registration results over state-of-the-art methods. Further, we show that image quality from a generative Visible-to-Thermal image translation task improves with regards to perceptual clarity and identity when training a GAN on Vista Morph registered pairs. We support our findings with thermal vessel maps and demonstrate Vista Morph can register non-facial domains. Future work includes assessment of generated thermal faces by thermal specialists and user studies, and finer investigation into the consistency of results for diverse demographic samples.

9. REFERENCES

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

[2] Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al., "Breaking the dilemma of medical image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1964–1978, 2021.

[3] Richard Zhang, "Making convolutional networks shift-invariant again," in *ICML*. PMLR, 2019, pp. 7324–7334.

[4] Seong G Kong, Jingu Heo, Faysal Boughorbel, Yue Zheng, Besma R Abidi, Andreas Koschan, Mingzhong Yi, and Mongi A Abidi, "Multiscale fusion of visible and thermal ir images for illumination-invariant face recognition," *International Journal of Computer Vision*, vol. 71, pp. 215–233, 2007.

[5] Palani Thanaraj Krishnan, Parvathavarthini Balasubramanian, Vijay Jeyakumar, Shriraam Mahadevan, and Alex Noel Joseph Raj, "Intensity matching through saliency maps for thermal and visible image registration for face detection applications," *The Visual Computer*, pp. 1–14, 2022.

[6] Jiayi Ma, Ji Zhao, Yong Ma, and Jinwen Tian, "Non-rigid visible and infrared face registration via regularized gaussian fields criterion," *Pattern Recognition*, vol. 48, no. 3, pp. 772–784, 2015.

[7] Lin Sun and Zengwei Zheng, "Thermal-to-visible face alignment on edge map," *IEEE Access*, vol. 5, pp. 11215–11227, 2017.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] Max Jaderberg, , et al., "Spatial transformer networks," *NeurIPS*, vol. 28, 2015.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] Giovana Augusta Benvenuto, Marilaine Colnago, and Wallace Casaca, "Unsupervised deep learning network for deformable fundus image registration," in *ICASSP*

2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 1281–1285.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[13] Roshan Reddy Upendra, Richard Simon, and Cristian A Linte, "Joint deep learning framework for image registration and segmentation of late gadolinium enhanced mri and cine cardiac mri," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. SPIE, 2021, vol. 11598, pp. 96–103.

[14] Yi Zhang, Shizhou Zhang, Ying Li, and Yanning Zhang, "Single-and cross-modality near duplicate image pairs detection via spatial transformer comparing cnn," *Sensors*, vol. 21, no. 1, pp. 255, 2021.

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

[16] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2021, pp. 01–06.

[17] Lam Phan, Hiep Thi Hong Nguyen, Harikrishna Warrier, and Yogesh Gupta, "Patch embedding as local features: Unifying deep local and global features via vision transformer for image retrieval," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 2527–2544.

[18] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, 2021.

[19] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52, pp. 128–143, 2019.

[20] Catherine Ordun, Edward Raff, and Sanjay Purushotham, "When visible-to-thermal facial gan beats conditional diffusion," *arXiv preprint arXiv:2302.09395*, 2023.

[21] Cunjian Chen et al., "Matching thermal to visible face images using a semantic-guided generative adversarial network," in *FG*, 2019.

[22] Gabriel Hermosilla, Diego-Ignacio Henríquez Tapia, Héctor Allende-Cid, Gonzalo Farías Castro, and Esteban Vera, "Thermal face generation using stylegan," *IEEE Access*, vol. 9, pp. 80511–80523, 2021.

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

[24] A Merla, L Di Donato, PM Rossini, and GL Romani, "Emotion detection through functional infrared imaging: preliminary results," *Biomedizinische Technick*, vol. 48, no. 2, pp. 284–286, 2004.

[25] Nithin Gopalakrishnan Nair et al., "T2v-ddpm: Thermal to visible face translation using denoising diffusion probabilistic models," *arXiv preprint arXiv:2209.08814*, 2022.

[26] Ioannis Pavlidis et al., "Interacting with human physiology," vol. 108, no. 1-2, pp. 150–170, 2007.

[27] Ioannis 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.

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

[29] Jonathan Ho et al., "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[30] Alexander Quinn Nichol et al., "Improved denoising diffusion probabilistic models," in *ICML*. PMLR, 2021, pp. 8162–8171.

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

[32] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[33] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13410–13419.

[34] Zekang Chen, Jia Wei, and Rui Li, "Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation," *arXiv preprint arXiv:2204.13656*, 2022.

[35] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," *arXiv preprint arXiv:2205.11876*, 2022.

[36] Yingxiao Xu, Jun Li, Chun Du, and Hao Chen, "Nbrnet: A nonrigid bidirectional registration network for multitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

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

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[39] Alexia Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standardGAN," *arXiv preprint arXiv:1807.00734*, 2018.

[40] Virginia Espinosa-Duró, Marcos Faundez-Zanuy, and Jiří Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognitive Computation*, vol. 5, no. 1, pp. 119–135, 2013.

[41] Domenick Poster et al., "A large-scale, time-synchronized visible and thermal face dataset," in *WACV*, 2021, pp. 1559–1568.

[42] Khawla Mallat and Jean-Luc Dugelay, "Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[43] FLIR, "Thermal dataset for algorithm training," 2019.

[44] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al., "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[46] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," *arXiv preprint arXiv:1907.06724*, 2019.

[47] Jeffrey P Kern and Marios S Pattichis, "Robust multispectral image registration using mutual-information models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1494–1505, 2007.

[48] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[49] Manu Francis, Akhil Jose, and KK Avinashe, "A novel technique for forearm blood vein detection and enhancement," *Biomedical Research*, vol. 28, no. 7, pp. 2913–2919, 2017.

[50] Yi-Bo Zhang, Qin Li, Jane You, and Prabir Bhattacharya, "Palm vein extraction and matching for personal authentication," in *Advances in Visual Information Systems: 9th International Conference, VISUAL 2007 Shanghai, China, June 28-29, 2007 Revised Selected Papers 9*. Springer, 2007, pp. 154–164.

[51] Pietro Perona and Jitendra Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[52] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[53] Alain Hore and Djemel Ziou, "Image quality metrics: Psnr vs. ssim," in *ICPR*. IEEE, 2010, pp. 2366–2369.