



# Toward a Thermal Image-Like Representation

Patricia L. Suárez<sup>1</sup> <sup>a</sup> and Angel D. Sappa<sup>1,2</sup> <sup>b</sup>

<sup>1</sup>*Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ingeniería en Electricidad y Computación, CIDIS, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador*

<sup>2</sup>*Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra, Barcelona, Spain*

**Keywords:** Contrastive Loss, Relativistic Standard GAN Loss, Spectral Normalization.

**Abstract:** This paper proposes a novel model to obtain thermal image-like representations to be used as an input in any thermal image compressive sensing approach (e.g., thermal image: filtering, enhancing, super-resolution). Thermal images offer interesting information about the objects in the scene, in addition to their temperature. Unfortunately, in most of the cases thermal cameras acquire low resolution/quality images. Hence, in order to improve these images, there are several state-of-the-art approaches that exploit complementary information from a low-cost channel (visible image) to increase the image quality of an expensive channel (infrared image). In these SOTA approaches visible images are fused at different levels without paying attention the images acquire information at different bands of the spectral. In this paper a novel approach is proposed to generate thermal image-like representations from a low cost visible images, by means of a contrastive cycled GAN network. Obtained representations (synthetic thermal image) can be later on used to improve the low quality thermal image of the same scene. Experimental results on different datasets are presented.

## 1 INTRODUCTION

In recent year, thermal imaging has increasingly being used in a range of different fields in industry, which has lately led to the manufacturing of low-cost thermal vision sensors. Low-cost thermal sensors are fast becoming available, and they are making their way into applications other than heavy industrial usage, such as surveillance, criminal investigation, military use, medical research, and building maintenance. Exploiting these alternate perspectives has the potential to play a significant role in computer vision by improving the accuracy of our existing conventional digital vision.


Although thermal sensors have come a long way, there is still a bottleneck related with the poor resolution, the cost of thermal cameras grows exponentially with greater resolution. In general, their resolutions are substantially lower than those of regular digital cameras working in the visible spectrum. Hence, it is critical to discover ways to leverage the information from sensors working at different spectral bands and combine them together to maximize their advantages.


Recently, some approaches have been proposed trying to use visible images information to enhance

the other domain images and hence produce images with a higher quality and close to human perception at a lower cost. These approaches are referred to in the literature as Guidance Image Processing (e.g., (Kopf et al., 2007), (Hui et al., 2016), (Barron and Poole, 2016)). Most of guidance based method fuse the provided information at different levels, but in almost all the cases the given images (e.g., thermal and visible) are used to feed the model without any concern on their nature/difference—they capture information from different spectral bands.

There are some approaches where the guidance is not performed at an image pixel level but at a feature level, for instance edges from one image are used to enhance the other image. The use of edge-based guiding facilitates the reconstruction of higher-frequency features (e.g., (Xie et al., 2015), (Zhou et al., 2018)). As mentioned before, the different nature of provided images reduce the possibility of taking the best from each representation. In the current work we propose to generate thermal image-like representations from visible spectrum images in order to facilitate the further guided process, since both images will be represented in a closer domain (thermal image domain).

Most of the approaches mentioned above are deep learning based solutions, which by means of efficient Convolutional Neural Networks (CNNs) significantly

<sup>a</sup>  <https://orcid.org/0000-0002-3684-0656>

<sup>b</sup>  <https://orcid.org/0000-0003-2468-0031>

outperform traditional methods. A fundamental aspect of CNNs is the large volume of data are required for their training. Furthermore, in cases like the one tackled in the current work, the existence of paired images (thermal and visible spectrum) is required. Having in mind these drawbacks (large amount of data, and the existence of paired set) in the current work a model capable of generating synthetic thermal images from its counterpart in visible space is proposed. The model is trained with unpaired set of thermal and visible images, which represent a great advantage with respect most of the state-of-the-art approaches. The contribution of this paper can be pointed out in:

- The usage of contrastive loss (Liu et al., 2021), to enhance the feature extraction of the generator. With the introduction of this loss, the learning of the model is favored from similar regions of the images contrasted with latent spaces with low affinity. In the proposed model, a combination of this loss with identity and adversarial loss is proposed.
- The modification of the architecture to use spectral normalization instead of batch normalization to improve the stylization of the images and avoiding fading of gradients (Miyato et al., 2018).
- The implementation of relativistic GAN to facilitate that the generated samples are closer to the decision limit of the model. This allows the model to generalize more quickly and improves the quality of the images.

The manuscript is organized as follows. Section 2 presents works related with the generation of synthetic images to solve related problems. Section 3 presents the proposed cycled GAN modified architecture. Experimental results and comparisons with different implementations are given in Section 4. Finally, conclusions are presented in Section 5.

## 2 RELATED WORK

Several approaches have been proposed to generate synthetic images to be used in the training of models that need thermal spectrum data sets, or to reinforce the training of other techniques that solve issues related to control, detection, classification, among others. In (Guo et al., 2019), an approach related with pedestrian detection in thermal imaging scenarios is proposed. It tackles the limitations of current data sets by generating synthetic thermal images from their widely available visible counterpart applying domain

matching. To generate the synthetic data set, a component has been created that performs the transformation from the visible to the far infrared domain together with the bounding boxes of the detected pedestrians. It is implemented through a cycled GAN network that serves as a data augments when training the pedestrian detection model in the thermal domain. Although interesting results are obtained the approach is not focused on the quality of the generated synthetic thermal images but on the pedestrian detection application.

Another use of the synthetic thermal images generated by CNN models is the one presented in Zhang et al. (Zhang et al., 2018), in that paper it is proposed to use the synthetic images to train a tracking model. The authors apply the transformation of paired and unpaired images. With these images, the results of the tracking model with thermal images are improved. Given the evolution of autonomous driving based on LIDAR sensors, there are some approaches that have been proposed. Therefore, in Lu et al. (Lu and Lu, 2021) the authors have designed a scheme to estimate the depth of the scene based on synthetic thermal images generated from RGB. The synthetic images are obtained by means of a cycled GAN network that performs the translation from visible to thermal domain together with a disparity map to maintain the consistency relation of the generated images.

In Liu et al. (Liu et al., 2021), the authors propose a method to improve scene context for night vision applications. This model uses synthetic images generated from visible spectrum images. A GAN network is used for mapping context information, generating synthetic images with higher quality. It allows improving the capture of fine details in synthetic images, used to enhance the context of scenes. In another approach, proposed by Li et al. (Li et al., 2020), a semantic image segmentation technique is presented. The authors propose to overcome lighting and environmental limitations by using images from both, real and synthetic thermal infrared cameras, to guide the contour extraction. A synthetic image dataset has been generated by a modified pix2pix image transformation proposed by (Isola et al., 2017). These synthetic images allow to improve the results of the training for the cases in which the visible images present limitations.

Another technique that uses thermal images is the one presented in (Kniaz et al., 2018), where a cross-modal generative network is proposed to generate synthetic thermal images that serve as a support for training a people reidentification model. This model introduces object notations to improve the results of people re-identification. Another application of syn-

thetic data is the one presented in Saleh et al., (Saleh et al., 2019) where the use of point cloud data from 3D LiDAR sensors for critical safety tasks applied to autonomous vehicle systems is introduced. Given the existing techniques to generate synthetic images, the authors propose to generate the point cloud from synthetic images using a cycle GAN combined with the real images obtained by 3D LiDAR sensors. This allows to improve the detection results of vehicles from a bird's eye view.

In the context of synthetic image generation, Generative Adversarial Networks (GANs) have facilitated the process and helped a lot in the area of computer vision. The transformation of information between domains is the main functional that these generative networks achieve. There are currently many types of generative networks, but we are going to mention only the once more related to the context of the problem to be solved in this paper. The learning of these classes of networks is done through the training process with a correctly registered data set. However, obtaining the paired information is sometimes not possible or, in case it is possible, some times it is difficult to acquire enough data to train the network. These limitations have motivated works such as the one presented in Zhu et al., (Zhu et al., 2017), to be able to generate images from one domain to another, without the need for them to be registered, the author in this work present the cycle consistency loss which allow the unpair image transformation known as cycled GAN. To address our proposed problem, we are going to use a set of images from the visible spectrum and map them to the thermal spectrum. Therefore, the mapping function is:  $G: X \rightarrow Y$  such that the image distribution of  $G(X)$  is indistinguishable from the distribution of  $Y$  using contradictory loss, originally proposed in (Goodfellow Ian et al., 2014). However, for the domain translation proposed in (Zhu et al., 2017) this mapping is very loosely constrained, a reverse  $F$  mapping is necessary:  $Y \rightarrow X$  and introduce a loop consistency loss to enforce  $F(G(X)) \approx X$  (and vice versa). The next section details the changes with respect to the original Cycled GAN (Zhu et al., 2017) implemented in the current work to enhance the translation of information. Our proposal is motivated by the fact that the mapping must include not only shape, but also the textures and should simulate temperatures of objects, making the design of the architecture more challenging. The pre-processing applied to the dataset for the training process is also presented.

### 3 PROPOSED APPROACH

The proposed approach is a combination of several state of the art techniques adapted to the problem of obtaining synthetic thermal images from visible spectrum images. The architecture is based on a Cycled GAN to perform the transfer of unpaired domains that is presented in (Zhu et al., 2017). To achieve a closer translation to the intensity of the pixels of the far infrared spectrum, it is proposed to use the contrastive loss. This allows to improve the quality of the images. The inclusion of this contrastive loss, presented in (Liu et al., 2021), allows the proposed architecture to focus on determining the relationship between input embeddings from regions close to the region being processed. This method tries to predict the missing information based on its environment instead of predicting the values per pixel. To determine the similarity of nearby regions, cosine similarity is used. This makes it easy to determine the difference based on its orientation and not just the magnitude like the L1 loss. It must be considered that the discriminator of the model maximizes the replicas closest to the real one and minimizes the differences between the target image and the various embeddings of the processed nearby image regions. In this way, the discriminator is updated minimizing the distances between the real image and the embeddings of the same class while being maximized otherwise. By forcing the embeds to relate through the loss of cosine similarity, the discriminator can learn the detailed representations of real images. Similarly, the generator exploits knowledge of the discriminator, such as intraclass features and higher order representations of the actual images, to generate more realistic images. The use of this class of contrastive loss is already quite widespread, especially in the computer field of vision, where different approaches apply it (e.g., (Yu et al., 2021), (Liu et al., 2021), (Suárez et al., 2019), (Park et al., 2020)) given the good results obtained in improving the quality of the images.

It is important to emphasize that given the results obtained in the experiments carried out, the RGB color space of the input images of the model had to be changed to HSV. This is due to the fact that with the H channel of this color space it was possible to perform the transformation of the visible information to thermal with greater accuracy in the simulation of temperatures of the objects presented in the images. This allows not only to reflect good contours and details, but also to represent the temperatures of the generated synthetic thermal images with a high fidelity.

Additionally, in this paper we include a relativistic GAN loss (Jolicoeur-Martineau, 2018), instead of the

standard GAN loss proposed by (Goodfellow et al., 2020). This relativistic GAN loss assumes that in each mini-batch at least half of the data generated are false, which can be observed by minimizing the learning divergence. Therefore, we include this relativistic loss because allows estimating that in a mini-batch of randomly generated data, more realistic than false samples are obtained. According to the authors of relativistic loss, they argue that the probability of real data being real  $D(x_r)$  should decrease as the probability of fake data being real  $D(x_f)$  increase. Therefore, this forces the generated samples to be closer to the real ones, avoiding model saturation and also accelerate the training process. For this reason, this loss is very well coupled to the transformation process from visible to thermal and, according to the results we obtained, it was adapted in our cycled transformation model. The standard GAN loss function is replaced with the relativistic standard GAN loss for discriminator and generator respectively and they are defined as:

$$L_D^{RGAN} = \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [f_1(C(x_r) - C(x_f))] \quad (1)$$

$$L_G^{RGAN} = \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [f_1(C(x_f) - C(x_r))] \quad (2)$$

where  $f$  and  $g$  are functions mapping a scalar input to another scalar and  $x_r, x_f$  is the real and fake image respectively.

To replace the consistency cycle loss of the GAN network, see Eq. 3, a contrastive loss has been implemented in our architecture:

$$\mathcal{L}_{CYCLE}(G, F) = \mathbb{E}_{x \sim p_{\text{data}(x)}} [\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}(y)}} [\|G(F(y)) - y\|_1]. \quad (3)$$

This loss allows the model to be trained based on learning the similarity of the latency spaces resulting from the generating network. According with (Andonian et al., 2021), for each model input image  $\mathbf{x}$ , contrastive learning approaches only need to define the similarity distribution to sample a positive input  $\mathbf{x}^+ \sim p^+(\cdot | \mathbf{x})$ , and a data distribution for a negative input  $\mathbf{x}^- \sim p^-(\cdot | \mathbf{x})$ , with respect to a sample input  $\mathbf{x}$ .

Furthermore, they argue that the shape of the tensor  $V_l \in \mathbb{R}^{S_l \times D_l}$  is determined by the architecture of the network, where  $S_l$  is the number of spatial locations of the tensor. Therefore, the tensor is indexed with the notation  $v_l^s \in \mathbb{R}^{D_l}$ , which is the  $D_l$ -dimensional feature vector at spatial location  $s^{\text{th}}$ . It has been denoted  $\bar{v}_l^s \in \mathbb{R}^{(S_l-1) \times D_l}$  as the collection of feature vectors at all other spatial locations. According to (Andonian et al., 2021) this loss can be written

as follows:

$$\mathcal{L}_{\text{contrastive}}(\hat{Y}, Y) = \sum_{l=1}^L \sum_{s=1}^{S_l} \ell_{\text{contr}}(\hat{v}_l^s, v_l^s, \bar{v}_l^s)$$

The main objective of this type of training based on contrastive loss is that the model matches similar and different samples. Those that are similar or called positive should be mapped as close together as possible. On the other hand, the negative or dissimilar pairs must be further away from the positive latency space. These similar representations will become unified, while the dissimilar ones will separate from the latency space of the positive pairs.

In addition, the model also implements the identity loss function so that the intensity levels of the pixels do not go outside the bounds of the objective domain during the transformation of the data. This implies that the generating network must preserve the most relevant characteristics, learn the level of thermal intensity, the shape of the objects and help maintain the stability of the formation model. That is, it is true that  $F(x) \approx x$  and  $G(y) \approx y$ .  $\lambda$  is an aggregate term to define the relative importance of the cycle and identity losses, compared to the GAN:

$$\mathcal{L}_{\text{identity}}(G, F) = \mathbb{E}_{c \sim P_{\text{data}}(c)} [\|F(c) - c\|] + \mathbb{E}_{n \sim P_{\text{data}}(n)} [\|G(n) - n\|].$$

Finally the multiple loss function implemented in our model is defined as:

$$\mathcal{L}_{\text{RGAN}}(G, D, X, Y) + \lambda_X \mathcal{L}_{\text{Lcontrastive}}(G, H, X) + \lambda_Y \mathcal{L}_{\text{Lcontrastive}}(G, H, Y) + \gamma \mathcal{L}_{\text{Identity}}(G, F) \quad (4)$$

where  $\lambda, \gamma$  are the weights of the contrastive and identity loss function respectively, and have been defined empirically according to the results of the experiments.

The architecture (see Fig. 1) also includes a spectral normalization to improve the quality of the generated synthetic thermal images. This normalization has been implemented, given the challenge involved in training a GAN network is to control the performance of the generator and discriminant networks. The main goal is to avoid the fading of the gradients, so as not to collapse the training of the model. To improve the multimodal translation from the visible to the far infrared spectrum, spectral normalization has been introduced in the discriminator. This improves control of the efficiency of this network, avoiding learning instability. It also contributes to the generalization of the model in less time. This occurs because the discriminator is more efficient at distinguishing the target distribution pattern. With this normalization it is

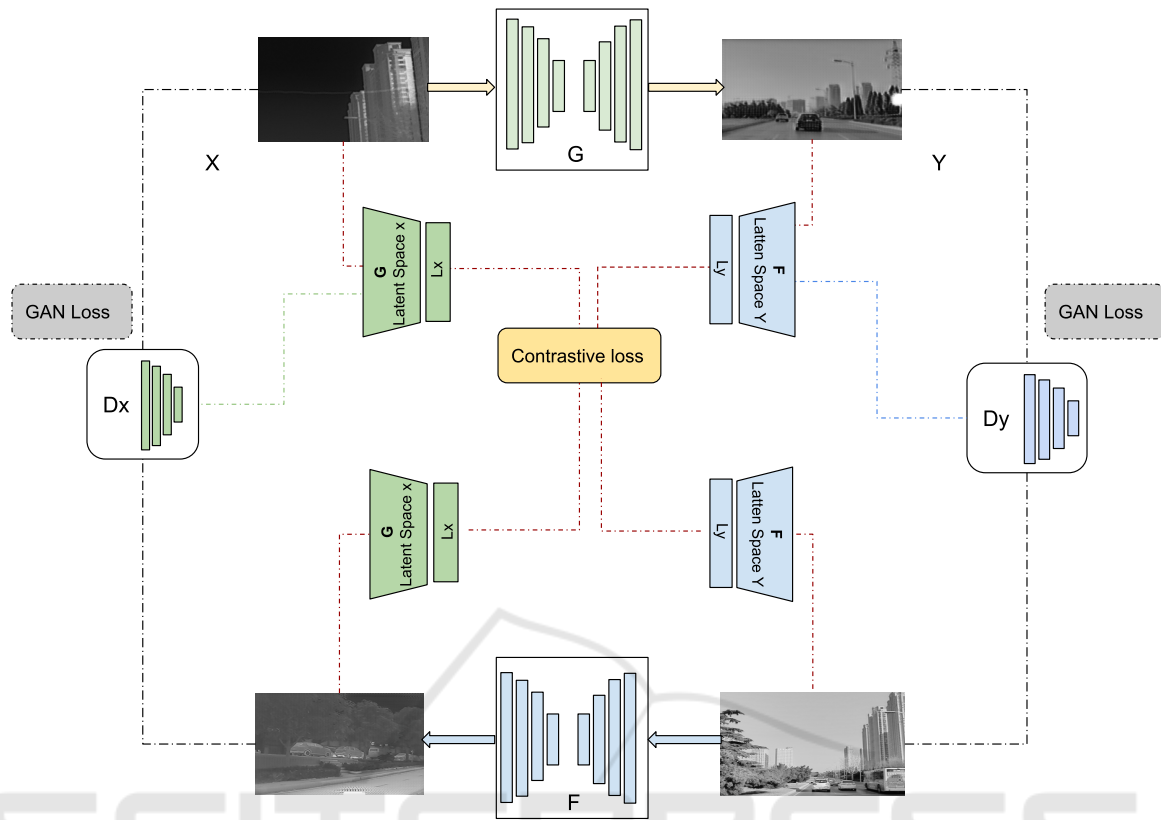


Figure 1: Cycle GAN proposed architecture.

avoided that the derivative of the discriminator network becomes zero and therefore the generating network stops its learning. This normalization acts as a choice constraint implemented in the discriminator network.

## 4 EXPERIMENTAL RESULTS

This section presents quantitative and qualitative results obtained with the proposed approach. It also includes a description of the data set used for training and the pre-processing applied to the images. Finally, a comparative analysis is carried out using the metrics of similarity and maximum noise present in the synthetic images.

### 4.1 Datasets

The designed model has been trained with the M3FD data set (Liu et al., 2022), whose acquisition has been carried out with a binocular optical and infrared sensor. The data set contains 4,500 image pairs of recorded outdoor scenes; from these images 3000 pairs were used for training, 890 pairs for testing and

the rest of the images for validation of the already trained model. It is worth mentioning that the images were pre-processed to generate the most realistic synthetic images possible in the far infrared spectrum. Therefore, the images were transferred to the HSV color space. Then the H channel that represents the (hue) has been selected as input to train model. Additionally, to validate the robustness of the model, a proprietary data set—referred to as Thermal Stereo—with 200 pairs of registered visible-thermal images has been considered. The results obtained with the model trained with the M3FD data set are included in the comparisons.

### 4.2 Training Settings

In order to train the model, the visible images have been converted to the HSV color space. Only the H channel has been considered to train the model. In addition, during the training the images have been resized to 256 x 256 pixels. For training the model the traditional GAN loss has been replaced with the relativistic GAN loss. The objective with this loss is that the discriminator globally evaluates the random samples against the real input data of the model.



Figure 2: Experimental results: (*1st. row*) results with state of the art technique (Zhu et al., 2017); (*2nd. row*) results from the proposed approach; (*3rd. row*) ground truth images from M3FD and Thermal Stereo datasets.

With this loss, a better quality of the generated images is achieved. The learning rate has been defined at 0.000273. The Adam optimizer has been used, where  $\beta_1$  and  $\beta_2$  have default values of 0.85 and 0.99, respectively. For quantitative evaluation, the maximum signal/noise ratio (PSNR) metric and the structural similarity index (SSIM) metric have been defined. A TITAN V GPU has been used for training. The training time of the model lasts about 96 hours.

### 4.3 Comparisons

The proposed approach has been evaluated by comparing it with the state-of-the-art model that performs the translation of unpaired images presented in (Zhu et al., 2017). This proposal allows generating synthetic images from images of the visible spectrum to another unpaired domain. Based on this concept, we have modified the loss functions and preprocessed the input images, in order to generate synthetic images of the thermal spectrum. Table 1 presents average results obtained with (Zhu et al., 2017) and the approach proposed in the current work. The model has been validated with samples from the M3FD data set and our own dataset taken from outdoor scenes. Figure 2, shows some illustrations of the synthetic thermal images obtained from these validation sets. Addition-

ally, for the purposes of quality comparison, Tables 2 and 3 show the best and worst results of the SSIM obtained with each data set—PSNR values are also provided. Furthermore, to illustrate these comparisons, Fig. 3 and Fig. 4 present the images that correspond to the metrics shown in these tables.

Table 1: Average results from the validation sets (M3FD-Thermal Stereo). Best results in **bold**.

Approaches	M3FD		Thermal Stereo	
	PSNR	SSIM	PSNR	SSIM
(Zhu et al., 2017)	12.589	0.501	11.939	0.419
Prop. Approach	<b>14.734</b>	<b>0.772</b>	<b>17.0989</b>	<b>0.733</b>

Table 2: Best and Worst SSIM results from the M3FD validation set. Best results in **bold**.

Approaches	M3FD			
	BEST		WORST	
	PSNR	SSIM	PSNR	SSIM
(Zhu et al., 2017)	17.381	0.631	8.90	0.3042
Prop. Approach	<b>22.899</b>	<b>0.869</b>	<b>11.279</b>	<b>0.638</b>



Figure 3: BEST and WORST results obtained with (Zhu et al., 2017): (1st. row) Images from M3FD dataset; (2nd. row) Images from Thermal Stereo dataset.



Figure 4: BEST and WORST results obtained with the proposed approach: (1st. row) Images from M3FD dataset; (2nd. row) Images from Thermal Stereo dataset.

Table 3: Best and Worst SSIM results from the Thermal Stereo validation set. Best results in **bold**.

Approaches	THERMAL STEREO			
	BEST		WORST	
	PSNR	SSIM	PSNR	SSIM
(Zhu et al., 2017)	17.434	0.641	5.56	0.0682
Prop. Approach	<b>31.502</b>	<b>0.950</b>	<b>11.538</b>	<b>0.4324</b>

## 5 CONCLUSIONS

This paper improves the domain transformation mechanism by generating synthetic images of the far

infrared (thermal) spectrum from visible spectrum images. In order to transfer not only shape, but also make the model simulate the temperature and texture of the thermal images, the unpaired cycled GAN network has been taken and modifications have been made in terms of the loss and normalization functions. As a further work we will explore with other state-of-the-art techniques based on transformers or diffusion models. The idea is to evaluate the generalization of the model to generate synthetic images with better quality.

## ACKNOWLEDGEMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0261; and partially supported by the ESPOL project CIDIS-12-2022; the Spanish Government under Project PID2021-128945NB-I00; and the "CERCA Programme / Generalitat de Catalunya". The authors gratefully acknowledge the NVIDIA Corporation for the donation of a Titan V GPU used for this research.

## REFERENCES

- Andonian, A., Park, T., Russell, B., Isola, P., Zhu, J.-Y., and Zhang, R. (2021). Contrastive feature loss for image prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1934–1943.
- Barron, J. T. and Poole, B. (2016). The fast bilateral solver. In *European conference on computer vision*, pages 617–632. Springer.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.
- Goodfellow Ian, J., Jean, P.-A., Mehdi, M., Bing, X., David, W.-F., Sherjil, O., and Courville Aaron, C. (2014). Generative adversarial nets. In *Proceedings of the 27th international conference on neural information processing systems*, volume 2, pages 2672–2680.
- Guo, T., Huynh, C. P., and Solh, M. (2019). Domain-adaptive pedestrian detection in thermal images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1660–1664. IEEE.
- Hui, T.-W., Loy, C. C., and Tang, X. (2016). Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision*, pages 353–369. Springer.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jolicœur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*.
- Kniaz, V. V., Knyaz, V. A., Hladuvka, J., Kropatsch, W. G., and Mizginov, V. (2018). Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.
- Kopf, J., Cohen, M. F., Lischinski, D., and Uyttendaele, M. (2007). Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es.
- Li, C., Xia, W., Yan, Y., Luo, B., and Tang, J. (2020). Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082.
- Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., and Luo, Z. (2022). Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811.
- Liu, R., Ge, Y., Choi, C. L., Wang, X., and Li, H. (2021). Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16377–16386.
- Lu, Y. and Lu, G. (2021). An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3833–3843.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.
- Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive learning for conditional image synthesis. In *ECCV*.
- Saleh, K., Abobakr, A., Attia, M., Iskander, J., Nahavandi, D., Hossny, M., and Nahavandi, S. (2019). Domain adaptation for vehicle detection from bird’s eye view lidar point cloud data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.
- Suárez, P. L., Sappa, A. D., and Vintimilla, B. X. (2019). Image patch similarity through a meta-learning metric based approach. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 511–517. IEEE.
- Xie, J., Feris, R. S., and Sun, M.-T. (2015). Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438.
- Yu, N., Liu, G., Dundar, A., Tao, A., Catanzaro, B., Davis, L. S., and Fritz, M. (2021). Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6731–6742.
- Zhang, L., Gonzalez-Garcia, A., Van De Weijer, J., Danelljan, M., and Khan, F. S. (2018). Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850.
- Zhou, D., Wang, R., Lu, J., and Zhang, Q. (2018). Depth image super resolution based on edge-guided method. *Applied Sciences*, 8(2):298.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.