Perspectives and limitations of visible-thermal image pair synthesis via generative adversarial networks

Danick Panchard[a], François Marelli[a,b], Edouard De Moura Presa[c], Peter Wellig[c], and Michael Liebling[a,d]

[a]Idiap Research Institute, Martigny, Switzerland
[b]École Polytechnique Fédérale de Lausanne, Switzerland
[c]armasuisse Science and Technology, Thun, Switzerland
[d]University of California, Santa Barbara, CA, USA

ABSTRACT

Many applications rely on thermal imagers to complement or replace visible light sensors in difficult imaging conditions. Recent advances in machine learning have opened the possibility of analyzing or enhancing images, yet these methods require large annotated databases. Training approaches that leverage data augmentation via simulated and synthetically-generated images could offer promising prospects. Here, we report on a method that uses generative adversarial nets (GANs) to synthesize images of a complementary contrast. Starting from a dual-modality dataset of co-registered visible and thermal images, we trained a GAN to generate synthetic thermal images from visible images and vice versa. Our results show that the procedure yields sharp synthesized images that might be used to augment dual-modality datasets or assist in visual interpretation, yet are also subject to the limitations imposed by contrast independence between thermal and visible images.

Keywords: thermal imaging, image synthesis, generative adversarial networks

1. INTRODUCTION

Many applications in defense, surveillance, rescue, or control rely on thermal imaging systems whose signal-to-noise or resolution are often limited. Limitations can be due to the physical nature of the imaged radiation or the specificities of the use case, such as low light or poor atmospheric conditions. Computational imaging approaches have the potential to improve both spatial and temporal resolution as well as to increase the dynamic range of the imaging systems. Various methods have been developed that take advantage of multiple views of the same object,[1,2] or a priori knowledge about the signal, for example, its sparsity in a predefined space[3–5] or properties inferred through machine learning-based approaches that rely on a large body of training examples.[6–9] Both end-to-end methods[10] and optics-constrained methods[11] are possible approaches. With the need for large sets of annotated training data, which is shared with other fields,[12] methods to generate diverse image datasets without requiring extensive collection and annotation campaigns are desirable. The prospect of being able to generate images of one contrast type from an image of a different contrast therefore seems appealing. This appears particularly important as the training can have a dramatic impact on performance.[11,13]

Various methods have been proposed, specifically to synthesize visible-like images from thermal images and vice-versa. These include works to synthesize RGB (visible) images from thermal images for face recognition applications[14] or driving scenes.[15] Conversely methods have also been proposed to generate thermal images from visible images.[16]

Synthesized images can have many applications of their own, for example to facilitate interpretation of a thermal contrast by complementing it with a visible image (e.g. to provide context by translating thermal images[17]), with which the viewer might be more familiar with, or to highlight possible discrepancies in the measured image compared to an expected one. Since our main goal is in improving resolution, we more specifically focus on the ability of the synthesized methods to be used for training purposes of machine learning models.

Further author information: michael.liebling (at) idiap.ch

In the present preliminary study, we have selected a generic image-to-image translation method[18] that we trained on visible-thermal image pairs from a dataset acquired for automated driving and pedestrian detection.[19,20] We were specifically interested in:

1. determining if the trained model could be used as-is with images of a similar contrast but different content from that in the training dataset (e.g. training a model on images of urban scenes yet applying it to natural scenes)
2. identifying specific cases in which synthesizing thermal images from visible images would be an inherently ill-posed problem.

This paper is organized as follows. In Section 2, we detail the implemented visible-to-thermal and thermal-to-visible methods. In Section 3, we provide visual examples of synthesized images using various image types as input (same and different from the type used for training). Finally, we discuss the results in Section 4 and we conclude in Section 5.

2. METHODS

2.1 Overview

Generative adversarial networks (GANs)[21] are at the core of several methods to produce photo-realistic single-image super-resolution, e.g.[8] Beyond photography, applications also include astronomy.[9] These methods offer the prospect of pushing imaging boundaries with robust and simple to use techniques. However, their reliance on many examples and their departure from physical models limits trust in the results they produce. Nevertheless, their use appears particularly promising in the area of noise reduction.[22] GANs are also used extensively for image synthesis.

Here, we trained an image-to-image translation conditional GAN proposed by Isola *et al.*,[18] which we adapted for contrast synthesis (thermal to visible or visible to thermal). For training and visual evaluation we identified the CVC-14[19,20] dataset that contains co-registered visible and thermal images (visible (gray scale) and thermal images, taken from a vehicle, 6500 images). We implemented the translation algorithm in PyTorch and trained it *de novo* on the above data set using a GPU grid. We further evaluated the generalizability using an arbitrary thermal image acquired with a dual sensor thermal imagers (SOPHIE LITE, Thalès, France).

2.2 Image Synthesis using cGAN

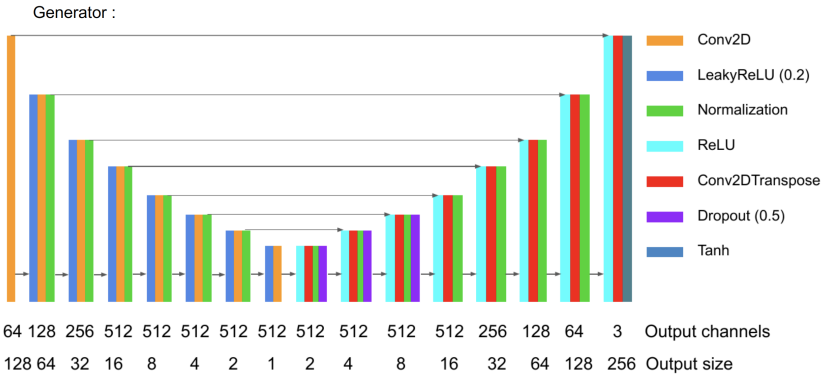

Figure 1. Architecture of the GAN Generator for image synthesis

We implemented an image to image translation algorithm based on the approach described by Isola *et al.*[18] using generator and discriminator networks described in Figures 1 and 2, respectively.

The translation could be done in either of two ways: (i) starting from thermal images, to synthesize images that look similar to those acquired with a camera in the visual range, or vice-versa (ii) starting from images in the visual spectrum, to synthesize thermal-alike images.

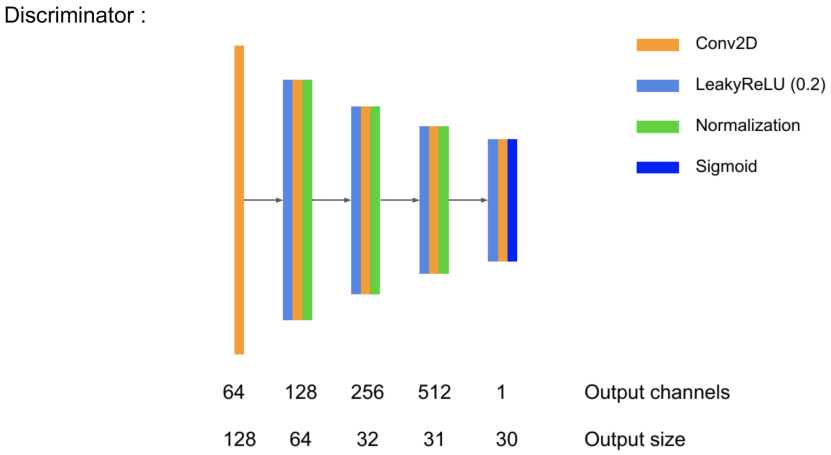

Figure 2. Architecture of the GAN Discriminator for image synthesis

The core idea is that one trains a model (i.e. a convolutional neural network) that takes in an image of the source modality as input and generates an image of the target modality as the output. This model is called the generator. Simultaneously, one trains a model (again, a neural network) that takes an image of the target modality as the input and whose output is to tell whether this images is synthetic (generated synthetically) or genuine (acquired by a camera). This latter model is the discriminator. During the simultaneous training of the generator and the discriminator, the generator is penalized whenever it is unable to synthesize an image that the discriminator identifies as likely being synthetic and is rewarded if it generates an image that the discriminator identifies as being likely genuine. During training, the discriminator is either fed synthesized images or genuine images, with correct and incorrect classifications as synthetic or genuine being rewarded and penalized, respectively. We specifically considered image-to-image conditional GANs, which further provide the discriminator with the source image corresponding to the target image (in addition to the image of the target modality) for it to proceed with the synthetic/genuine classification. In such a model, the generator is penalized using the L_1 distance between the synthetic and the genuine target image, in addition to the penalty related to the decision of the discriminator.

Concretely, for the case where one wishes to synthesize visible-alike images from thermal images, the training proceed as follows. The generator is fed a thermal image and produces (given the current state of its parameters) a visible-alike image. This synthetic visible-alike image is entered into the discriminator (in its current training state), along with the thermal image. The discriminator returns the probabilities that the entered image is synthetic. Since, in this example, the image is synthetic, depending on the probability score, both the discriminator and the generator will be adjusted. Further rounds of training (with other input images) are necessary to complete the training. The discriminator is also regularly fed genuine visible images, as if they had been generated from the thermal images (to train the discriminator to get better at discriminating synthetic images).

3. RESULTS

We used the CVC-14[19,20] dataset and split it as follows: 3140 images that come from an IR camera and corresponding visible camera for training and 550 images that come from an IR camera and corresponding grayscale camera for testing. All pictures were taken from a vehicle and had a resolution of 640 by 480 pixels.

The results of the generation are described in Figures 3 and 4 for synthesis of thermal from visible images and synthesis of visible from thermal images, respectively. We further applied the technique on thermal-visible pair that we acquired with a dual visible and thermal imager (SOPHIE-LITE; Fig. 5).

Visible | Thermal (measured) | Synthetic Thermal (generated from visible)

Figure 3. Generation of synthetic thermal images forma visible camera image. Results for models trained over 150 epochs on GPU (epoch duration about 5 minutes, image input sized 512 × 512). Note that although the test set is formally separate from the training set, when the imaging vehicle is stopped (first row), very similar images may still appear in the test and training sets, leading to results that suggest overly good performance on reconstruction of static objects.

4. DISCUSSION

Image-to-image translation is an area where learning-based methods have an important role to play and where their generic structure has practical relevance as the same architecture could be used for a variety of contrast conversions.

Figure 4. Generation of synthetic visible images from a visible camera image. Results for models trained over 150 epochs on GPU (epoch duration about 5 minutes, image input sized 512 × 512). Note that although the test set is formally separate from the training set, when the imaging vehicle is stopped (second row), very similar images may still appear in the test and training sets, leading to results that suggest overly good performance on reconstruction of static objects.

GAN-based contrast synthesis (thermal to visible or visible to thermal) exploits common information in training image pairs. While the synthesis methods only approximately emulate the physical measurements unique to each thermal range, they may provide a baseline to highlight anomalies in a common modality. Both the training data set size and its similarity to intended application are important to successfully train the models. Pre-training on large image datasets offers benefits when few application-specific images are available.

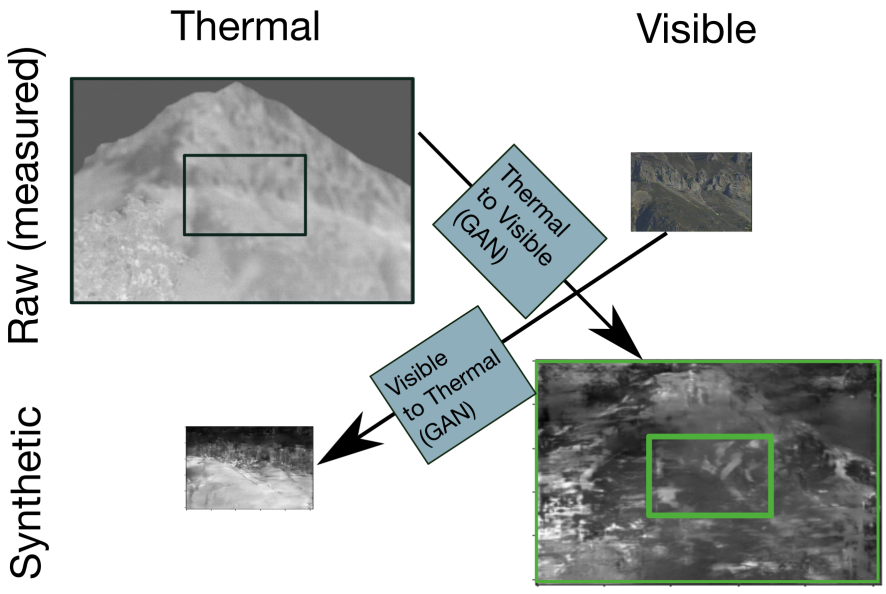

Figure 5. Example of synthetic generation of cross contrast using GANs. Images of the Grand Chavalard mountain in Switzerland (from a distance of 8.4km); raw images were acquired with SOPHIE LITE thermal imager; the synthetic images were generated with a conditional GAN model[18] trained on the CVC-14 (visible-thermal) dataset.[19]

In the case of thermal to visible or visible to thermal synthesis, we noticed that the synthesized images gave a "qualitative look" that indeed resembles the actual target modality, yet without offering much accuracy. This result is explainable by the fact that there is no one-to-one correspondence between features in one modality with features in another modality. The wavelength-ranges (thermal or visible) being different, they provide unique features that capture unique physical characteristics. As such one should therefore not expect that contrast synthesis can make one or the other modality redundant. We nevertheless see applications for such conversions. Synthetic images could, in particular, be used to generate "expected images" from one modality, which could be compared to the measured images in order to highlight the specific areas that provide information that is unique to one or the other modality, thereby potentially facilitating detection.

While synthetic images produced with models trained on datasets that contained a large number of examples that were similar (same cameras) as the target application are generally good (Figure 3 and Figure 4), the results of fairly naively applying the trained models to images from a different camera and from a different type of non-urban scene (Figure 5) appear to provide much less accurate results. For such applications to be truly relevant in practice, the training sets should likely match the target application and be of sufficient size. Nevertheless, taking the models trained on images of a specific use case as a starting point, transfer learning would likely permit to reduce the amount of data necessary for training the model for a different use case.

Another area of concern is that the visible to thermal and thermal to visible image conversion is inherently ill-posed. To illustrate this point, we acquired thermal images and their visible counter-part, with several visually similar objects in the scene, yet whose temperature was variable (Fig. 6). Specifically, the images, shot on a dry Winter morning, contain two cars whose recent use is not apparent in the visible-light image. Thermal imaging reveals that the temperature of one car is uniform (as it has remained parked overnight) whereas the wheels and the hood of the other car exhibit a higher temperature than the rest of the car's body. While training sets that contain either situation (e.g. through manual curation) might be used for training in order to reproduce one or the other scenario, generation of thermal images from visible images with a network trained on mixed data could likely create unpredictable results.

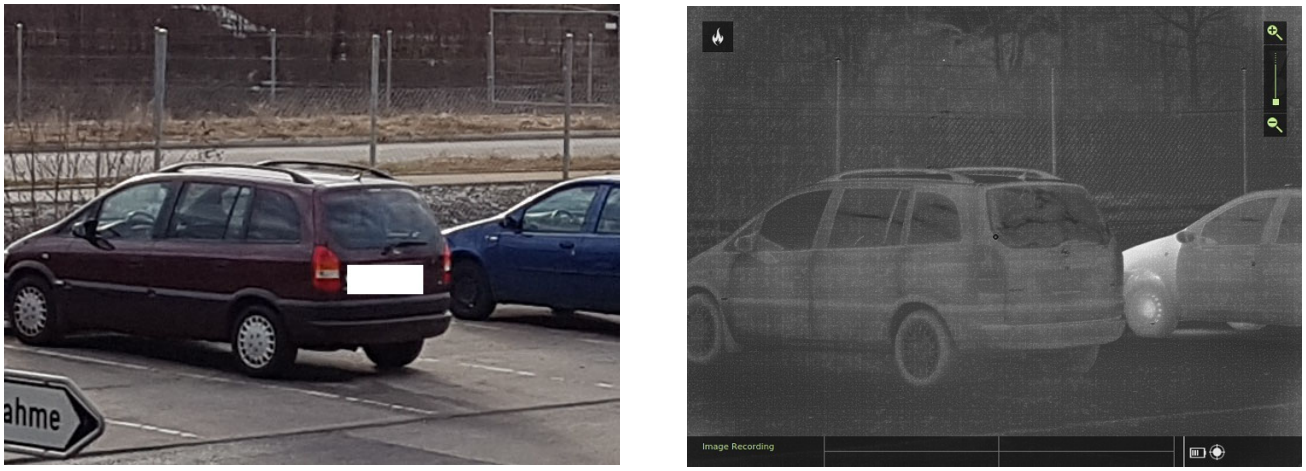

Figure 6. Illustration of ill-posedness of the image-to-image translation problem for visible to thermal conversion. The two images (thermal on the left, visible on the right) were taken at around 9am at an outside temperature of 6°C. The car on the left had remained parked overnight while the car on the right had just recently been used.

5. CONCLUSION

This preliminary study highlights both the encouraging potential of image synthesis but also reminds of the inherent limitations that the image translation task bears. Generalization limitations could be alleviated by using use-case specific images, but these may precisely be lacking if the goal is to use image synthesis to expand a training dataset. A less constraining approach would be to consider Cycle-Consistent Adversarial Networks,[23] which do not require paired training data. Nevertheless, limitations related to the inherent physical ill-posedness of the problem remain. The ability to produce reliable radiometric data appears compromised with an unconstrained translation process.

REFERENCES

[1] Šroubek, F., Cristóbal, G., and Flusser, J., "A unified approach to superresolution and multichannel blind deconvolution," *IEEE Trans. Image Process.* **16**, 2322–32 (Sep 2007).

[2] Takeda, H., Milanfar, P., Protter, M., and Elad, M., "Super-resolution without explicit subpixel motion estimation," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* **18**, 1958–75 (Sep 2009).

[3] Donoho, D. L., "Superresolution via sparsity constraints," *SIAM J. Math Analysis* **23**, 1309–1331 (Sept. 1992).

[4] Candès, E., Romberg, J., and Tao, T., "Stable signal recovery from incomplete and inaccurate measurements," *Commun Pur Appl Math* **59**, 1207–1223 (Jan 2006).

[5] Tropp, J., "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE T Inform Theory* **52**, 1030–1051 (Jan 2006).

[6] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *Nature* **521**(7553), 436–444 (2015).

[7] Dong, C., Loy, C. C., He, K., and Tang, X., "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(2), 295–307 (2016).

[8] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," (2016).

[9] Schawinski, K., Zhang, C., Zhang, H., Fowler, L., and Santhanam, G. K., "Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit," *Monthly Notices of the Royal Astronomical Society: Letters* **467**(1), L110–L114 (2017).

[10] Dong, C., Loy, C. C., He, K., and Tang, X., "Image super-resolution using deep convolutional networks," *CoRR* **abs/1501.00092** (2015).

[11] Shajkofci, A. and Liebling, M., "Spatially-variant CNN-based point spread function estimation for blind deconvolution and depth estimation in optical microscopy," *IEEE Trans. Image Proces.* **29**, 5848–5861 (2020).

[12] Shajkofci, A. and Liebling, M., "Free annotated data for deep learning in microscopy? a hitchhiker's guide," *Photoniques* , 30–33 (September-October 2020).

[13] Ciolino, M., Noever, D., and Kalin, J., "Training set effect on super resolution for automated target recognition," in [*Automatic Target Recognition XXX*], Hammoud, R. I., Overman, T. L., and Mahalanobis, A., eds., **11394**, 105 – 117, International Society for Optics and Photonics, SPIE (2020).

[14] Kezebou, L., Oludare, V., Panetta, K., and Agaian, S., "TR-GAN: thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition," in [*Mobile Multimedia/Image Processing, Security, and Applications 2020*], Agaian, S. S., Asari, V. K., DelMarco, S. P., and Jassim, S. A., eds., **11399**, 158–168, International Society for Optics and Photonics, SPIE (2020).

[15] Berg, A., Ahlberg, J., and Felsberg, M., "Generating visible spectrum images from thermal infrared," in [*Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR) Workshops*], (June 2018).

[16] Mizginov, V. A., Kniaz, V. V., and Fomin, N. A., "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).

[17] Liu, S., John, V., Blasch, E., Liu, Z., and Huang, Y., "Ir2vi: Enhanced night environmental perception by unsupervised thermal image translation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (June 2018).

[18] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*Proc. IEEE conf computer vision pattern recognition*], 1125–1134 (2017).

[19] ADAS-CVC, "CVC-14: Visible-FIR day-night pedestrian sequence dataset," (2016).

[20] González, A., Fang, Z., Socarras, Y., Serrat, J., Vázquez, D., Xu, J., and López, A. M., "Pedestrian detection at day/night time with visible and fir cameras: A comparison," *Sensors* **16**(6), 820 (2016).

[21] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial networks," (2014).

[22] Krull, A., Buchholz, T., and Jug, F., "Noise2void - learning denoising from single noisy images," *CoRR* **abs/1811.10980** (2018).

[23] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in [*Proceedings of the IEEE international conference on computer vision*], 2223–2232 (2017).