

T2V-DDPM: Thermal to Visible Face Translation using Denoising Diffusion Probabilistic Models

Nithin Gopalakrishnan Nair and Vishal M. Patel

Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA

Abstract—Modern-day surveillance systems perform person recognition using deep learning-based face verification networks. Most state-of-the-art facial verification systems are trained using visible spectrum images. But, acquiring images in the visible spectrum is impractical in scenarios of low-light and nighttime conditions, and often images are captured in an alternate domain such as the thermal infrared domain. Facial verification in thermal images is often performed after retrieving the corresponding visible domain images. This is a well-established problem often known as the Thermal-to-Visible (T2V) image translation. In this paper, we propose a Denoising Diffusion Probabilistic Model (DDPM) based solution for T2V translation specifically for facial images. During training, the model learns the conditional distribution of visible facial images given their corresponding thermal image through the diffusion process. During inference, the visible domain image is obtained by starting from Gaussian noise and performing denoising repeatedly. The existing inference process for DDPMs is stochastic and time-consuming. Hence, we propose a novel inference strategy for speeding up the inference time of DDPMs, specifically for the problem of T2V image translation. We achieve the state-of-the-art results on multiple datasets. The code and pretrained models are publically available at <http://github.com/Nithin-GK/T2V-DDPM>

I. INTRODUCTION

Many surveillance systems include sensors to capture images at multiple wavelengths to accommodate day and nighttime settings. Under low-light settings, the images captured by visible spectrum cameras fail to capture visual details in the scene due to the increasing amount of additive Poisson noise on the captured image. Hence, the use of infrared cameras to capture an additional thermal image is quite prominent. The images captured by these surveillance systems are further used for person recognition using facial recognition algorithms, and this falls under the broad area of research called Heterogeneous Face Recognition (HFR) [12], [13], [14]. Deep Convolutional Neural Network (CNN) based algorithms have produced state-of-the-art results for facial recognition with almost perfect accuracy in multiple benchmarks [1], [18].

Although the existing facial verification algorithms work remarkably well during daytime, the performance of these algorithms fall drastically under low-light settings, this is due to the large domain discrepancy between thermal and visible images. One solution is to retrain a network for facial recognition using just thermal images. But this is impractical since the normal CNN-based facial recognition networks [1] often require large amounts of face images to obtain

good performance and there does not exist many publically available large scale thermal images to train such networks. Hence the problem of HFR is very relevant and has multiple practical implications.

The recent rise of conditional generative models [5] has enabled an alternate approach to address this problem through a two-step process. First, the problem is simplified to an image-to-image translation problem from thermal to visible domains. After the corresponding visible images are obtained, facial recognition algorithms can be directly applied to them. Multiple works utilize GANs for tackling this problem [16], [2], [8]. Conditional GANs attempt to learn the translation from the thermal domain to the visible domain through a min-max approach. Although with sufficient training, GAN-based T2V translation networks can produce good results, training of GANs is a tedious process and may result in phenomena like mode collapse. Also, given limited data, the convergence of GANs for accurate T2V translation is not guaranteed.

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [7] have gained a significant attention due to their ability to generate high-quality images. Like Variational Autoencoders [11], DDPMs attempt to learn the variational lower bound of the log-likelihood of the data distribution. DDPMs have already beaten GANs in the task of image generation. Multiple methods have proposed DDPMs for low-level vision tasks like image super resolution, colorization, deblurring, and denoising [22], [21], [10], [15]. These methods have performed much better in terms of quality of the output produced based on realness metrics like Fréchet Inception Distance (FID) [6] and structural similarity metrics like Structural Similarity Index (SSIM). Moreover, there are works connecting the ability of DDPM to learn the conditional distribution to the optimal transport theory of the most efficient transformation [24]. Despite its immense potential in modeling conditional distributions, there are no works yet published for the problem of T2V face image translation. In this paper, we propose a solution for T2V face translation using DDPMs. Since the model attempts to learn the transformation, very few image pairs are required for training the model to achieve good results. We show this in Figure 4 where we can clearly see the performance of DDPM-based models with just a few training image pairs.

One disadvantage of DDPMs over conventional CNNs is that the inference process is quite slow because of the underlying Markovian chain, which requires multiple forward passes through a single neural network. But through

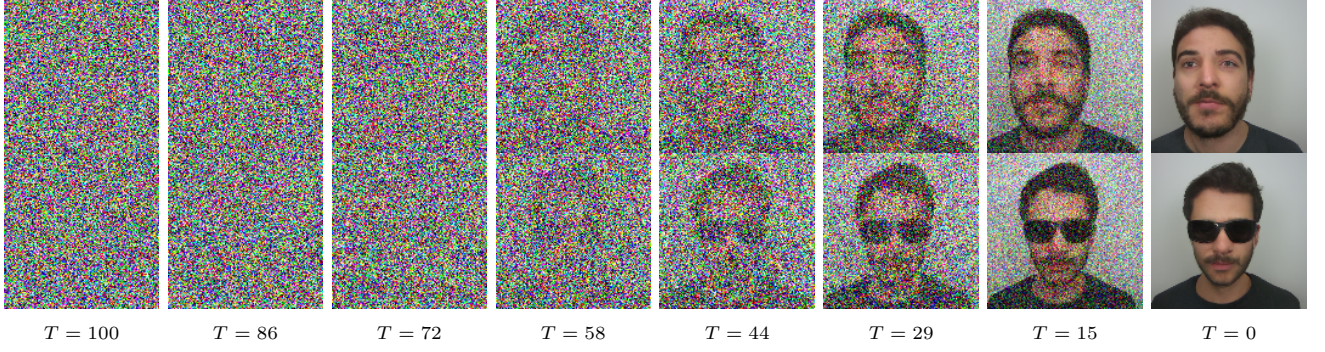


Fig. 1: Visualization of time re-scaled diffusion process during inference.

our experiments we have observed that the initial steps of the DDPM inference process aim to recover the high-level image details, like structure and color. This information can be easily captured from the corresponding thermal image; hence, rather than starting from isotropic Gaussian noise, we start with the noised thermal image. Thereby skipping some steps and speeding up the inference process.

Infrared cameras capture images in different wavelengths within the infrared spectrum based on the application. Two common modalities are the Near Infrared (NIR) and Long wave Infrared (LWIR). The choice between NIR and LWIR is based on whether it is a short-range or long-range surveillance application. NIR images consist of images close to the visible spectrum; hence the resolution of these images is naturally high and can capture finer details in the facial image captured. LWIR images are often used for long-range surveillance applications, but these images are often shot in low-resolution and fail to capture essential facial details. This could be clearly seen when we compare the facial recognition accuracies when HFR algorithms are applied to these images. NIR face datasets [3], [4] produce accuracies close to 99% but those of LWIR datasets [8], [19] are significantly lower. In this paper, we mainly focus on HFR for LWIR datasets.

In short, this paper makes the following contributions

- We propose a DDPM-based solution for the HFR problem. We focus on the image translation problem from LWIR thermal images to their corresponding visible images.
- We introduce a novel inference strategy for DDPMs mainly applicable to the T2V face translation problem. Through this we achieve upto $(2\times)$ speed up on the inference times without any drop in performance.
- We evaluate our method qualitatively and quantitatively on real-world datasets and show that it performs better than the existing methods on T2V face translation problem.

II. PROPOSED METHOD

A. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic models [23], [7] belong to a class of generative models where the models learn the distribution of data through a Markovian process. DDPMs

consist of a forward process and a reverse process. The joint distribution of the data is represented by the reverse process. The forward process is a Markovian process where the next state is obtained by sampling from a Gaussian distribution whose mean depends on the current state of the system, a predefined variance schedule as well as the time t . The sampling operation for the forward step is given by

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where $\{\beta_t\}$ is a predefined variance schedule. In practical scenarios, t usually takes values between 10^{-4} to 10^{-2} . This could also be considered as a noising operation, where the next state is obtained from the current state by adding a small Gaussian noise with variance schedule $\{\beta_t\}$. The state at timestep t can also be computed from the initial state x_0 . the distribution of the state at a particular timestep t given x_0 is,

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (2)$$

Equivalently,

$$x_t = \bar{\alpha}_t x_0 + (1 - \bar{\alpha}_t)\epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (3)$$

where, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ and $\alpha_t = 1 - \beta_t$. As can be seen from Eq. 2, for large t , $\bar{\alpha}_t$ becomes zero. Hence it results in a standard isotropic Gaussian. The reverse process denotes a generation step where we start from a standard Gaussian and iteratively perform denoising through t timesteps to generate an image corresponding to the training distribution. [23] has shown that when the number of timesteps is large and the increment in $\{\beta_i\}$ is small, then the reverse distribution could be also approximated by a Gaussian. Each step in the reverse process is performed by sampling from a distribution whose parameters are modelled using a neural network with parameters θ and each reverse step is defined by,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (4)$$

The parameters θ are obtained by minimizing the variational lower bound of the negative log-likelihood of the data distribution defined by

$$E[-\log(p(x_0))] \leq E_q\left[\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] = L. \quad (5)$$

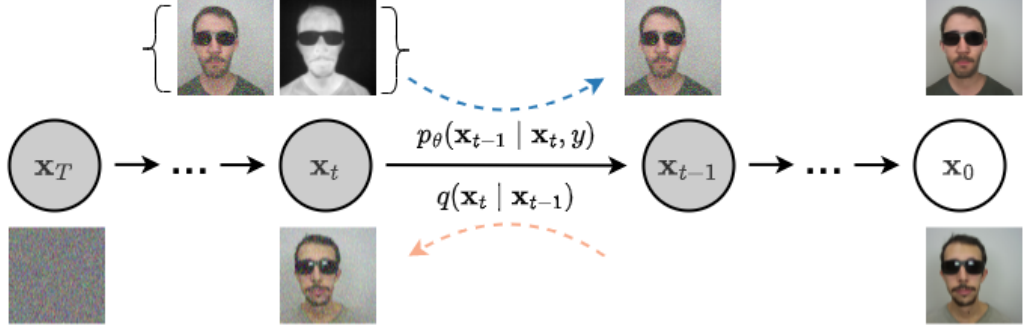


Fig. 2: An overview of the proposed method. During training the thermal image is conditioned along with the noisy visible image. During inference, we skip some timesteps by starting with a coarse image consisting of the background colour and the coarse features.

Further simplifying,

$$L = \mathbb{E}_q \left[\underbrace{D_{KL}(q(\mathbf{x}_T | \mathbf{x}_0) || p(\mathbf{x}_T))}_{L_T} - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1) + \sum_{t>1} \underbrace{D_{KL}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1} | \mathbf{x}_t))}_{L_{t-1}} \right]. \quad (6)$$

Recent works [7] have shown the variance schedule for the reverse step could be kept the same as the forward process and only the means need to be learned using a neural network. Furthermore using the input at the current timestep, the mean of the reverse step could be reparametrized according to

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(\mathbf{x}_t, t) \right). \quad (7)$$

Also given an x_0, x_t , the mean of the distribution $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ can be written as

$$\mu(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon(\mathbf{x}_t, t) \right). \quad (8)$$

Equating the means of the distributions, the training objective could be further[7] simplified to,

$$L_{simple} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2]. \quad (9)$$

Once the data distribution is learned, new samples can be derived from the data distribution, starting from a Gaussian random sample and following a Markovian process.

B. Conditional Diffusion Models for T2V Generation

Saharia et al. [22] proposed a method for conditional generation using DDPM where the images are generated based on a constraint, and the model learns the conditional distribution given a condition y . Learning the conditional distribution rather than the unconditional distribution enables the use of DDPMs for low-level vision problems like image restoration. To make the model learn the conditional distribution, we condition the neural network with the thermal image y at all timesteps t . Here the effective training objective is defined by,

$$L_{simple} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, y, t)\|^2]. \quad (10)$$

Once the conditional distribution is learnt, the model can be used for inference, i.e., given a thermal image, we can sample the corresponding visible image by starting with pure Gaussian noise and sampling from a Markov chain of T steps. The corresponding mean of the conditional distribution for the transition $p_\theta(x_{t-1} | x_t, y)$ is

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(\mathbf{x}_t, y, t) \right). \quad (11)$$

and the variance schedule remains the same as that during training.

C. Speeding up conditional diffusion models

From our experiments we made a key observation that forms the basis for this section. As can be seen from Figure 3, which shows how the generated image varies with time, the low-level semantics of the image is learnt much further into the diffusion process. The initial steps are aimed at learning coarse features like the shape of the underlying face and background color. This means if we have an image with coarse facial features but lacks finer details, it can possibly be utilized to speed up the reconstruction process. Motivated by this empirical observation, we generate an image with coarse features through the following process. Given the thermal image y and the visible image, x normalized in the range $[0, 1]$, we binarize the thermal image to generate a mask m through the following operation

$$m = \begin{cases} 1 & \text{if } x > \epsilon \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

Since the background colour is constant for almost all images, we sample the background colour from a random image in the training dataset and form a grid c of the size of the image. The coarse image y^c is created according to

$$y^c = m \cdot y + (1 - m) \cdot c. \quad (13)$$

Using y^c , the noisy version of y^c denoted by $y_{T_r}^c$ corresponding to time T_r is created. Figure 3 shows a visualization of y^c , $y_{T_r}^c$ as well as x_{T_r} starting from an isotropic Gaussian noise. From the visualization, one can note that x_{T_r} and $y_{T_r}^c$

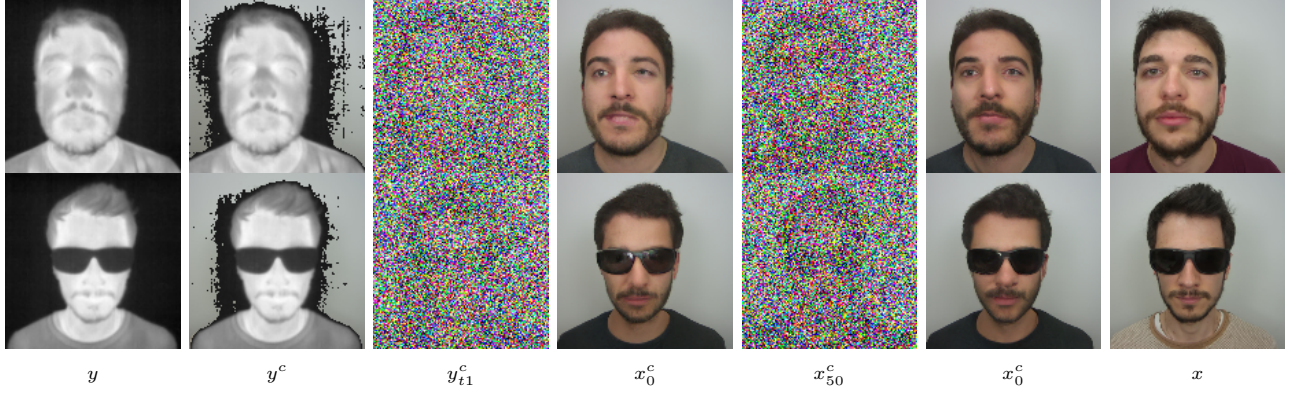


Fig. 3: (a) Thermal Image, (b) Masked thermal image (c) Noise masked thermal image($T=50$) (d) Reconstructed masked thermal image($T=100$) (e) Intermediate diffusion process result image starting from Gaussian noise visualized at ($T=50$) (f) Reconstructed image starting from ($T=0$). (g) Visible Image

for $T_r = 50$ does not differ much semantically. Moreover the SNR values of both the images are almost the same.

Algorithm 1 Training

Input: Thermal image and visible image pairs $P = \{(\mathbf{y}^k, \mathbf{x}^k)\}_{k=1}^K$

- 1: **repeat**
- 2: $(\mathbf{y}, \mathbf{x}) \sim P$
- 3: $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 5: $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_c + \sqrt{1 - \alpha_t} \epsilon$
- 6: Gradient descent step on $\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, \mathbf{y}, t)\|^2$
- 7: **until** converged

Algorithm 2 Inference

Input: Thermal image \mathbf{x} , background colour c

- 1: Create coarse image x^c according to Equation 13
- 2: Sample $x_{T_r}^c \sim q(x_{T_r}^c | x^c)$
- 3: **for** $t = T_r, \dots, 1$ **do**
- 4: sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$ else $\mathbf{z} = 0$
- 5: compute $\mathbf{x}_{t-1}^c = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t^c - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_{\theta}(\mathbf{x}_t^c, \mathbf{x}, t) \right) + \sigma_t \mathbf{z}$
- 6: **end for**
- 7: **return** x_0^c

Algorithm 1 summarises the training procedure of the proposed method. The thermal image and their corresponding noisy visible images are used for training a DDPM model for T timesteps. During inference, we reduce the inference time of the diffusion process by starting from a noisy coarse image than isotropic Gaussian noise. The inference procedure is given in Algorithm 2.

III. EXPERIMENTS

In our experiments, we focus on the translation problem from thermal images of dimension 128×128 to visible images of size 128×128 . The main emphasis of our experiments is on improving the facial recognition accuracy of

the reconstructed images, but we also perform comparisons with different metrics. Until now, there are no standardized baselines for T2V face translation. Hence we make use of the datasets used in the work [16] for our experiments. We perform experiments on three different datasets for T2V face translation. More details about the individual datasets and their evaluation criteria are presented in this section.

VIS-TH dataset [16]: The VIS-TH dataset consists of facial images corresponding to 50 different identities. Images in the VIS-TH dataset are captured through a dual-sensor camera in Long Wave Infrared (LWIR) modality and are aligned. The facial images in the dataset consist of various poses. We create the train-test split for the dataset in the same fashion as followed by [16], [8]. We use all the images corresponding to 40 randomly selected identities as the training set and the remaining images as the test set. VIS-TH is quite a challenging dataset owing to the low number of identities and the diversity of the dataset.

ARL-VTF dataset[20]: Like the VIS-TH dataset, the RL-VTF dataset also consists of facial images captured in the LWIR modality. The dataset also provides the image capture settings for aligning the faces. Visible images in the ARL-VTF dataset are severely overexposed. Hence we correct this overexposure through exposure matching with the VIS-TH dataset. We create a subset of the original ARL-VTF dataset for all our experiments and choose 100 identities with different expressions as the training dataset, and data corresponding to 40 identities as the testing set. In total, there are 3,200 training pairs and 985 testing pairs.

Evaluation metrics: For evaluating the effectiveness of our method, we utilize two different schemes like [3], [16]. We evaluate the facial verification performance of the reconstructed images by using and comparing our method with the existing methods in terms of the Rank-1 accuracy, Verification Rate (VR) @ False Accept Rate (FAR)=1% and VR@FAR=0.1%. All Facial verification experiments are conducted using pretrained ArcFace facial recognition system[1]. For evaluating the quality of the reconstructed outputs, we use the following metrics: Learned Perceptual



Fig. 4: Qualitative results for the thermal to visible translation tasks for facial images on a challenging dataset (VIS-TH dataset). Due to the low number of training image pairs, existing methods fail to reconstruct facial images accurately, see columns(1,2,3,4) in the results and doesn’t generalize well over unseen poses. On contrast Our method is able to generalize well given a low number of training image pairs.

TABLE I: Verification results on VIS-TH dataset

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
Pix2Pix[9]	20.88	1.73	0.0
SAGAN[2]	13.04	0.0	0.0
GANVFS[26]	30.98	3.43	0.0
HiFaceGAN[25]	68.48	37.14	17.14
AxialGAN[8]	58.15	23.43	6.86
T2V-DDPM	70.5	45.5	23.5

Image Patch Similarity (LPIPS) [27], Deg (cosine distance between LightCNN features), Peak Signal to Noise ratio (PSNR) of the underlying grayscale image and Structural Similarity Index (SSIM).

Training settings: For the diffusion model, we use the same parameters as for the iamgeNet superresolution model as used in improved diffusion [17]. We initialize the model with the imageNet pretrained weights. The model is trained for $T = 1000$ timesteps. During inference, we use timestep rescaling as in [17] and reduce the number of inference steps to 100. $T_r = 60$, ϵ is set equal to 0.1 for all experiments

Comparison methods: We evaluate our method by comparing with different generative model based approaches for image-to-image translation. The following methods are used for comparison: Pixel2Pixel[9], Self-Attention GAN (SAGAN)[2], GANVFS[26], HiFaceGAN[25] and AxialGAN[8].

A. Results on the VIS-TH dataset

Figure 4 shows the qualitative results on the VIS-TH dataset for four different poses. As we can see, all the

TABLE II: Image Quality metrics results on VIS-TH dataset

Method	LPIPS(↓)	Deg.(↑)	PSNR(↑)	SSIM(↑)
TH	0.5780	0.3635	6.3516	0.1841
Pix2Pix[9]	0.4847	0.2794	13.45	0.353
SAGAN[2]	0.4678	0.3367	14.14	0.4183
GANVFS[26]	0.4116	0.3292	13.83	0.4186
HiFaceGAN[25]	0.1950	0.6268	19.16	0.6963
AxialGAN[8]	0.1756	0.6157	22.80	0.7458
T2V-DDPM	0.1718	0.6441	19.419	0.7178

existing methods except HiFaceGAN fail to reconstruct attributes like hats and sunglasses. Whereas our method, as well as HiFaceGAN work well even with these occlusions. If we have a close look at the last row of Figure 4 we can see that when the frontal pose changes, HiFaceGAN fails to reconstruct the face properly, but our method works well even with different attributes as well as poses. For evaluating the methods quantitatively, we use two criteria as mentioned earlier. The quantitative results can be found in Table II and I. For the rank-1 accuracy, we can see that we gain an improvement by 2%. For VR@FAR=0.1% and VR@FAR=1%, we gain performance improvements of 8.3% and 6.3% respectively. For the image quality metrics, in terms of the perceptual similarity metrics (LPIPS and Deg), we gain an improvement of 0.0038 and 0.03, respectively. Our values for PSNR and SSIM are a bit lower than AxialGAN. However, note that PSNR often does not often represent the amount of facial details present in the reconstructed images since a blurry image could have a higher PSNR compared to a much sharper image.

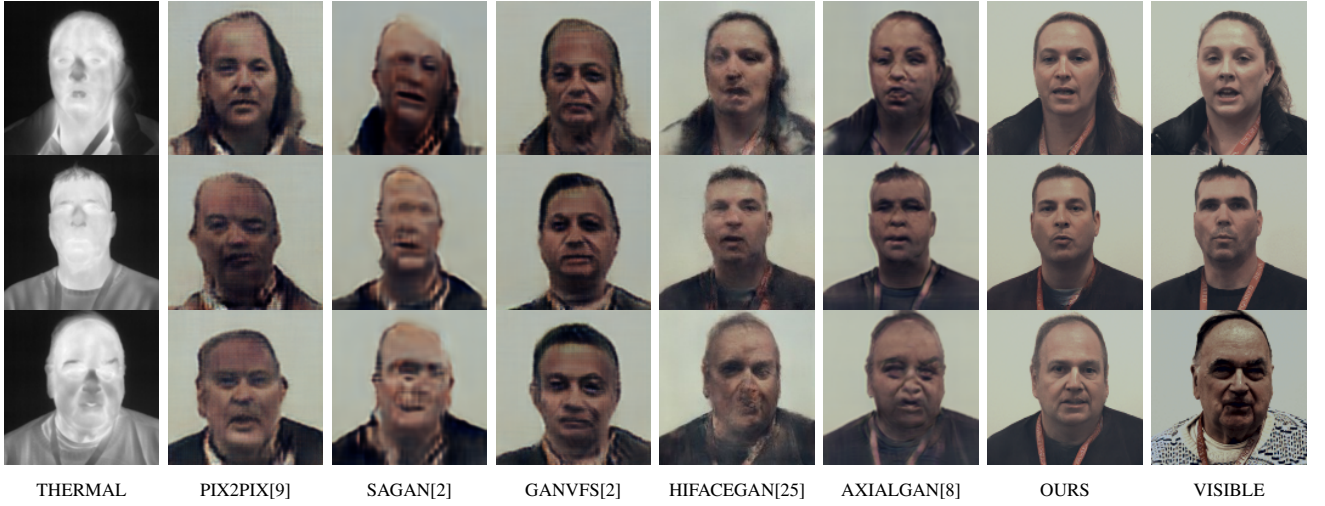


Fig. 5: Qualitative results on the ARL-VTF-dataset.

TABLE III: Verification results on the ARL-VTF dataset

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
Pix2Pix[9]	18.88	5.09	0.33
SAGAN[2]	13.46	5.42	0.33
GANVFS[26]	21.84	12.97	2.63
HiFaceGAN[25]	65.35	41.33	20.89
AxialGAN[8]	66.67	42.86	18.62
T2V-DDPM	75.37	43.51	19.87

TABLE IV: Image quality metrics results on the ARL-VTF dataset

Method	LPIPS(↓)	Deg.(↑)	PSNR(↑)	SSIM(↑)
TH	0.5551	0.4201	5.674	0.1095
Pix2Pix[9]	0.4467	0.3489	13.12	0.3804
SAGAN[2]	0.4044	0.4244	14.25	0.4490
GANVFS[26]	0.3924	0.3935	13.68	0.4313
HiFaceGAN[25]	0.1937	0.6563	19.62	0.6937
AxialGAN[8]	0.2123	0.6699	20.04	0.7179
T2V-DDPM	0.2010	0.6606	19.70	0.6775

B. Results on the ARL-VTF dataset

Figure 5 shows the qualitative results on the ARL-VTF dataset. This dataset consists of more identities and much more training and testing images. As we can see in Fig. 5, only our method can properly reconstruct the salient underlying facial features, and it can create realistic features that have coarse features corresponding to the thermal image. Further quantitative evaluations can be found in Table IV and III.

C. Ablation study: Performance variation on changing T_r :

To note the performance variation on changing T_r , we vary T_r in steps of 10 from $T_r = 40, \dots, 100$ on the VIS-TH dataset, where $T_r = 100$ denotes starting from pure Gaussian noise. The corresponding results are shown in Table V. As can be seen, the facial verification performance doesn't vary much even for $T_r = 40$. This clearly validates our claim and observation that in the initial steps of the diffusion process, only the coarse features are learnt and if this could

TABLE V: Ablation study on the TH-VIS dataset. The number of inference timesteps are varied here and the performance is noted.

T_r	Rank-1	VR@FAR=1%	VR@FAR=0.1%
40	66.5	34.0	18.0
50	70.5	45.5	23.5
60	71	42.5	31.0
70	70.5	40.0	21.0
80	68.5	39.0	23.5
90	72.0	37.0	22.0
100	70.0	40.5	18.5

be inputted properly to the diffusion model, we can achieve major boost for the inference time.

IV. CONCLUSION

We presented a solution for V2T face translation using DDPMs by treating it as a conditional image generation problem. We find that with the immense power in modeling probabilistic distributions, DDPMs prove as an ideal solution for generating samples from the conditional distribution of visible images given thermal images. We also introduce a novel sampling strategy to reduce the inference time for DDPMs. Our experiments on multiple datasets show that DDPMs perform better than GANs for the T2V face translation problem. To the best of our knowledge, this is the first work utilizing DDPMs to reconstruct visible facial images from thermal images.

V. ACKNOWLEDGMENTS

This research was supported by NSF CAREER award 2045489.

REFERENCES

- [1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

- [2] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Polarimetric thermal to visible face verification via self-attention guided synthesis. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.
- [3] B. Duan, C. Fu, Y. Li, X. Song, and R. He. Cross-spectral face hallucination via disentangling independent factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7930–7938, 2020.
- [4] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He. Dual variational generation for low shot heterogeneous face recognition. *Advances in Neural Information Processing Systems*, 32, 2019.
- [5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [7] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [8] R. Immidisetti, S. Hu, and V. M. Patel. Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.
- [9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [10] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.
- [11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [12] S. Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on pattern analysis and machine intelligence*, 29(4):627–639, 2007.
- [13] S. Z. Li, Z. Lei, and M. Ao. The hfb face database for heterogeneous face biometrics research. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2009.
- [14] S. Liu, D. Yi, Z. Lei, and S. Z. Li. Heterogeneous face image matching using multi-scale features. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 79–84. IEEE, 2012.
- [15] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.
- [16] Y. Mei, P. Guo, and V. M. Patel. Escaping data scarcity for high-resolution heterogeneous face hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18676–18686, 2022.
- [17] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.
- [18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.
- [19] N. Peri, J. Gleason, C. D. Castillo, T. Bourlai, V. M. Patel, and R. Chellappa. A synthesis-based approach for thermal-to-visible face verification. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.
- [20] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, et al. A large-scale, time-synchronized visible and thermal face dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1559–1568, 2021.
- [21] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.
- [22] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021.
- [23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [24] X. Su, J. Song, C. Meng, and S. Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.
- [25] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, and P. Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1551–1560, 2020.
- [26] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, 127(6):845–862, 2019.
- [27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.