An Approach to Synthesize Thermal Infrared Ship Images

Doan Thinh Vo Duc Phan Anh Thao Nguyen Nhu Huy Nguyen Quoc
Luu Nguyen Phong Huong Ninh
Hai Tran
Optoelectronics Center, Viettel Aerospace Institute, Viettel Group

Abstract

In this paper, we propose a method to produce synthetic thermal infrared (TIR) images using a diffusion-based image-to-image translation model. The model translates the abundantly available RGB images into synthetic TIR data closer to the domain of authentic TIR images. For this purpose, we explore the usage of an unpaired image translation neural model based on Schrödinger bridge algorithms. In addition, the visual characteristic of the object in the image is an important consideration when generating the results. Thus, we take advantage of a segmentation module before the image-to-image translation model to discriminate the background and object regions. We practice the model's performance with a self-proposed dataset comprising unpaired realistic RGB-TIR images. When incorporated into the training set, the synthesized images of our model significantly increase the classification accuracy by 15% and F1-score by 18% when only using realistic TIR images.

1. Introduction

The thermal image is crucial for various applications, such as car and pedestrian surveillance systems and defense mechanisms, due to its independence from illumination fluctuations and its effectiveness in differentiating objects from backgrounds, especially in total darkness [20, 25, 29]. However, a significant challenge persists: the scarcity of publicly available thermal image datasets. This shortage hinders the advancement of system performance in diverse scenarios. Additionally, the limited availability of annotated thermal datasets exacerbates the issue, impeding the development and validation of robust machine-learning models for real-world environments. Addressing these data limitations is critical to enhancing the efficacy and applicability of thermal infrared base tasks.

To address the limitations caused by a lack of data, scientists have looked into some approaches, such as generation and simulation [32] techniques, for producing thermal images. Although simulation provides a regulated setting for producing synthetic temperature data, it frequently has built-in drawbacks. There may be differences between synthetic and real thermal images as a result of the simulation's inability to fully represent the subtleties and complexity of real-world thermal dynamics. Furthermore, there are major computing problems in simulating a variety of environmental variables and object interactions, which limits the scalability and usefulness of simulated datasets for training and testing.

On the other hand, generative techniques such as Generative Adversarial Networks (GANs) show promise in synthesizing realistic thermal images [14]. However, they face challenges such as mode collapse and training instability. Additionally, there is a risk of generating artifacts that deviate from physical realism. Diffusion-based methods offer a promising alternative, simulating thermal diffusion within materials to produce stable and reliable thermal images resembling real-world scenarios. Unlike GANs, they mitigate mode collapse and artifacts, ensuring higher fidelity and consistency. They also provide scalability and flexibility in simulating diverse environmental conditions, enhancing their applicability. As such, diffusion-based techniques represent a significant advancement in thermal image generation, offering solutions to challenges associated with traditional generative approaches.

In sum, our paper introduces two following contributions that aim to enhance the landscape of thermal imaging and maritime surveillance that aim to enhance the landscape of thermal imaging and maritime surveillance:

- We present a novel approach for generating thermal object images based on diffusion-based Image-to-Image (I2I) techniques. Leveraging the principles of thermal diffusion within materials, our method offers a systematic and principled framework for synthesizing synthetic thermal objects with remarkable realism and accuracy.
- We describe a comprehensive RGB dataset comprising various types of vessels encountered in maritime environments. Through rigorous curation and annotation, our dataset encompasses a diverse range of vessel classes with full class coverage, ensuring completeness and accuracy

in vessel detection and classification algorithms

2. Related Works

Thermal image synthesis In recent times, deep neural networks have been employed for image-to-image translation across different spectral ranges. In this paper, we present a solution for converting color images into thermal images.

Based on the modification of Residual SqueezeNet, ThermalNet in [12] incorporates fourteen layers, comprising convolution, deconvolution, and fire modules. Subsequent to thermal image generation, postprocessing is performed using a trained VGG-16 network, acting as a reference for similarity during the synthesis of the output thermal image to match a predetermined ground truth image.

With the progression of Generative Adversarial Networks (GANs), several investigations leverage this advancement for thermal image generation. [13] introduced an innovative training methodology, expanding the conventional GAN training framework from a two-player adversarial game to a three-player setup, where the third player supplies true negative samples to the discriminator network. [18] put forth an approach employing GANs, semantic segmentation, and 3D modeling to synthetically generate thermal images. Building upon the work of [8], [19] employs GANs to synthesize thermal imaging images, incorporating modifications to the algorithm, a novel neural network architecture and an innovative training approach. The study demonstrates the efficacy of the proposed method in producing infrared images closely aligned with the ground truth model in terms of both thermal emissivity and geometric shape. The synthesis of visible to thermal images is also achieved through the application of generative adversarial networks (GAN) using Style-GAN2 in [4], incorporating different variants of StyleGAN2 along with the updated StyleGAN that features adaptive discriminator augmentation (ADA).

I2I translation algorithm Image-to-image translation (I2I) aims to learn a mapping between two distinct image domains. This field has gained attention due to its applications, with Generative Adversarial Network (GAN)-based methods proposed for tasks like unpaired translation [33], unsupervised cross-domain generation [28], multi-domain translation [1], and few-shot translation [15]. However, prevailing GAN models may face limitations in achieving consistent structural and textural regularity. Diffusion models [27] have shown notable achievements in image generation [2, 5], super-resolution [7, 21], unpaired I2I translation [23], and image editing [17, 26]. Palette [22] surpassed strong GAN and regression benchmarks across colorization, inpainting, uncropping, and JPEG restoration without task-specific adjustments. DiffI2I [31] is an effective diffusion model framework for I2I tasks, comprising a compact prior extraction network, dynamic transformer, and denoising network. It achieves SOTA performance while reducing computational burdens. The Unpaired Neural Schrödinger Bridge (UNSB) [9] addresses the Gaussian prior assumption constraints in unpaired I2I translation by framing the Schrödinger Bridge problem as adversarial learning, facilitating scalability across diverse tasks.

3. Method

3.1. Dataset

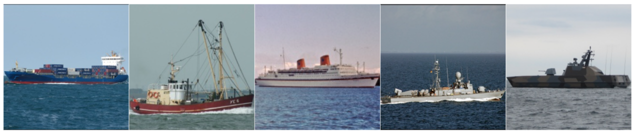

Figure 1. Example of RGB ship dataset

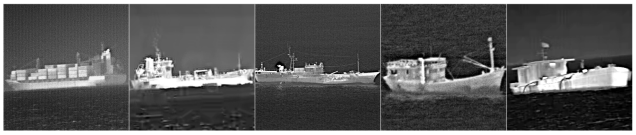

Figure 2. Example of TIR ship dataset

Our goal is to obtain two large-scale and high-diversity ship datasets which are used for training the I2I model. The first dataset contains RGB images and the second includes TIR images. For the RGB dataset, we choose base images on the Spotship website. This dataset contains various types of ships from civilian ships such as fishing vessel, cargo liner, bulk carrier to military ships, i.e. frigate, destroyer, battleship, etc. To eliminate the impact of the image background and converging solely on the object features, images without complex backgrounds are chosen. In view of the fact that public TIR ship datasets are limited, we utilize our LWIR camera to collect ship images in 3 types of scenarios(i.e., video surveillance, ship-mounted, hand-held). This dataset includes a variety of classes (i.e. fishing boat, container ship, cruise, ferry, etc.) in different weather conditions (i.e. day, night, mist, sunny, rainy, etc.). Finally, the base dataset contains 25000 images for the RGB ship and 2100 images for TIR ship.

Both datasets require all images to be in square size and focus on the object. We set up a two steps processing to crop the ship and remove unnecessary areas:

Initially, YoloV5 [6] model pretrained on VOC [3] dataset will be used to detect and extract the bounding box of the ship. For some images that the model fails to detect, we manually create a bounding box for the ship. After that, if some bounding boxes created automatically from YoloV5 model are not fit with the ship, especially with the TIR dataset. In this case, we manually fix the bounding box of the target with the support of CVAT [24] annotation tools. Furthermore, our method will crop the ship base on the bounding box and then add zero padding to make a square-size image.

After data processing phase, we obtain 31900 images for RGB dataset and 2100 images for TIR dataset. Since there is more than one ship in some images, the number of images in the final datasets is greater than the number of images in the base datasets. Some images in both datasets are shown in Figs. 1 and 2.

3.2. Synthetic TIR Generative Framework

We adopt powerful **U**npaired image-to-image translation via **N**eural **S**chrodinger **B**ridge, denoted as **UNSB** [10] for synthetic TIR Generative framework. Our goal is to learn the unpaired translation between the RGB and thermal image domain.

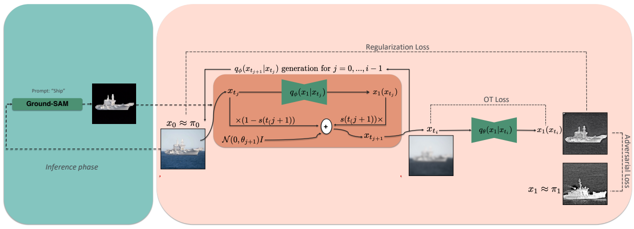

Figure 3. Overview of the proposed framework. **Left**: Grounded-SAM segmentation module. **Right**: UNSB image-to-image translation module

Let π_0 and π_1 denote the probability distributions of RGB and thermal image domains, respectively. The goal is to learn to translate samples from one domain to another domain. Due to the source distribution is already known as π_0, the main interest lies in revealing the transition probabilities $p(x_{t+1} \mid x_t)$ for $i = 0, .., N-1$. Fig. 3 illustrates the overall architecture.

We first sample $x_0 \sim \pi_0$ as x_{t_j}. To obtain the intermediate sample x_{t_i}, it will pass the to a generative network, G, which representing the transition mapping $p(x_{t+1} \mid x_t)$. This generates $x_1(x_{t_j})$. After that, x_0 and $x_1(x_{t_j})$ are interpolated by adding Gaussian noise to yields $x_{t_{j+1}}$.

We repeat this process by a NFE (Number of Function Evaluation) N = 5 times to obtain the intermediate sample x_{t_i}. Finally, we pass the intermediate x_{t_i} to the generative network G, representing the translation $p(x_1 \mid x_{t_i})$, to obtain our final prediction in the thermal domain π_1, denoted as $x_1(x_{t_i})$.

The image-to-image (I2I) translation framework encounters a significant limitation: it has minimal awareness of individual instances due to the absence of an attention mechanism that correlates images with their semantic context at the instance level. This shortfall is evident in the first two columns of Fig. 4, where the I2I framework's tendency to focus on areas of high intensity results in suboptimal object retention. To tackle this problem, we employ a simple yet effective approach that utilize a segmentation module to identify the instance region and black-out other region. We utilize the pre-trained Grounded-Segment Anything (Grounded-SAM) model, which facilitates open-domain segmentation via textual prompts. Grounded-SAM synergizes the GroundingDINO [16] and SAM [11] models, eliminating the need for manual bounding box annotations. When provided with an RGB image x_0 and a corresponding text prompt such as "Ship," Grounded-SAM precisely isolates the ship's region, represented as $S(x_0)$. This isolated segment, $S(x_0)$, is subsequently input into the UNSB I2I translation framework to produce the refined output image. The enhancements in object retention facilitated by the segmentation module are showcased in Fig. 4.

The overall loss is defined as the weighted summation of the Adversarial Loss, Schrödinger Bridge Loss, Regularization Loss. The Adversarial Loss ensures that the generated thermal image $x_1(x_{t_i})$ conforms to the target thermal image distribution. This loss computes the Kullback-Leibler Divergence between the predicted thermal distribu-

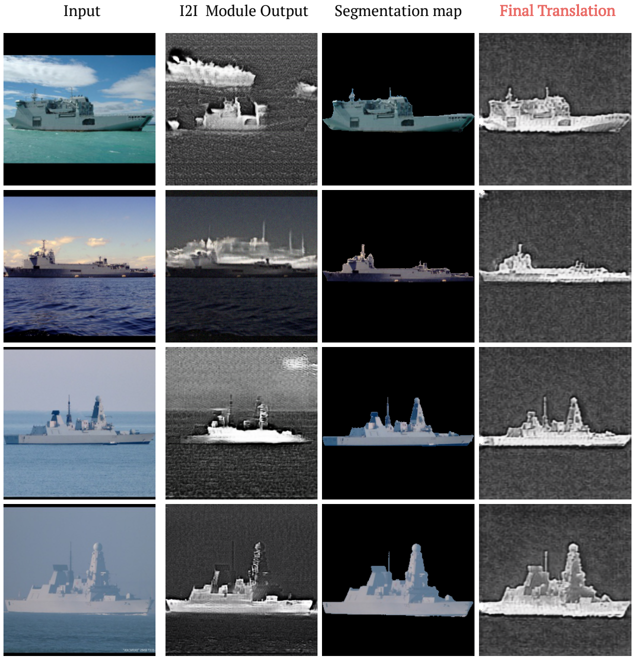

Figure 4. The I2I module, when used independently, frequently misinterprets image content. This can manifest as incorrectly identifying non-object regions (e.g., clouds) as targets for translation or partially converting objects to the thermal domain, with darker regions overlooked. By incorporating a robust segmentation module, the model can accurately delineate objects within the image, enabling robust translation

tion $q_{\phi_i}(\boldsymbol{x}_1)$ and the real thermal distribution $p(\boldsymbol{x}_1)$. The Schrödinger Bridge Loss measures the energy needed to transport the image domain through the transition $p(x_{t+1} \mid x_t)$. This loss function aims to find an optimal transformation map between two domains. It is defined as the expected squared distance between $\boldsymbol{x}_{t_i}$ and $\boldsymbol{x}_1$, minus a term involving the entropy of the joint distribution between xt_i and x_1. The Regularization Loss is augmented to the framework as regularization, which enforces the generator network G to satisfy consistency between the predicted x_1 and x_0. It is defined as the expected value of a scalar-valued function $\mathcal{R}$ that measures the similarity between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$.

4. Experiments

Training We first summarize the details of the implementation of the I2I UNSB module. We conduct an experiment using a single 24GB A5000 GPU. We employ the dataset previously outlined in the methodology section. we define the real RGB images to be the source domain π_0 and the thermal TIR images as the target domain π_1. For training, we randomly choose 10000 RGB images and all 2100 thermal images, which are not paired together. We train this module using images of three sizes: 128x128x3, 256x256x3, and 512x512x3. We employ an Adam optimizer for 20 epochs of batch size 1 using a step decay learning rate scheduler. The initial learning rate of 2e-4. For loss weights, we set $\lambda_{SB} = 1$, $\lambda_{Reg} = 1.5$, and $\lambda_{Adv} = 1$.

Metrics We assess the quality of the translated image considering its faithfulness. We compute the standard image distance metric the Structural Similarity Index Metric SSIM [30]. These are calculated for each generated image and its corresponding real image. Rather than comparing the entire images, our analysis specifically targets the preservation of the ship object. To achieve this, we employ the Grounded-SAM model with the prompt **”Ship”** to segment the ship object. Subsequently, we modify the segmented object’s color to white and the background to black. This approach effectively eliminates unnecessary details, allowing us to solely concentrate on the preservation of the ship object. Higher values of SSIM indicate greater structural similarity between the translated and real images. Fig. 5 illustrates an example of SSIM between the original images and the translated thermal images.

Synthetic thermal set on classification To assess the efficacy of synthetic datasets, we explore a classification task involving five types of ships: Cargo, Cruise, Fishing, Tanker, and Destroyer. We train models on three image resolutions: 128x128x3, 256x256x3, and 512x512x3 using EfficientNet_B3, MixNet_XL, and ViT. We train 1000 epochs for each scenario and choose the best checkpoint based on accuracy. We divide our 2100 real thermal images into 1470 train and 630 test images. We fix test set and consider three training set scenarios: (1) using only the 1470 real images, (2) using only 10000 synthetic images, and (3) combining both real and synthetic images for a total of 11470 images. The results are presented in Table 1.

Table 1. Comparison of classification performance across image sizes and models.

	Metrics	**Accuracy**			**F1-Score**		
	Image Size	128	256	512	128	256	512
w/o synthetic	EfficientNet	0.84	0.81	0.82	0.84	0.80	0.84
	MixNet	0.85	**0.87**	0.86	**0.89**	0.90	0.86
	ViT	0.86	0.82	0.86	0.88	0.85	0.79
synthetic only	EfficientNet	0.90	0.89	0.91	0.88	0.90	0.89
	MixNet	0.89	0.90	0.88	0.91	0.90	0.89
	ViT	0.93	0.94	**0.95**	0.92	**0.94**	**0.94**
synthetic + original	EfficientNet	0.93	0.95	0.92	0.96	**0.97**	0.91
	MixNet	0.91	0.92	0.96	0.93	0.95	0.91
	ViT	0.93	**0.98**	0.97	0.94	0.94	0.95

The inclusion of synthetic thermal images, whether exclusively or in conjunction with real images, significantly improves classification accuracy and the F1 score. ”The combination of synthetic and original images in the training set shows the best results, with accuracy improvements ranging from 10% to 15% and F1 score improvements ranging from 5% to 18%. Notably, the 256x256x3 resolution models demonstrated superior improvement, with ViT achieving a 15% increase in Accuracy and EfficientNet_B3 an 18% improvement in F1 score.

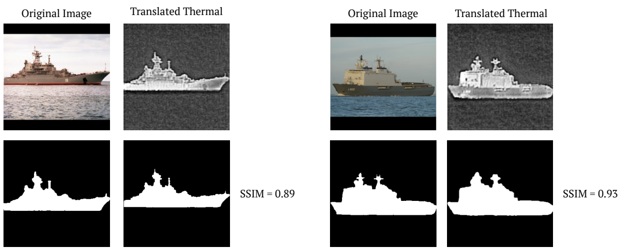

Figure 5. Example of SSIM between the original and translated thermal images.

5. Conclusion

In this paper, we have presented a new approach based on I2I translation and segmentation model, which can directly and accurately create synthetic thermal infrared images from visible images. Using a segmentation module to process the data before the I2I translation model, we aim to distinguish between the background and object areas. For practical purpose, we create a dataset comprising unpaired realistic RGB-TIR images. Our method demonstrates good performance in generating synthetic TIR images, which subsequently improves classification performance over the use of solely real TIR images. In the future, we intend to refine our method’s capability to produce thermal images with detailed background patterns.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2018. 2

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 2

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2

[4] Gabriel Hermosilla, Diego-Ignacio Henríquez Tapia, Héctor Allende-Cid, Gonzalo Farías Castro, and Esteban Vera. Thermal face generation using stylegan. *IEEE Access*, 9: 80511–80523, 2021. 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[6] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu , changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, 2020. 2

[7] Zahra Kadkhodaie and Eero P. Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser, 2021. 2

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 2

[9] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge, 2023. 2

[10] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *ICLR*, 2024. 3

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[12] Vladimir Kniaz, V. Gorbatsevich, and Vladimir Mizginov. Thermalnet: A deep convolutional network for synthetic thermal image generation. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W4:41–45, 2017. 2

[13] V. V. Kniaz and V. A. Mizginov. Thermal texture generation and 3d model reconstruction using sfm and gan. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2:519–524, 2018. 2

[14] Vladimir V. Kniaz, Vladimir A. Knyaz, Jiří Hladůvka, Walter G. Kropatsch, and Vladimir A. Mizginov. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In *Computer Vision – ECCV 2018 Workshops*. Springer International Publishing, 2018. 1

[15] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation, 2019. 2

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. 2

[18] V. A. Mizginov and S. Y. Danilov. Synthetic thermal background and object texture generation using geometric information and gan. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W12:149–154, 2019. 2

[19] V. A. Mizginov, V. V. Kniaz, and N. A. Fomin. A method for synthesizing thermal images using gan multi-layered approach. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-2/W1-2021:155–162, 2021. 2

[20] F. Piras, E. De Moura Presa, P. Wellig, and M. Liebling. Local estimation of parametric point spread functions in thermal images via convolutional neural networks. *Target and Background Signatures VIII*, 2022. 1

[21] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 2

[22] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022. 2

[23] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models, 2021. 2

[24] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, 2020. 3

[25] Ivana Shopovska, Ljubomir Jovanov, and Wilfried Philips. Deep visible and thermal image fusion for enhanced pedestrian visibility. *Sensors*, 19(17), 2019. 1

[26] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-denoising models for few-shot conditional generation, 2021. 2

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the*

32nd International Conference on Machine Learning, pages 2256–2265, Lille, France, 2015. PMLR. 2

[28] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation, 2016. 2

[29] Emirhan Tosun, Omer Faruk Dinc, Berfin Arli, and Serhat Tozburun. A classifier for dynamic thermal imaging. In *Translational Biophotonics: Diagnostics and Therapeutics III*, page 126271H. Optica Publishing Group, 2023. 1

[30] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 4

[31] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, Radu Timotfe, and Luc Van Gool. Diffi2i: Efficient diffusion model for image-to-image translation, 2023. 2

[32] Ramin Zaeim and Meysam Taheri. Computer simulation of thermal imaging system. In *UKACC International Conference on Control 2010*, pages 1–4, 2010. 1

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 2