ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096281

THERMAL INFRARED IMAGE INPAINTING VIA EDGE-AWARE GUIDANCE

Zeyu Wang[a], *Haibin Shen*[a], *Changyou Men*[b], *Quan Sun*[b], *Kejie Huang*[a,*]

[a]Zhejiang University, Hangzhou, China
[b]Hangzhou Vango Technologies, Inc., Hangzhou, China

ABSTRACT

Image inpainting has achieved fundamental advances with deep learning. However, almost all existing inpainting methods aim to process natural images, while few target Thermal Infrared (TIR) images, which have widespread applications. When applied to TIR images, conventional inpainting methods usually generate distorted or blurry content. In this paper, we propose a novel task—*Thermal Infrared Image Inpainting*, which aims to reconstruct missing regions of TIR images. Crucially, we propose a novel deep-learning-based model *TIR-Fill*. We adopt the edge generator to complete the canny edges of broken TIR images. The completed edges are projected to the normalization weights and biases to enhance edge awareness of the model. In addition, a refinement network based on gated convolution is employed to improve TIR image consistency. The experiments demonstrate that our method outperforms state-of-the-art image inpainting approaches on *FLIR* thermal dataset.

Index Terms— Thermal Infrared Image Inpainting, Edge Awareness, Gated Convolution, State-Of-The-Art.

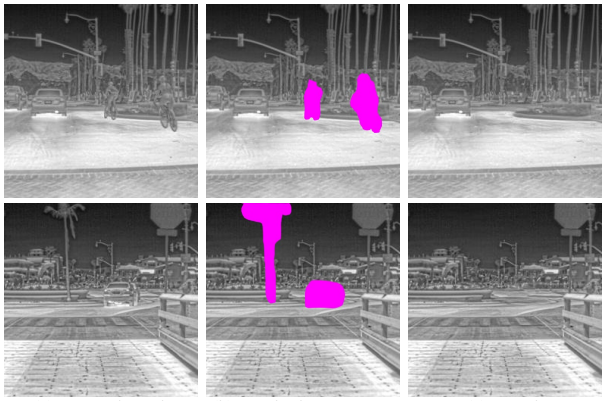
1. INTRODUCTION

Image inpainting refers to filling missing regions of broken images, which has excellent usage in image processing. Driven by the advances of Deep Learning (DL), this technique has progressed significantly in the past few years [1, 2, 3, 4, 5, 6, 7]. However, most inpainting methods aim to reconstruct natural images in the visible spectrum.

Recently, Thermal Infrared (TIR) technology has been increasingly vital in remote sensing [8], medicine [9], and so on. Unlike visible-spectrum cameras, TIR cameras can capture infrared radiation at the wavelength of $2-1000\mu m$, enabling them to detect low-light objects. In this paper, we propose a novel task—*Thermal Infrared Image Inpainting*, which aims to fill the missing regions of broken TIR images. This technique can contribute to image editing, artifact repair, and object removal for TIR images, as shown in Fig. 1.

This research has been supported by the China National Key R&D Program (Grant No. 2022YFB4400704) and Hangzhou Major Technology Innovation Project of Artificial Intelligence (Grant No. 2022AIZD0060).

*Corresponding author: huangkejie@zju.edu.cn



(a) Ground Truth (b) Masked TIR Image (c) Result

Fig. 1. Examples of *Thermal Infrared Image Inpainting*. The results are generated by our *TIR-Fill*.

However, conventional DL-based inpainting methods are not applicable to TIR image inpainting, as they aim to reconstruct natural images but usually create distorted or blurry structures for TIR images. Compared with natural images, TIR images have lower chromatic contrast but contain richer thermal information and sharper edge contours. Therefore, we propose a novel DL-based model *TIR-Fill*. We first extract the canny edges of broken TIR images and build an edge generator to reconstruct them. The completed edges are inserted into our Edge-Aware Guidance (EAG) normalization in *TIR-Fill*, which enhances the model awareness of edge information. A refinement network based on gated convolution is employed to improve the generated TIR image quality and consistency. We demonstrate that our method quantitatively outperforms state-of-the-art image inpainting approaches and generates visually appealing results on *FLIR* thermal dataset [10], as shown in Fig. 1(c).

2. RELATED WORKS

2.1. Image Inpainting

Traditional image inpainting methods include diffusion-based methods [11] and patch-based methods [12]. In recent years, Deep Learning (DL) has exhibited superior performance in

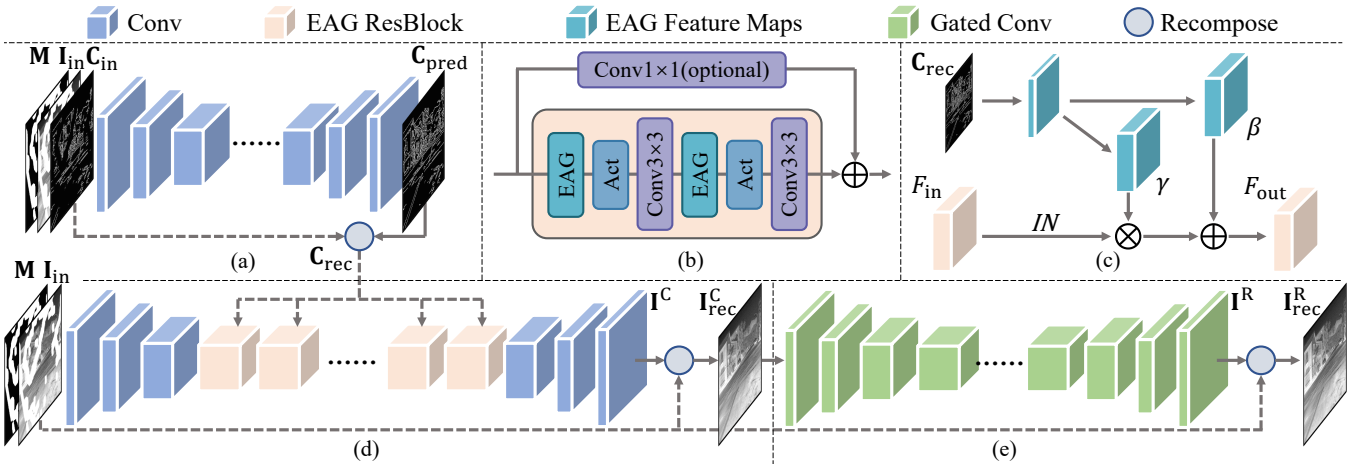


Fig. 2. Illustration of our proposed *TIR-Fill*. (a) The CNN-based edge generator E_θ [4]. (b) Our EAG ResBlock. (c) Our EAG normalization layer. (d) Our TIR image completion network G_ϕ, which integrates EAG ResBlocks. (e) Our TIR image refinement network R_ψ based on gated convolution [3].

image inpainting. Conventional DL-based inpainting methods are based on Convolutional Neural Network (CNN) [1, 2, 3, 4, 6] and Transformer [7, 5]. The majority of these works are proposed for natural image inpainting.

2.2. Thermal Infrared Image Processing

TIR image processing has attracted extensive research for its vital and widespread applications. Previous works targeted at TIR image super-resolution [13] and semantic segmentation [14]. Currently, some works aim to tackle TIR image colorization, which maps a single-channel grayscale TIR image to an RGB image [15]. To our knowledge, no DL-based work has been proposed for TIR image inpainting.

3. METHODS

3.1. Overview

To perform the task of *Thermal Infrared Image Inpainting*, we propose the novel DL-based model *TIR-Fill*, which is illustrated in Fig. 2. It comprises three stages: Edge connection, TIR image completion, and TIR image refinement. Given the mask $\mathbf{M}$, the goal of the task is to reconstruct the ground-truth TIR image $\mathbf{I}_{gt}$ based on the input $\mathbf{I}_{in} = \mathbf{I}_{gt} \odot \mathbf{M}$. The mask $\mathbf{M}$ is a binary matrix, where 0 and 1 denote the hole and valid pixels, respectively.

3.2. Edge Connection

We first extract the canny edge $\mathbf{C}_{in}$ of the broken image $\mathbf{I}_{in}$. The CNN-based edge generator E_θ [4] is adopted for the edge connection, as illustrated in Fig. 2(a). The goal of this stage is to reconstruct the ground-truth edge $\mathbf{C}_{gt}$ based on $\mathbf{M}$, $\mathbf{I}_{in}$, and $\mathbf{C}_{in}$, as formulated below:

$$\mathbf{C}_{pred} = \varepsilon(E_\theta(\mathbf{M}, \mathbf{I}_{in}, \mathbf{C}_{in}) - t_0) \tag{1}$$

where $t_0 = 0.5$ and $\varepsilon(\cdot)$ denotes the unit step function: if $t < t_0$, $\varepsilon(t - t_0) = 0$, else $\varepsilon(t - t_0) = 1$.

The predicted edge $\mathbf{C}_{pred}$ is recomposed with $\mathbf{C}_{in}$ into $\mathbf{C}_{rec} = \mathbf{C}_{in} + \mathbf{C}_{pred} \odot (\mathbf{1} - \mathbf{M})$, which will be adopted for enhancing the edge awareness.

3.3. TIR Image Completion

The TIR image completion network G_ϕ takes $\mathbf{M}$, $\mathbf{I}_{in}$, and $\mathbf{C}_{rec}$ as inputs to generate $\mathbf{I}^{C}$, as illustrated in Fig. 2(d).

$$\mathbf{I}^{C} = G_\phi(\mathbf{M}, \mathbf{I}_{in}, \mathbf{C}_{rec}) \tag{2}$$

In detail, the intermediate layers of G_ϕ are our Edge-Aware Guidance (EAG) ResBlocks, as illustrated in Fig. 2(b). Compared with conventional ResNet Block, EAG ResBlock replaces conventional normalization with our EAG normalization, which inserts the recomposed edge $\mathbf{C}_{rec}$, as shown in Fig. 2(c). Inside the EAG normalization layer, $\mathbf{C}_{rec}$ is projected through convolutional layers into the modulation parameters γ and β. Then, γ and β are adopted as the normalization weight and bias, respectively, which is formulated as:

$$F_{out} = IN(F_{in}) \odot \gamma_{x,y,c}(\mathbf{C}_{rec}) + \beta_{x,y,c}(\mathbf{C}_{rec}) \tag{3}$$

where *IN* denotes the instance normalization, $\gamma = \gamma_{x,y,c}(\mathbf{C}_{rec})$, and $\beta = \beta_{x,y,c}(\mathbf{C}_{rec})$. It is inspired from SPADE [16] normalization for semantic synthesis. The predicted coarse result is recomposed into $\mathbf{I}^{C}_{rec} = \mathbf{I}_{in} + \mathbf{I}^{C} \odot (\mathbf{1} - \mathbf{M})$.

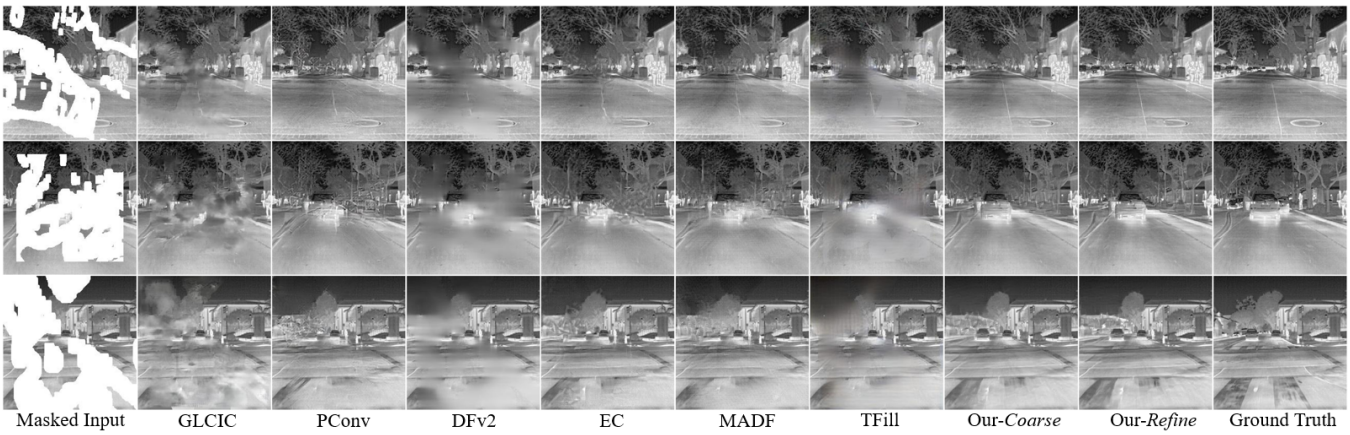


Fig. 3. Qualitative comparison between our *TIR-Fill* and the baseline inpainting methods on *FLIR* dataset.

3.4. TIR Image Refinement

Based on coarse recomposed result $\mathbf{I}^{\mathrm{C}}_{\mathrm{rec}}$, a refinement network R_ψ is employed to further improve the TIR image quality, as shown in Fig. 2(e).

$$\mathbf{I}^{\mathrm{R}} = R_\psi(\mathbf{M}, \mathbf{I}^{\mathrm{C}}_{\mathrm{rec}}) \tag{4}$$

The refinement network R_ψ consists of stacked gated convolutional layers [3], which can learn dynamic feature gating for each channel and each spatial location:

$$F_{\mathrm{out}} = \sigma(W_g * F_{\mathrm{in}}) \odot \phi(W_f * F_{\mathrm{in}}) \tag{5}$$

where W_g and W_f denote two different convolutional filters, σ denotes Sigmoid activation, and ϕ denotes Swish activation. The predicted $\mathbf{I}^{\mathrm{R}}$ is further recomposed with the input into the final result $\mathbf{I}^{\mathrm{R}}_{\mathrm{rec}} = \mathbf{I}_{\mathrm{in}} + \mathbf{I}^{\mathrm{R}} \odot (\mathbf{1} - \mathbf{M})$.

3.5. Loss Functions

To train E_θ, we adopt adversarial training with Patch-GAN discriminator D_{patch} [17] and hinge loss.

$$\mathcal{L}_{E_\theta} = \mathcal{L}_{\mathrm{adv}}(\mathbf{C}_{\mathrm{pred}}) = -\mathbb{E}[D_{\mathrm{patch}}(\mathbf{C}_{\mathrm{pred}})] \tag{6}$$

$$\begin{aligned}\mathcal{L}_{D_{\mathrm{patch}}} =& \mathbb{E}[relu(1 - D_{\mathrm{patch}}(\mathbf{C}_{\mathrm{gt}}))] \\ &+ \mathbb{E}[relu(1 + D_{\mathrm{patch}}(\mathbf{C}_{\mathrm{pred}}))]\end{aligned} \tag{7}$$

To train G_ϕ and R_ψ, we define a reconstruction loss consisting of ℓ_1 loss, perceptual loss [18], and style loss [19].

$$\begin{aligned}\mathcal{L}_{\mathrm{rec}}(\mathbf{I}, \mathbf{I}_{\mathrm{gt}}) = \ell_1(\mathbf{I}, \mathbf{I}_{\mathrm{gt}}) &+ \sum\nolimits_i \|\mathcal{F}_i(\mathbf{I}) - \mathcal{F}_i(\mathbf{I}_{\mathrm{gt}})\|_1 \\ &+ \sum\nolimits_j \|\mathcal{G}_j(\mathbf{I}) - \mathcal{G}_j(\mathbf{I}_{\mathrm{gt}})\|_1\end{aligned} \tag{8}$$

where $\mathcal{F}_i (i \in \{2, 7, 12, 21, 30\})$ denotes the intermediate feature map of the i-th layer in the VGG-19 [20], and $\mathcal{G}_j(\cdot) = \mathcal{F}_j(\cdot)\mathcal{F}_j(\cdot)^T (j \in \{9, 18, 27, 32\})$ denotes the Gram matrix [19]. Based on $\mathcal{L}_{\mathrm{rec}}$, the loss functions of G_ϕ and R_ψ are formulated as:

$$\mathcal{L}_{G_\phi} = \mathcal{L}_{\mathrm{rec}}(\mathbf{I}^{\mathrm{C}}, \mathbf{I}_{\mathrm{gt}}) \tag{9}$$

$$\mathcal{L}_{R_\psi} = \mathcal{L}_{\mathrm{rec}}(\mathbf{I}^{\mathrm{R}}, \mathbf{I}_{\mathrm{gt}}) + \mathcal{L}_{\mathrm{adv}}(\mathbf{I}^{\mathrm{R}}) \tag{10}$$

where another Patch-GAN discriminator is implemented, along with the adversarial loss $\mathcal{L}_{\mathrm{adv}}(\mathbf{I}^{\mathrm{R}}) = -\mathbb{E}[D_{\mathrm{patch}}(\mathbf{I}^{\mathrm{R}})]$ and the same discriminator loss as Eq. 7.

4. EXPERIMENTS

4.1. Dataset

We adopt *FLIR* thermal dataset [10], which consists of 8862 training images and 1366 testing images. For training images, we randomly crop and resize them to 256×256. For testing images, we resize them to 300×375 and crop them from the center to 256×256 for evaluation. The irregular masks with arbitrary mask ratios provided by Liu *et al.* [2] are adopted.

4.2. Experimental Settings

We implement *TIR-Fill* with PyTorch 1.8.1 on one NVIDIA RTX A6000 GPU with 40G memory. The low and high thresholds of canny edge detection are set to 80 and 160, respectively. The learning rates for training E_θ, G_ϕ, and R_ψ are set to $1e^{-3}$, $1e^{-4}$, and $1e^{-4}$, respectively. The Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ is employed.

4.3. Quantitative Comparison

We report PSNR, SSIM, LPIPS, and FID for quantitative comparison. The results of our *TIR-Fill* and the baseline inpainting methods are shown in Table 1. Among the baseline methods, DFv2 (DeepFillv2) performs well in PSNR, and even outperforms our coarse result under large mask ratios. TFill is the SOTA method for natural image inpainting, but it performs poorly for TIR image inpainting. By contrast,

Table 1. Quantitative comparison between our *TIR-Fill* and the baseline inpainting methods on *FLIR* dataset.

	Mask Ratio	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%
PSNR↑	GLCIC [1]	36.61	31.02	27.53	25.11	23.04	20.28
	PConv [2]	37.87	31.79	28.56	26.50	24.90	22.85
	DFv2 [3]	37.82	32.24	29.20	27.23	25.72	23.69
	EC [4]	38.21	32.25	29.01	26.92	25.31	23.21
	MADF [6]	37.99	31.87	28.78	26.84	25.11	23.06
	TFill [7]	36.49	30.51	28.20	26.43	25.07	23.31
	Our-*Coarse*	39.36	32.96	29.45	27.13	25.38	23.17
	Our-*Refine*	39.65	33.48	30.06	27.79	26.06	23.82
SSIM↑	GLCIC [1]	0.975	0.932	0.872	0.808	0.740	0.653
	PConv [2]	0.977	0.935	0.881	0.824	0.766	0.682
	DFv2 [3]	0.977	0.939	0.890	0.841	0.791	0.727
	EC [4]	0.979	0.941	0.891	0.837	0.781	0.699
	MADF [6]	0.976	0.938	0.885	0.833	0.772	0.693
	TFill [7]	0.974	0.931	0.880	0.829	0.778	0.715
	Our-*Coarse*	0.983	0.950	0.904	0.853	0.799	0.716
	Our-*Refine*	0.983	0.952	0.908	0.859	0.807	0.728
LPIPS↓	GLCIC [1]	0.040	0.099	0.172	0.238	0.302	0.364
	PConv [2]	0.029	0.074	0.127	0.177	0.228	0.290
	DFv2 [3]	0.041	0.102	0.172	0.236	0.297	0.362
	EC [4]	0.028	0.072	0.123	0.173	0.224	0.289
	MADF [6]	0.036	0.089	0.144	0.196	0.267	0.342
	TFill [7]	0.038	0.094	0.154	0.208	0.263	0.322
	Our-*Coarse*	0.021	0.055	0.097	0.139	0.184	0.246
	Our-*Refine*	0.021	0.054	0.095	0.137	0.181	0.242
FID↓	GLCIC [1]	8.41	26.27	54.78	83.82	114.85	145.63
	PConv [2]	5.19	13.96	26.49	41.71	60.36	85.33
	DFv2 [3]	8.23	26.15	54.24	85.10	115.96	141.67
	EC [4]	4.78	12.39	23.10	35.83	52.02	80.48
	MADF [6]	6.31	18.52	35.62	46.82	64.87	90.13
	TFill [7]	9.41	25.97	50.13	77.49	108.41	136.61
	Our-*Coarse*	3.28	8.12	14.29	20.33	27.36	39.84
	Our-*Refine*	3.16	7.58	12.86	17.76	23.13	31.86

↑: Higher is better. ↓: Lower is better. The values marked in red and blue denote the best and the second best results, respectively.

our *TIR-Fill* achieves the best metrics under all mask ratios. Obviously, our advantage in FID score is the most significant.

4.4. Qualitative Comparison

The qualitative results are shown in Fig. 3. We can see that these baseline methods generate poor inpainting results for TIR images. DFv2 and TFill generate blurry content, while GLCIC and PConv create distorted structures. EC (Edge Connector), which also utilizes edge information, can generate somewhat good results but with a few damaged details. By contrast, our *TIR-Fill* generates delicate structures, including pedestrians, trees, and cars. The comparison illustrates that our model can generate visually appealing inpainting results for TIR images. Therefore, the previous baseline methods are not applicable to TIR image inpainting, which is a worth-discussing task requiring the specifically designed method.

Table 2. Quantitative results of the ablation study. The results are shown as the average scores under all mask ratios.

Model	Stage	PSNR↑	SSIM↑	LPIPS↓	FID↓
w/o EAG	*Coarse*	29.24	0.844	0.135	25.61
w/o EAG	*Refine*	29.73	0.849	0.133	21.42
TIR-Fill	*Coarse*	29.57	0.867	0.124	18.87
TIR-Fill	*Refine*	30.14	0.873	0.122	16.06

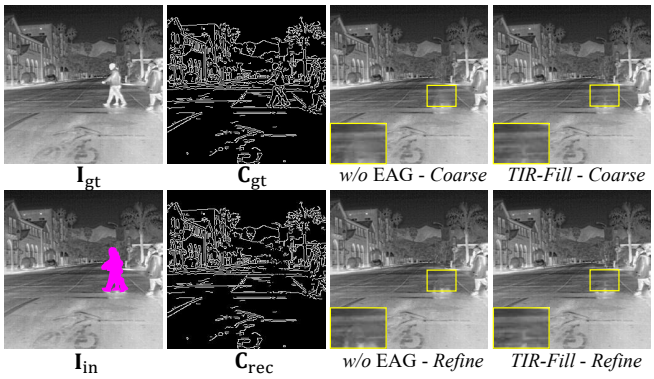


Fig. 4. Qualitative results of the ablation study and visualization of the reconstructed canny edges.

4.5. Ablation Study

We conduct the ablation study to illustrate the effectiveness of our EAG normalization. The variant "*w/o* EAG" denotes another *TIR-Fill* which replaces the EAG normalization with conventional instance normalization. The results shown in Table 2 demonstrate that EAG normalization dramatically improves the quantitative performance, especially the FID score. In Fig. 4, we aim to remove the pedestrians on the road. The visualization of the reconstructed edge $\mathbf{C}_{rec}$ shows that the edges of the pedestrians are naturally removed and completed. The qualitative comparison between the variant and the complete *TIR-Fill* demonstrates that *TIR-Fill* can generate more delicate structures.

5. CONCLUSION

In this work, we propose a novel image processing task—*Thermal Infrared Image Inpainting*, which aims to reconstruct missing regions of TIR images. In addition, we propose an effective model *TIR-Fill* to deal with the novel task, which integrates our EAG normalization for enhancing edge awareness. The experiments demonstrate that our *TIR-Fill* outperforms the baseline inpainting methods. Its visually appealing inpainted results demonstrate its ability in TIR image editing. The ablation study illustrates that EAG normalization performs better than conventional normalization. In the future, we expect the worth-discussing task will attract extensive research and contribute to widespread applications.

6. REFERENCES

[1] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.

[2] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[3] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[4] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[5] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao, "High-fidelity pluralistic image completion with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4692–4701.

[6] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang, "Image inpainting by end-to-end cascaded refinement with mask awareness," *IEEE Transactions on Image Processing*, vol. 30, pp. 4855–4866, 2021.

[7] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung, "Bridging global context interactions for high-fidelity image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11512–11522.

[8] John M Norman and Francois Becker, "Terminology in thermal infrared remote sensing of natural surfaces," *Agricultural and Forest Meteorology*, vol. 77, no. 3-4, pp. 153–166, 1995.

[9] EFJ Ring and Kurt Ammer, "Infrared thermal imaging in medicine," *Physiological measurement*, vol. 33, no. 3, pp. R33, 2012.

[10] F.A.Group, "Flir thermal dataset for algorithm training," *https://www.flir.co.uk/oem/adas/adas-dataset-form/*, 2019.

[11] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 417–424.

[12] Maxime Daisy, David Tschumperlé, and Olivier Lézoray, "A fast spatial patch blending algorithm for artefact reduction in pattern-based image inpainting," in *SIGGRAPH Asia 2013 Technical Briefs*, pp. 1–4. 2013.

[13] Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Anjali Sarvaiya, Kishor P Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch, "Channel split convolutional neural network (chasnet) for thermal image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4368–4377.

[14] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[15] Fuya Luo, Yijun Cao, and Yongjie Li, "Nighttime thermal infrared image colorization with dynamic label mining," in *International Conference on Image and Graphics*. Springer, 2021, pp. 388–399.

[16] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[19] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.