2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3503-1892-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/WACV57701.2024.00164

PETIT-GAN: Physically Enhanced Thermal Image-Translating Generative Adversarial Network

Omri Berman[1,2], Navot Oz[1,2], David Mendlovic[1], Nir Sochen[1], Yafit Cohen[2], Iftach Klapp[2]
{omriberman,navotoz}@mail.tau.ac.il david@eng.tau.ac.il sochen@tauex.tau.ac.il
{yafitush,iftach}@volcani.agri.gov.il
[1]Tel Aviv University, Tel Aviv, Israel
[2]Agricultural Research Organization - Volcani Institute, Rishon LeZion, Israel

Abstract

Thermal multispectral imagery is imperative for a plethora of environmental applications. Unfortunately, there are no publicly-available datasets of thermal multispectral images with a high spatial resolution that would enable the development of algorithms and systems in this field. However, image-to-image (I2I) translation could be used to artificially synthesize such data by transforming largely-available datasets of other visual modalities. In most cases, pairs of content-wise-aligned input-target images are not available, making it harder to train and converge to a satisfying solution. Nevertheless, some data domains, and particularly the thermal domain, have unique properties that tie the input to the output that could help mitigate those weaknesses. We propose PETIT-GAN, a physically enhanced thermal image-translating generative adversarial network to transform between different thermal modalities - a step toward synthesizing a complete thermal multispectral dataset. Our novel approach embeds physically modeled prior information in an UI2I translation to produce outputs with greater fidelity to the target modality. We further show that our solution outperforms the current state-of-the-art architectures at thermal UI2I translation by approximately 50% *with respect to the standard perceptual metrics, and enjoys a more robust training procedure. The code and data used for the development and analysis of our method are publicly available and can be accessed through our project's website: https://bermanz.github.io/PETIT*

1. Introduction

The emergence of nanosatellites has revolutionized the field of remote sensing, allowing to acquire ground images with higher spatial resolution and at lower costs. Many sensing systems involving nanosatellites are focused at the visible and near infrared spectrum (between 380-1400 nm) - MISC [11] (between 380-700), Charybdis [18] (between 412-870 nm) and SuperDOVE [30] (between 431-885 nm) to name a few. However, very little to no attention was given to the longwave-infrared (LWIR) spectrum (7000 - 14000 nm), a.k.a., the thermal spectrum, which plays a crucial role in various environmental aspects, such as climate and water monitoring, fires prediction, *etc*. Identifying such phenomena requires thermal multispectral imaging, *i.e.*, a collection of several image layers of the same scene where each layer is acquired at a particular wavelength band belonging to the thermal spectrum. Unfortunately, while panchromatic (wide-band) thermal images were relatively easy to obtain, no off-the-shelf high-resolution thermal multispectral images were available to develop proofs of concepts for those applications. To tackle this deficiency, we used a light-plane, a thermal camera and a set of infrared bandpass filters to collect and assemble an aerial thermal multispectral images dataset. Due to inherent setup limitations, the amount of collected images per spectral channel was limited. Moreover, images of different wavelengths were not spatially registered, which is essential for a complete multispectral dataset. To overcome both the sample size and registration issues, we developed an unpaired image to image translation algorithm to transform the relatively abundant panchromatic images into pixel-wise-aligned multispectral images.

Image to image (I2I) translation is the task of transforming the style of an image to that of a different domain while preserving its content. Many methods have been developed to tackle this task, utilizing various deep neural architectures such as auto-encoders [36], generative adversarial networks (GANs) [23, 37, 39], diffusion models [25, 26] and more. Those methods have countless applications and are being used for numerous purposes, such as synthetic dataset generation for learning tasks in fields such as autonomous cars [1, 3], medical imaging [2, 29], *etc*. In most cases, as in our problem, there are no pairs of content-equivalent images

2642-9381/24/$31.00 ©2024 IEEE
DOI 10.1109/WACV57701.2024.00164

in the input and output domains. The I2I transformation in those cases is termed *unpaired*. As the unpaired image-to-image (UI2I) translation task is unsupervised and highly ill-posed, those models are usually very hard to train.

In contrast to other visual domains, thermal imaging has unique underlying physical properties that are shared across all thermal modalities. These properties enable the design of closed-form transformations between panchromatic and multispectral images. In turn, those transformations can be embedded in deep UI2I architectures to improve their statistical performance and robustness.

To test this novel hypothesis, we first train two different state-of-the-art (SOTA) GANs to perform thermal UI2I translation to establish a baseline. We then provide the generators with an additional physical property of the acquired images, allowing the GANs generator's output to be conditioned on that property. Finally, we design a physical estimator and fuse it with the generator, resulting in our proposed method, named a physically enhanced thermal image-translating (PETIT) GAN.

Statistical analysis showed that our solution achieves an improvement of approximately 50% compared to the SOTA GANs w.r.t. the conventional evaluation metrics. Furthermore, our method exhibits a more robust training procedure, possibly indicating convergence to flatter minima. These improvements are further demonstrated qualitatively through our method's greater visual fidelity to the desired target domain.

Our paper's contribution is three-fold: (1) Application of UI2I between different thermal image modalities; (2) Development and utilization of an analytic-physical-UI2I translation model; (3) Introduction of a novel thermal aerial images dataset with unpaired images of different spectral bands.

2. Related work

GAN is a class of deep generative models, first introduced by Goodfellow *et al*. [6]. Many improvements and extensions have been made in the field of GANs in the last few years, and many of these form the underlying principles and architecture of numerous SOTA models in several generative tasks. GANs are typically made up of two components: (1) a generator, which samples random vectors from some predefined probability density function as inputs and transforms them into meaningful outputs of some target modality; (2) a discriminator, which has access to both real images from the target modality and the generator's outputs, and needs to tell them apart. The generator and discriminator are trained in an adversarial fashion where one's improvement comes at the expense of the other's. If successful, the training procedure converges when the generator and discriminator reach a Nash equilibrium [6].

Among the various tasks performed by GANs is I2I translation, where the output is conditioned on an input image. I2I translation has a plethora of applications, such as image segmentation [17, 31], pose estimation [4, 16], colorization [10, 27, 34], super resolution [33, 35] and many more. The I2I translation task can be roughly classified into supervised I2I (paired I2I), where each image in the input domain has a content-aligned equivalent in the output domain, and unsupervised I2I (UI2I), where there are no content-equivalent pairs in the input and output domains. Most practical I2I tasks are performed in an unsupervised fashion, as fully registered pairs of images in two different modalities are extremely difficult to obtain.

The great challenge in UI2I translation is that no ground truth is available as a reference for the transformed output. Thus, in contrast to paired I2I, pixel-level loss cannot be used to steer the training toward a better content-preserving solution. Therefore, content preservation of the transformation must be enforced by an alternative mechanism. The most popular strategy to ensure content preservation is to use cycle consistency [15]. This approach relies on two translators: one from domain A to domain B ($G_{A\to B}$), and one in the opposite direction ($G_{B\to A}$). In addition to the standard adversarial loss, a cycle-consistency loss is used to penalize for discrepancies between input x_A and its reconstruction by the roundtrip transformation from A to B and then back to A:

$$\mathcal{L}_{cyc} = \mathcal{L}\left(x_A, G_{B\to A}\left(G_{A\to B}(x_A)\right)\right) \tag{1}$$

CycleGAN [39], along with DiscoGAN [13] and DualGAN [32], additionally impose cycle consistency over images originating in domain B, resulting in two simultaneous cyclic losses.

While successfully eliminating the need for ground truth, cycle consistency inherently encourages the transformation to encode information about the input that serves solely for the purpose of cyclic reconstruction. This encoded information comes at the expense of fidelity to the target modality, which is clearly undesirable. In an attempt to eliminate the need for cycle consistency, several approaches have implemented a one-sided translation that manages to preserve content in a different fashion. Typically, this is done by embedding both input and target in some shared style-agnostic space. The geometric distance between the embeddings is treated as a measure of content discrepancy, and then minimized to improve content preservation. Fu *et al*. [5] encouraged preservation of the geometric relationship between an input and its geometrically transformed versions and their outputs. Both F-LSeSim [38] and contrastive unpaired translation (CUT) [23] used contrastive representation learning by maximizing the similarity between pairs of corresponding patches in the input and output, and minimizing it for non-matching patches.

3. Proposed method

Toward synthesizing a multispectral dataset, we showcase the UI2I transformation between a panchromatic modality, *i.e*., a wide-band thermal image, and a single monochromatic modality, *i.e*., a narrow-band thermal image. More concretely, we transform images with a bandwidth of $7-14\mu m$ to images with a central wavelength of $9\mu m$ and a full width at half maximum (FWHM) of $0.5\mu m$. For ease of notation, we will use the subscripts *pan* to describe panchromatic data, and *mono* for monochromatic. This concept could be easily extended to synthesize a complete multispectral dataset by applying it repeatedly for disjoint sub-bands of the thermal spectrum to full coverage.

3.1. Physical estimator

3.1.1 Background

Black-body radiation is the thermal electro-magnetic signal that is emitted by an ideal opaque object due to its temperature. Planck's law of black-body radiation states that:

$$B_\lambda(T)=\frac{2\pi hc^2}{\lambda^5}\frac{1}{e^{\frac{hc}{\lambda kT}}-1}\quad\left[Wsr^{-1}m^{-2}\mu m^{-1}\right]\tag{2}$$

where $B_\lambda(T)$ is the ideal object's spectral radiance density, h is Planck's constant, c is the speed of light in a vacuum, k is the Boltzmann constant, λ is the electromagnetic wavelength and T is the object's absolute temperature [14]. The Stefan-Boltzmann law [28] ties the power radiated from an object (which is the result of integrating $B_\lambda(T)$ over the entire spectrum of wavelengths from zero to infinity) to the object's temperature:

$$P(T)=\int_0^\infty B_\lambda(T)d\lambda=\frac{\sigma}{\pi}T^4\quad\left[Wsr^{-1}m^{-2}\right]\tag{3}$$

where P is the radiated power and σ is the Stephan-Boltzmann constant.

A real-world opaque object emits less power than an ideal black body at the same temperature. The ratio between the radiation emission of an object and that of an ideal blackbody at the same temperature is called *emissivity*. The emissivity is a function of various a-priori unpredictable characteristics of the object, such as material type, surface structure, viewing angle, *etc*. Thus, the Stephan-Boltzmann law for practical objects is given by:

$$P(T)=\frac{\sigma}{\pi}\epsilon T^4\quad\left[Wsr^{-1}m^{-2}\right]\tag{4}$$

where ϵ is the object's emissivity. In general, the emissivity is a function of the wavelength [12], but it is used here as a constant that reflects its expected value over the thermal bandwidth for simplification.

According to [14], when acquired by a thermal microbolometric camera with a finite bandwidth, the incident power on the microbolometer (the thermal camera's sensor) can be described by:

$$\phi(T)=\gamma F_{pan}\epsilon T^4\quad\left[Wsr^{-1}m^{-2}\right]\tag{5}$$

where ϕ is the incident power on the microbolometer, γ is a constant governed by the camera's geometrical properties and F_{pan} represents the fraction of power that is within the camera's bandwidth. When applying an IR bandpass filter over the camera lens, equation 5 still holds except that F_{pan} is replaced by F_{mono}, reflecting the fraction of power that is strictly within the bandpass region of the applied filter [14].

Since $\gamma, F_{pan}, \epsilon$ are all constants, they can be reduced into a single coefficient:

$$\phi(T)=aT^4\quad\left[Wsr^{-1}m^{-2}\right]\tag{6}$$

suggesting that the incident power is proportional to T^4. Consequently, a thermally stabilized camera operating in radiometric mode, *i.e*., when the image intensity levels are linear in the incident power, the intensity of a pixel is obtained by an affine transformation of T^4:

$$I(T)=b+aT^4\tag{7}$$

where I is the intensity level of the pixel and b is the digital bias level.

Based solely on equation 7, we can seemingly infer an object's temperature directly from the thermal image intensity and vice versa. However, when dealing with an uncooled (non-thermally stabilized) microbolometric camera, the coefficients in the equation are highly sensitive to the camera's intrinsic temperature. According to [21], the dependency of those coefficients on the intrinsic temperature can be faithfully approximated by a third-order polynomial, concluding that a more accurate description of a pixel's intensity level is

$$I(T_{obj},T_{int})=p_c^{(0)}(T_{int})+p_c^{(1)}(T_{int})T_{obj}^4\tag{8}$$

where T_{obj} is the object's absolute temperature, T_{int} is the camera's intrinsic temperature at the time of acquisition, and:

$$p_c^{(i)}(T_{int})=\sum_{k=0}^{3}c_{i,k}T_{int}^k\tag{9}$$

where the superscript $^{(i)}$ indicates that the two polynomials in equation 8 have different coefficients. Plugging equation 9 into 8 and simplifying all of the terms gives:

$$\begin{aligned}I(T_{obj},T_{int})=&\,c_{0,0}+c_{0,1}\cdot T_{int}+c_{0,2}\cdot T_{int}^2\\&+c_{0,3}\cdot T_{int}^3+(c_{1,0}+c_{1,1}\cdot T_{int}\\&+c_{1,2}\cdot T_{int}^2+c_{1,3}\cdot T_{int}^3)\cdot T_{obj}^4\\&=<F,C>\end{aligned}\tag{10}$$

where in the last transition, we factorize the relationship as an inner product by stacking all monomials in a single feature vector F and all coefficients in a vector C. Overall, the estimator in equation 10 is made up of eight different monomials and parametrized by eight corresponding coefficients.

3.1.2 Estimator modeling

Traditionally, the coefficients from equation 8 are calibrated to estimate an object's temperature given a measured intensity. However, we noticed that it could also be used in the opposite direction, *i.e.*, to produce a thermal image intensity given a known object temperature. Our innovation is in combining the two directions of applying equation 8 in a cascade to assemble an analytic UI2I translation model in the following way:

Given a set of calibrated panchromatic coefficients, we can estimate the object's temperature using the panchromatic intensity and the intrinsic temperature at acquisition:

$$\hat{T}_{obj} = \sqrt[4]{\frac{I_{pan} - p^{(0)}_{c_{pan}}(T_{pan})}{p^{(1)}_{c_{pan}}(T_{pan})}} \tag{11}$$

With the estimated object temperature at hand, we can invoke equation 8 once again, this time using calibrated monochromatic coefficients, to estimate the monochromatic intensity:

$$\hat{I}_{mono} = p^{(0)}_{c_{mono}}(T_{mono}) + p^{(1)}_{c_{mono}}(T_{mono})\hat{T}^4_{obj} \tag{12}$$

We treat the cascaded utilization of equations 11 and 12 as the physical estimator, and denote it by G_{phys}. Formally:

$$\hat{I}_{mono} = G_{phys}(I_{pan}, T_{pan}, T_{mono}) \tag{13}$$

3.1.3 Estimator coefficient calibration

Our in-house-designed calibration setup consists of a thermal camera, blackbody target (to control the scene's temperature) and an environmental chamber (to control the camera's ambient temperature). The setup was used to capture images of varying scenes and ambient temperatures, to cover the three-dimensional T_{int}-T_{obj}-intensity space. We then used the measurements to solve for the physical estimator's coefficients using a least-squares minimization criterion. The calibration results of an exemplary pixel can be visualized as a surface in the three-dimensional T_{int}-T_{obj}-intensity space, as shown in Figure 1. Since the physical estimator requires both panchromatic and monochromatic calibrated coefficients, the calibration process was conducted twice, with and without applying an IR bandpass filter over the camera lens. For a more elaborate description of the calibration process, please refer to section 6 in the supplementary material.

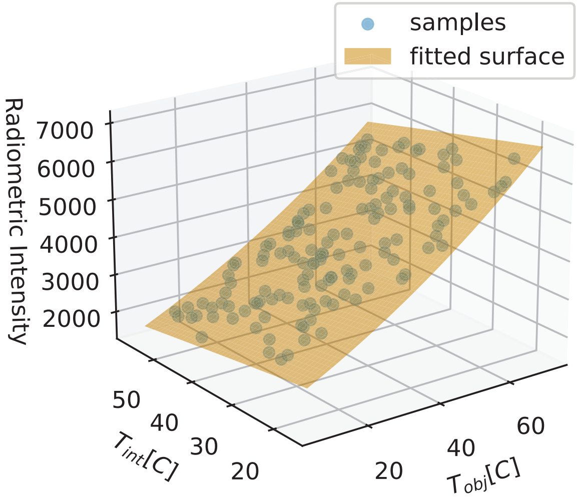


Figure 1. An example surface fit visualizing the calibrated polynomial coefficients of a single pixel.

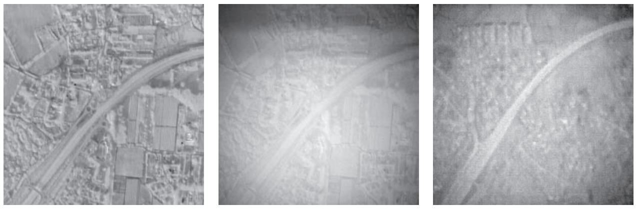
As it turns out, the calibrated physical estimator was not very accurate, and in particular suffered from a significant first-order error. Those inaccuracies might have had to do with issues related to the calibration setup, which would normally require an exhaustive investigation to find its root cause. To circumvent this cumbersome effort, we applied a pixel-wise affine transformation to the estimator, *i.e.*:

$$\tilde{G}_{phys} = A \circ G_{phys} + B \tag{14}$$

where A, B are matrices and $\circ$ is the Hadamard product operator. By constructing the elements of the matrices A, B as learnable parameters, back-propagation can be utilized to implicitly correct the physical estimator's prediction of the monochromatic output.

3.2. Deep estimator

Given the calibrated physical estimator, one might wonder why this is not enough to solve the UI2I task altogether. Unfortunately, as evident from equation 4, the emissivity can utterly change the incident thermal radiation on the camera's sensor. Therefore, two objects sharing the exact same temperature might result in significantly different bolometric readouts, and thus different intensity levels [9]. In addition, the application of an IR filter over the lens results in a scene-dependent spatial distortion known as the narcissus effect [14]. This effect is easily observable in real monochromatic images, such as those in Figure 2 in the monochromatic image. Hence, the physical estimator alone cannot accurately predict the intensity levels of a real-world scene, because it has no capacity to handle scene conditions that are different from its calibration setup. A demonstra-



(a) Panchromatic (b) Physical (c) Monochromatic

Figure 2. Demonstration of the Narcissus effect and the difference between the physical estimator's output and real monochromatic images. (a) Panchromatic (Pan) input. (b) Physical estimator (Phys) output. (c) Real unpaired monochromatic image for reference.

tion of the gap between the estimated output of the physical estimator and a real monochromatic image appears in figure 2.

This is where the family of deep generative I2I translation models, which have significantly greater capacity than the polynomial physical estimator, is brought into play. As baselines, we examined the architectures of CycleGAN [39] and CUT [23], which are considered to achieve SOTA results for the task of UI2I translation. Both CycleGAN's and CUT's generators consist of a convolutional encoder-decoder scheme with a bottleneck of residual blocks [7] in between, as schematized in Figure 3a.

As previously stated in equation 8, the intensity levels of both our input and target modalities are affected by the camera's intrinsic temperature. Fortunately, each image acquired by the FLIR Tau2 is saved along with the intrinsic thermal sensor readouts. Specifically for our activity, we chose to use the focal plane array temperature as our intrinsic temperature (T_{int}). Hence, it made sense to design an architecture that could accept this temperature as additional input. More concretely, we provide the panchromatic (input) intrinsic temperature to the encoder, and the monochromatic (output) intrinsic temperature to the decoder[1]. In doing so, we attempt to disentangle the intrinsic-temperature-dependent transformations (handled by the encoder and decoder) and the more general inter-modal transformation (handled by the bottleneck).

Our use of the intrinsic temperatures as inputs can be thought of as an extension of the concept of conditional GAN (CGAN) [19]. In our case, the output is conditioned on two continuous variables (panchromatic and monochromatic intrinsic temperatures) as opposed to the original paper where a single discrete conditional variable was used. Since both the encoder and decoder are convolutional networks, the intrinsic temperatures (scalars) were reshaped as constant matrices before being concatenated to the corresponding tensors.

3.3. Fusion of estimators

Although somewhat mitigated by the learnable affine transformation, the calibrated physical estimator still suffers from inaccuracies. Nevertheless, its prediction is much closer to the expected monochromatic output than a sheer random guess, which is the initial state of all ordinary GAN generators. Hence, we can use the physical estimator to produce a prior approximation of the desired output, and let the deep estimator learn the residual w.r.t. the desired result. This approach is expected to facilitate the deep estimator's pursuit of the optimal solution and make it more robust w.r.t. random weight initialization. Therefore, our proposed method fuses the physical estimator (augmented with the affine transformation) with the deep estimator:

$$G_{PETIT}(x) = \tilde{G}_{phys}(x) + G_{deep}(x) \tag{15}$$

where G_{deep} is used to describe the generator of the deep estimator. Schematics comparing the generator architecture of the deep baseline models (CycleGAN, CUT) and our model (PETIT) are shown in Figure 3.

4. Experiments

4.1. Dataset preparation

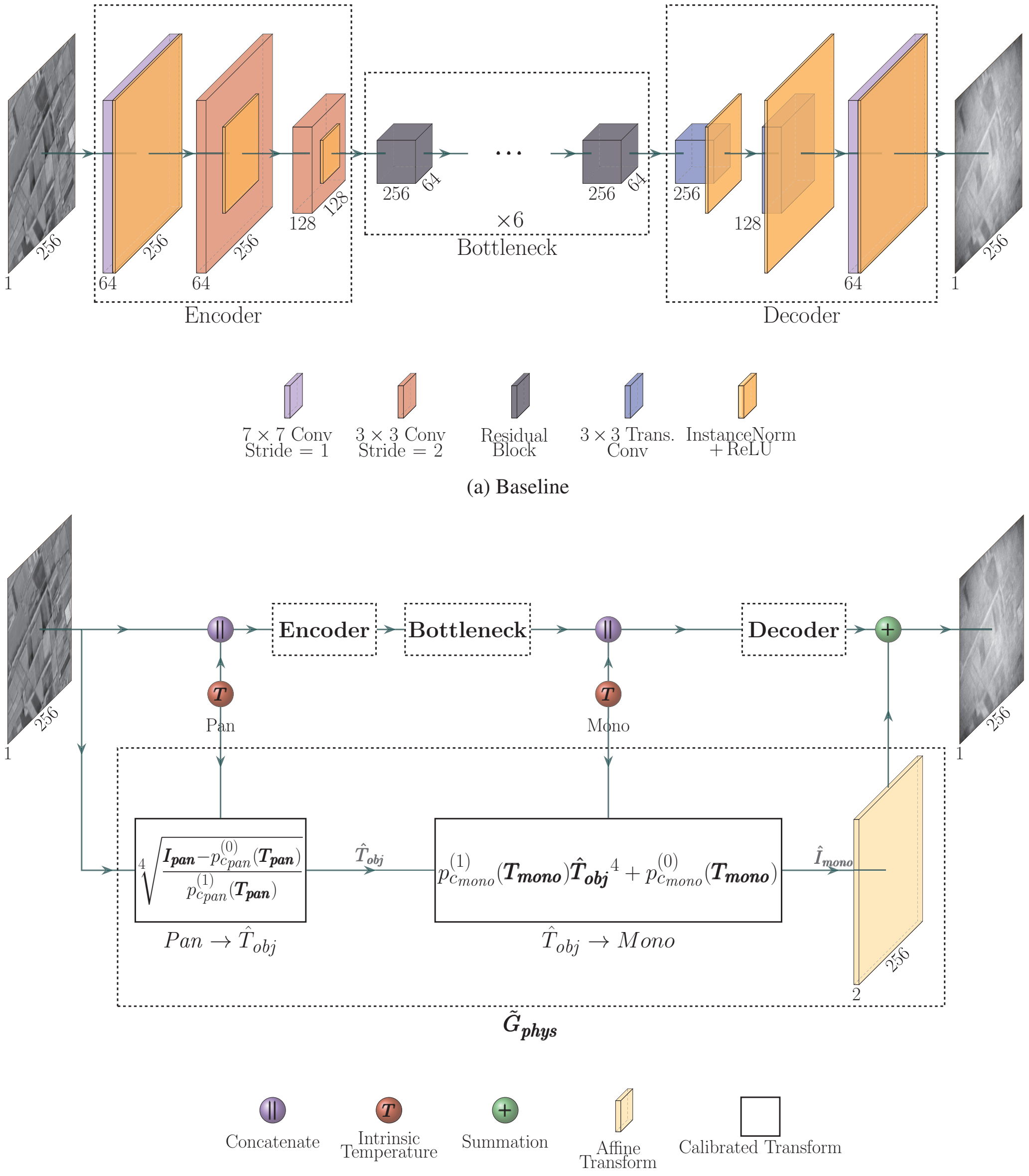
As mentioned in the introduction, there were no off-the-shelf thermal multispectral datasets available for training and testing our proposed method. The only remotely related available datasets were of satellite missions, but those are not open-sourced and only provide post-processed data (*e.g.*, estimated humidity) maps rather than raw thermal images. Therefore, we assembled a dedicated dataset in-house using a light airplane equipped with the same camera that was used to calibrate the physical model (FLIR Tau2). The pilot performed several flights, with a $9\mu m$ IR bandpass filter applied to the camera lens (which we refer to as *monochromatic* images), or without this filter (which we refer to as *panchromatic* images).

Ensuring that the plane trajectory and camera position are the same for two different flights is physically infeasible. Therefore, the monochromatic and panchromatic sets are necessarily unpaired. This guided us toward basing our method on an UI2I translation model as described in section 3.

4.2. Metrics

Since our dataset is unpaired, pixel-based metrics are not fit for performance evaluation. This implies that only statistically based metrics are viable. One such metric is the Fréchet Inception Distance (FID) [8], which is widely used

[1] In CycleGAN, the generator translating back from the monochromatic to the panchromatic domain receives the inputs in the reverse order, *i.e.*, monochromatic temperature to the encoder and panchromatic temperature to the decoder



(a) Baseline

(b) PETIT

Figure 3. Comparison between the baseline (CycleGAN, CUT) and PETIT (our method) generators. The architectures of the encoder, bottleneck and decoder blocks are identical in all models.

for evaluating the quality of generated images and specifically GANs. Typically, a lower FID score indicates greater fidelity of the generated images to the real target modality. We note here that while the density and coverage (DC) [20] metric is the most recent and popular approach for evaluating GANs, its coverage component is irrelevant for I2I tasks because the output's content is necessarily conditioned on the input.

One could argue that FID isn't an optimal metric of choice in our case, as the underlying network providing its score was pre-trained on RGB images rather than thermal ones. However, as shown Oz *et al*. [22] for super-resolution, neural networks trained with RGB data generalize very well for thermal images as well. Following this argument, it made sense to rely on FID as a proper measure of fidelity between thermal image modalities. Therefore, and in the absence of compellingly better alternatives, we use FID as our sole numerical evaluation metric.

4.3. Experimental setup

As a rule of thumb, the FID metric requires test sets of about $10,000$ images from each modality to be indicative of their true statistical distribution. This clearly limits the amount of data left for training and validation. Thus, only 1000 images were used for training and 100 for validation per modality, resulting in a total of 2200 images.

To fairly assess the method's contribution, we first trained our baseline models (CycleGAN, CUT) and fine-tuned their hyper-parameters and design choices. Consequently, the following changes were made (w.r.t. the original implementation) to achieve optimal FID score on our dataset: the generator's bottleneck was implemented using 6 residual blocks (instead of 9), the learning rate was set to 5×10^{-4} (instead of 2×10^{-4}) and a batch size of 2 was used (instead of 1). In addition, we applied a statistically based input-normalization technique, *i.e.*, the mean and standard deviation of each modality were calculated based on the entire training set, and then used to normalize all training, validation and test images.

The same set of hyper-parameters and design choices were applied as is for training all our proposed methods and their ablative configurations. For each model configuration, FID was calculated at the end of every epoch. The minimum (best) FID score over all epochs was treated as the score of that configuration. We note that since the FID is an evaluation metric, it should in essence be calculated only at test time, after training has already ended. Nevertheless, as the hyper-parameters were prefixed for all configurations, calculating the FID score had no impact over the optimization procedure. Therefore, there was no flaw in calculating it at the end of every epoch.

4.4. Results

4.4.1 Quantitative

Typically, numerical scores are obtained by training a network and evaluating it once. However, due to the highly non-convex nature of the deep networks loss functions, the local minimum to which a network converges is highly dependent on its weight initialization. The findings of [24] further illustrate that a network trained once (*i.e.*, with a single weight initialization) can have an outlier score that is much better or much worse than the average.

Therefore, we chose a statistical approach to evaluate and compare the different configurations. For each configuration, we trained and evaluated the FID score 10 consecutive times, where we randomly initialized the weights in every training-evaluation cycle. We then calculated the mean and standard deviation of the 10 FID scores and used them as criteria for comparison with the other configurations. This approach reduces the sensitivity to the randomness of the weight initialization, and provides a measure for the model's robustness w.r.t. random initialization, which can be inferred from the standard deviation.

The numerical comparison between the different configurations can be found in Table 1. We tested every possible configuration of our propositions, *i.e.*, with and without providing the intrinsic temperatures to the deep generator, and with and without fusing the deep generator with the physical estimator. All configurations were tested on top of both baseline models (CycleGAN and CUT).

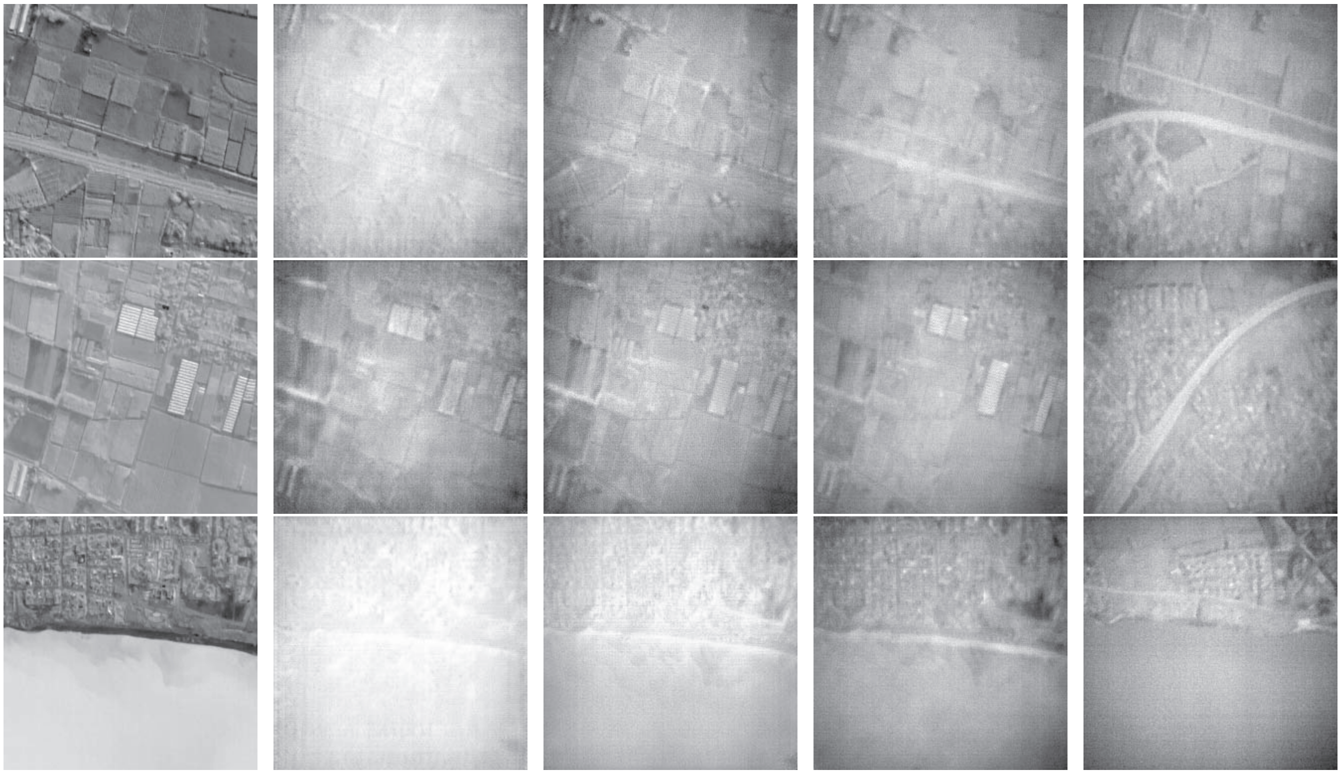
Configuration				FID	
Backbone	Int	Phys	Caption	Mean	Std
CycleGan	✗	✗	Baseline	51.05	9.82
	✗	✓		35.54	3.72
	✓	✗		50.17	8.89
	✓	✓	PETIT	**33.8**	**1.23**
CUT	✗	✗	Baseline	38.43	1.52
	✗	✓		29.85	**0.99**
	✓	✗		48.88	1.46
	✓	✓	PETIT	**27.35**	1.01

Table 1. Comparison of FID score statistics between the different configurations (the lower the better). *Int* stands for intrinsic temperature and *Phys* for physical estimator.

PETIT was found to dominate all other configurations with both backbones. Specifically, compared to the baseline configurations, PETIT achieved an approximately 50% mean FID improvement and a significantly improved standard deviation, indicating a more robust solution. Another interesting observation was that all configurations involving the physical estimator outperformed their counterparts by a large margin, in terms of both mean and standard deviation. On the other hand, the intrinsic temperature information alone did not seem to have a significant impact on the results, and was only helpful when combined with the physical estimator. This suggests that the physically estimated prior information steers the optimization procedure toward points on the manifold where the intrinsic temperature information is locally beneficial.

4.4.2 Qualitative

In accordance with the quantitative results, the monochromatic outputs produced by PETIT seem to be of superior quality compared to all other configurations. Generally speaking, PETIT's outputs incur less spurious artifacts and exhibit stronger fidelity to real monochromatic modality. An impression of the discussed superiority can be obtained from the examples in Figure 4. Due to space limitations, we only display the outputs of CycleGAN and CUT baselines *vs*. the CUT-backbone-based PETIT configuration. Real unpaired monochromatic images are also displayed in juxtaposition to the generated outputs for an impression of the modality's true nature. For additional examples, please refer to section 8 in the supplementary material.



(a) Pan (input) (b) CycleGAN (baseline) (c) CUT (baseline) (d) PETIT (ours) (e) Mono (ref)

Figure 4. Qualitative comparison. (a) Panchromatic (Pan) input. (b) CycleGAN output. (c) CUT output. (d) PETIT output. (e) Real unpaired monochromatic image (Mono) for reference.

5. Summary and conclusions

In this paper, we propose a novel method that takes advantage of the unique physical properties of our data domain to improve the performance of a generative model in an unpaired inter-modal translation task. Statistical analysis revealed a significant improvement in both performance and robustness with our approach. In particular, the significant contribution to both performance and robustness appears to be due to the fusion of a calibrated physical estimator with a deep generative architecture. This observation supports our hypothesis regarding the added value of a physically estimated prior information to a deep-generative model in generating more realistic outputs. The improved robustness might also indicate that the physical prior guides the deep model's convergence toward flatter minima. While the physical model enhancement was demonstrated for the backbones of CycleGAN and CUT, there is no limitation tying it to those specific models. Therefore, out method could in principle be harnessed to any type of generative architecture to improve its performance and robustness.

As in every work, our study leaves room for improvement and further investigation. While sufficiently helpful for improving the deep estimator results, any improvement in the calibration process or in the physical modeling could potentially improve our physical estimator's prediction and provide the deep estimator with a better initial approximation of the desired output. Another issue that requires investigation is the redundancy of the intrinsic temperature in the absence of the physical estimator.

Lastly, as mentioned in the methods section, our method is designed to transform an image from the panchromatic modality to a single monochromatic modality. By repeating the process for other monochromatic channels, a complete thermal multispectral dataset can be synthesized. Furthermore, instead of a single monochromatic target modality, our approach could be extended to apply multi-modal translation, *i.e.*, from panchromatic to several monochromatic modalities simultaneously, resulting in a complete multispectral transformation.

Acknowledgements

This work has been supported by the Israeli ministry of science and technology, research number 3-17381.

References

[1] Sebastian Bujwid, Miquel Martí, Hossein Azizpour, and Alessandro Pieropan. Gantruth - an unpaired image-to-image translation method for driving scenarios, 2018. 1

[2] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. 1

[3] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *ArXiv*, abs/1807.09384, 2018. 1

[4] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4354–4362, 2017. 2

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. 5

[9] Jack Philip Holman. *Heat Transfer [SI Metric Ed.]*. McGraw-Hill, 1989. 4

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei 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. 2

[11] Andrew Kalman, Adam Reif, Dan Berkenstock, Julian Mann, and James Cutler. Misc™–a novel approach to low-cost imaging satellites. 2008. 1

[12] John P Kerekes, Kristin Strackerjan, and Carl Salvaggio. Spectral reflectance and emissivity of man-made surfaces contaminated with environmental effects. *Optical Engineering*, 47(10):106201, 2008. 3

[13] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 2

[14] Michael Vollmer & Klaus-PeterMöllmann. *Fundamentals of Infrared Thermal Imaging*, chapter 1, pages 1–106. John Wiley & Sons, Ltd, 2017. 3, 4

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[16] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 2

[17] Rui Li, Wenming Cao, Qianfen Jiao, Si Wu, and Hau-San Wong. Simplified unsupervised image translation for semantic segmentation adaptation. *Pattern Recognition*, 105:107343, 2020. 2

[18] Christopher Lowe, Malcom Macdonald, Steve Greenland, and David Mckee. 'charybdis'–the next generation in ocean colour and biogeochemical remote sensing. 2012. 1

[19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 5

[20] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 6

[21] Paul W. Nugent, Joseph A. Shaw, and Nathan J. Pust. Correcting for focal-plane-array temperature dependence in microbolometer infrared cameras lacking thermal stabilization. *Optical Engineering*, 52(6):1 – 8, 2013. 3

[22] Navot Oz, Nir Sochen, Oshry Markovich, Ziv Halamish, Lena Shpialter-Karol, and Iftach Klapp. Rapid super resolution for infrared imagery. *Opt. Express*, 28(18):27196–27209, Aug 2020. 6

[23] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 1, 2, 5

[24] David Picard. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021. 7

[25] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1

[26] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. UNIT-DDPM: unpaired image translation with denoising diffusion probabilistic models. *CoRR*, abs/2104.05358, 2021. 1

[27] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–23, 2017. 2

[28] L.M. Surhone, M.T. Timpledon, and S.F. Marseken. *Stefan-Boltzmann Law: Stefan-Boltzmann Law, Black Body, Irra-*

diance, Thermodynamic Temperature, Ultraviolet Catastrophe, History of Quantum Mechanics, Thermodynamics. Betascript Publishing, 2010. 3

[29] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A. Hicks, Hugo L. Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A. Riegler. SinGAN-seg: Synthetic training data generation for medical image segmentation. *PLOS ONE*, 17(5):e0267976, may 2022. 1

[30] Yu-Hsuan Tu, Kasper Johansen, Bruno Aragon, Marcel M El Hajj, and Matthew F McCabe. The radiometric accuracy of the 8-band multi-spectral surface reflectance from the planet superdove constellation. *International Journal of Applied Earth Observation and Geoinformation*, 114:103035, 2022. 1

[31] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I Chang, Yan Xu, et al. Mri cross-modality neuroimage-to-neuroimage translation. *arXiv preprint arXiv:1801.06940*, 2018. 2

[32] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[33] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 2

[34] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 2

[35] Yongbing Zhang, Siyuan Liu, Chao Dong, Xinfeng Zhang, and Yuan Yuan. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE transactions on Image Processing*, 29:1101–1112, 2019. 2

[36] Yang Zhao and Changyou Chen. Unpaired image-to-image translation via latent energy transport. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16418–16427, 2021. 1

[37] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *European Conference on Computer Vision*, pages 800–815. Springer, 2020. 1

[38] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16407–16417, 2021. 2

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1, 2, 5