A Preliminary Study on Object-Aware RGB-to-Thermal Image Translation for Cooking Scenes

Ziqi Li[1,2,a)] Motoharu Sonogashira[2,1] Kazuya Kitano[1] Yuki Fujimura[1] Yasuhiro Mukaigawa[1] Yasutomo Kawanishi[3,2,1,b)]

Abstract

Thermal image generation in cooking scenes is challenging because food temperatures continuously change during the heating process, while visual appearance changes are often limited. In this study, we propose an RGB-to-thermal image translation framework for cooking scenarios that incorporates temporal and cooking-state information. The proposed method utilizes an initial RGB image, a current RGB image, a normalized heating-progress condition, and a flipping-state condition to model temperature evolution during cooking. A conditional diffusion model is employed to generate thermal images from these inputs. Experiments are conducted on a paired RGB-thermal cooking dataset consisting of beef-heating sequences. Quantitative and qualitative evaluation demonstrate that the proposed framework can generate realistic thermal distributions and capture temperature trends during cooking. Furthermore, the results show that incorporating temporal and flipping-state information improves thermal image generation and enables more accurate modeling of temperature dynamics throughout the cooking process.

1. Introduction

Thermal imaging provides valuable temperature information for a wide range of applications, including industrial inspection, autonomous driving, and human activity analysis [1], [9]. Thermal imaging has also been widely used in food-related applications, including food quality assessment, temperature monitoring, and meat analysis [10]. In cooking environments, thermal images can be used to monitor food temperatures and estimate cooking states, which are important for cooking assistance systems and food quality assessment. However, thermal cameras are generally expensive and difficult to deploy in everyday environments, motivating the development of methods that estimate thermal information from conventional RGB images.

[1] NARA Institute of Science and Technology
[2] RIKEN
[3] Ritsumeikan University
[a)] ziqi.li@a.riken.jp
[b)] kwns@fc.ritsumei.ac.jp

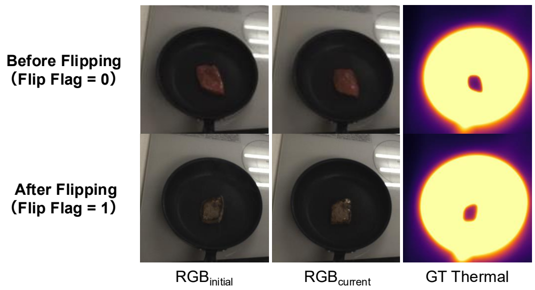


Fig. 1: Examples of RGB-to-Thermal image translation in cooking scenes. Similar RGB appearances may correspond to different thermal distributions. GT denotes the ground-truth thermal image. The proposed method utilizes temporal and cooking-state information to generate thermal images.

Recent advances in image-to-image translation have enabled the generation of thermal images from RGB images using deep learning techniques [3], [4]. Existing RGB-to-thermal image translation methods have demonstrated promising performance in outdoor and autonomous driving scenarios [1], where temperature distributions are relatively stable and strongly correlated with object categories. These approaches have shown that visual appearance can provide useful cues for estimating thermal information.

As illustrated in Fig. 1, similar RGB appearances may correspond to substantially different thermal distributions during the cooking process. Therefore, thermal image generation in cooking scenes present unique challenges with conventional RGB-to-thermal translation tasks. Unlike outdoor environments, the temperature of food continuously changes during the heating process. Furthermore, temperature distributions are influenced not only by object appearance but also by cooking duration and heating conditions. As a result, the relationship between RGB appearance and temperature is considerably more complex than that in conventional RGB-to-thermal translation tasks.

To generate accurate thermal images in cooking environments, additional information beyond visual appearance should be considered. Food at different heating stages may exhibit very similar RGB appearances while corresponding to substantially different temperature distributions. Therefore, relying solely on the current RGB observation may be insufficient for estimating thermal distributions. Comparing the current RGB image with the initial RGB image provides additional cues about the cooking progress and temperature evolution. Moreover, cooking operations such as food flipping can introduce abrupt changes in thermal evolution that are difficult to infer solely from RGB observations. Therefore, both temporal information and cooking-state information are important for accurate thermal image generation.

In this study, we propose a cooking-oriented RGB-to-thermal image translation framework that generates a thermal image corresponding to the current cooking stage. The proposed method utilizes the current RGB image as the primary visual observation, together with an initial RGB image, a normalized heating-progress condition, and a flipping-state condition as additional inputs. The initial RGB image provides a reference before heating, allowing the model to capture appearance changes and cooking progress that are difficult to infer from the current RGB image alone. Experiments conducted on a paired RGB-thermal cooking dataset demonstrate that incorporating these cooking-related cues improves thermal image generation and enables more accurate modeling of temperature changes throughout the cooking process.

2. Proposed Method

In this study, we propose a cooking-oriented RGB-to-thermal image translation framework that generates thermal images from RGB observations and cooking-related conditions. The framework incorporates visual information, temporal information, and cooking-state information to model temperature evolution during the cooking process. An overview of the proposed framework is shown in Fig. 2.

2.1 Cooking-Oriented RGB-to-Thermal Framework

Generating thermal images in cooking environments is challenging due to continuous temperature changes during heating and food flipping. The proposed framework utilizes visual observations and cooking-related conditions to estimate thermal distributions.

Let R_0 denote the RGB image captured before heating, R_k denote the RGB image captured at frame index k, γ_k denote the normalized heating-progress condition, and f_k denote the flipping-state condition. The current RGB image R_k provides the primary visual observation corresponding to the target thermal image, while the initial RGB image R_0 serves as a reference before heating. Comparing R_0 and R_k provides cues about cooking progress and temperature evolution that are difficult to infer from a single RGB observation. The RGB-to-thermal image translation task is formulated as

$$Y_k = F(R_0, R_k, \gamma_k, f_k), \tag{1}$$

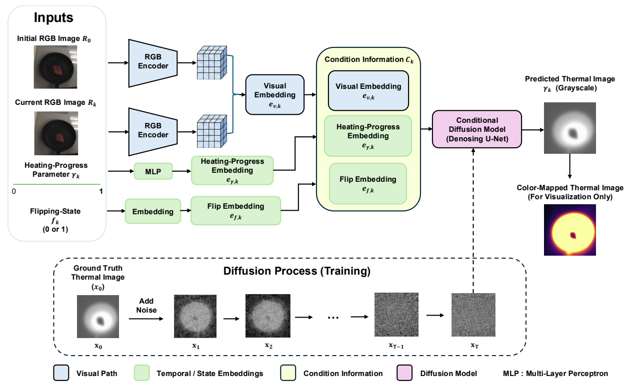


Fig. 2: Overview of the proposed framework. Visual, temporal, and cooking-state information are integrated to generate thermal images using a conditional diffusion model.

where Y_k denotes the thermal image corresponding to the current cooking stage and $F(\cdot)$ represents the translation model.

As shown in Fig. 2, the initial RGB image R_0 and the current RGB image R_k are first processed by an RGB encoder to extract visual representations. The extracted features are aggregated as a visual embedding $e_{v,k}$, which captures appearance information of the food before and during heating.

In addition to visual observations, the framework incorporates two cooking-related conditions. The heating-progress condition γ_k is transformed into a heating-progress embedding $e_{\gamma,k}$ to represent temporal progression during cooking. Furthermore, the flipping-state condition f_k is converted into a flip embedding $e_{f,k}$ to distinguish thermal characteristics before and after food flipping. These embeddings are combined with the visual embedding to construct condition information for thermal image generation.

We adapt a conditional diffusion model due to its strong capability in image generation and image-to-image translation tasks [5], [8]. The model receives both the noisy thermal image and the condition information. During training, Gaussian noise is progressively added to the ground-truth thermal image to construct the diffusion process. The model then learns to recover the corresponding thermal distribution through an iterative denoising process. By jointly utilizing visual, temporal, and cooking-state information, the proposed framework learns the relationship between food appearance and temperature distributions throughout the cooking process.

2.2 Temporal and Cooking-State Conditioning

Thermal distributions in cooking scenes are strongly influenced by both heating duration and cooking operations. Food at different cooking stages may exhibit similar visual appearances despite having substantially different temperature distributions. Furthermore, food flipping introduces abrupt changes in heat transfer and temperature evolution that are difficult to infer solely from RGB observations.

To provide the diffusion model with temporal information, a normalized heating-progress condition γ_k is introduced.

The condition is defined as

$$\gamma_k = \frac{k}{K}, \tag{2}$$

where k denotes the current frame index and K denotes the total number of frames in the cooking sequence. Consequently, γ_k takes values in the range $[0, 1]$, where $\gamma_k = 0$ corresponds to the beginning of the heating process and $\gamma_k = 1$ corresponds to the final stage.

Since γ_k is represented by a scalar value, it is transformed through a multilayer perceptron (MLP) to obtain a heating-progress embedding. The resulting heating-progress embedding enables effective conditioning on cooking progress into the conditional diffusion model.

In addition to temporal information, a binary flipping-state condition f_k is introduced to represent whether the food has been flipped during cooking. Specifically,

$$f_k = \begin{cases} 0, & \text{before flipping,} \\ 1, & \text{after flipping.} \end{cases} \tag{3}$$

The flipping-state condition is converted into a learnable embedding $e_{f,k}$ through an embedding layer. This embedding provides explicit information about the cooking state and helps the model distinguish thermal characteristics before and after flipping.

The heating-progress embedding $e_{\gamma,k}$ and flip embedding $e_{f,k}$ are combined with the visual embedding $e_{v,k}$ extracted from RGB observations to construct the condition information

$$c_k = [e_{v,k}, e_{\gamma,k}, e_{f,k}], \tag{4}$$

where c_k denotes the condition information used by the conditional diffusion model. By integrating visual, temporal, and cooking-state information into a unified representation, the proposed framework can effectively model temperature evolution throughout the cooking process.

2.3 Diffusion-Based Thermal Image Generation

The proposed framework employs a conditional diffusion model to generate thermal images from RGB observations and cooking-related conditions. During training, Gaussian noise is progressively added to the ground-truth thermal image, while the model learns to recover the original thermal distribution through an iterative denoising process [5], [6], [7], [8].

Given a ground-truth thermal image x_0, the forward diffusion process generates a noisy image x_t by gradually adding Gaussian noise according to

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise and $\bar{\alpha}_t$ is the cumulative noise schedule.

The denoising process is performed by a conditional U-Net, which receives the noisy thermal image x_t together with the condition information c_k defined in Section 2.2.

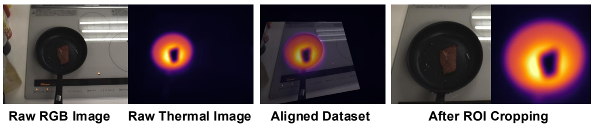


Fig. 3: Preprocessing procedure of the RGB-thermal dataset. The RGB and thermal images are first geometrically aligned to establish pixel-wise correspondence and then cropped to a fixed region of interest for training and evaluation.

The network predicts the noise component contained in x_t as

$$\hat{\epsilon} = \epsilon_\theta(x_t, t, c_k), \tag{6}$$

where $\epsilon_\theta(\cdot)$ denotes the noise prediction network parameterized by θ.

Using the predicted noise, the model progressively removes noise from the thermal image and reconstructs the corresponding thermal distribution. By conditioning the denoising process on c_k, the model can exploit visual appearance, heating progression, and flipping-state information throughout the generation process.

The model is trained by minimizing the noise prediction loss

$$L_{\text{diff}} = \mathbb{E}_{x_0,\epsilon,t} \left[\| \epsilon - \epsilon_\theta(x_t, t, c_k) \|_2^2 \right]. \tag{7}$$

where $\mathbb{E}_{x_0,\epsilon,t}$ denotes the expectation over training samples x_0, Gaussian noise ϵ, and diffusion step t. This process encourages the model to learn the mapping from RGB observations and thermal distributions while incorporating temporal and cooking-state information. During inference, thermal images are generated by iteratively removing noise from an initial random Gaussian sample under the guidance of the condition information.

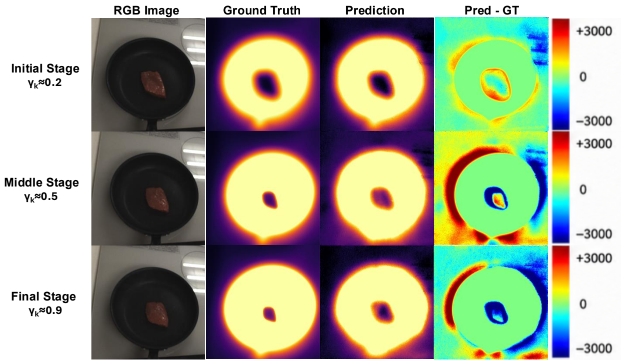
3. Experiments

3.1 Dataset and Experimental Settings

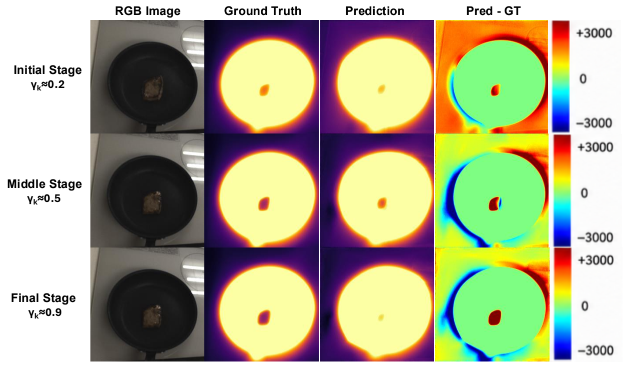
We constructed a paired RGB-thermal cooking dataset consisting of synchronized beef-heating sequences. We captured synchronized RGB and thermal image sequences during the heating process of beef, where RGB images record food appearance and thermal images provide the corresponding temperature distributions.

As shown in Fig. 3, RGB and thermal images were geometrically aligned using a homography transformation estimated from manually selected corresponding points between the two modalities. The estimated transformation was applied to all image pairs to establish pixel-wise correspondence between RGB and thermal images. The aligned image pairs were then cropped to a fixed region of interest (ROI) and resized to a resolution of 260×260 pixels for training and evaluation. Finally, the dataset was divided into training and validation sets using different cooking sequences.

For each frame index k, the model receives an initial RGB image R_0, a current RGB image R_k, a normalized heating-progress condition γ_k, and a flipping-state condition f_k as



(a) Before flipping ($f_k = 0$)



(b) After flipping ($f_k = 1$)

Fig. 4: Qualitative thermal image generation results before and after flipping.

Table 1: Quantitative comparison of thermal image generation performance. Heat denotes the heating-progress condition and Flip denotes the flipping-state condition.

Method	MAE↓	PSNR↑	SSIM↑
Baseline RGB-Thermal	2252.71	11.43	0.77
+ Heat	1617.68	16.65	0.86
Proposed (Heat + Flip)	**1025.90**	**18.56**	**0.91**

inputs. The corresponding thermal image is used as the supervision signal during training.

The proposed framework was implemented using a conditional diffusion model based on UNet2DConditionModel [2]. Performance was evaluated using mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). The thermal images are represented using 16-bit raw thermal values rather than calibrated temperature values in degrees Celsius. Therefore, MAE is reported in the raw thermal value domain.

3.2 Experimental Results

Table 1 summarizes the quantitative evaluation results of different conditioning strategies. Compared with the baseline model, introducing the heating-progress condition consistently improves all evaluation metircs, demonstrating the importance of temporal information for modeling temperature evolution during cooking.

Incorporating the flipping-state condition achieves the best overall performance, highlighting the importance of cooking-state awareness for thermal image generation.

Fig. 4 presents qualitative thermal image generation results at different cooking stages before and after food flipping. The generated thermal images successfully reproduce the overall thermal distributions observed in the ground-truth images. In particular, the proposed framework preserves the spatial structure of the food while capturing temperature variations throughout the heating process.

Most prediction errors are concentrated around food boundaries and surrounding high-temperature regions, while the generated thermal images remain largely consistent with the ground-truth thermal distributions.

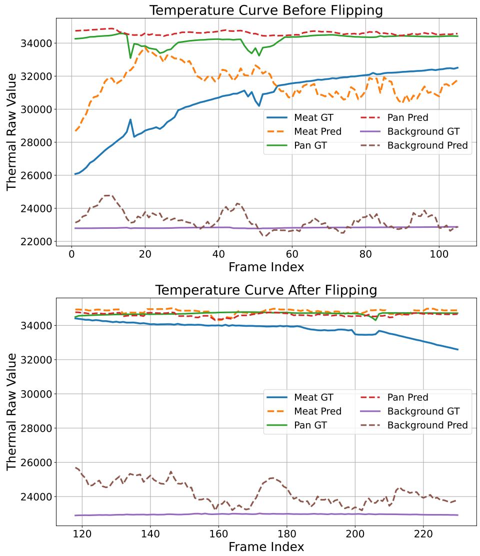


Fig. 5: Temperature evolution analysis before and after food flipping.

3.3 Temporal Analysis of Thermal Generation

Fig. 5 shows the temperature evolution of different regions during cooking. The average thermal values of the meat, pan, and background regions were computed from both the generated and ground-truth thermal images.

The proposed framework successfully captures the temperature trends before and after food flipping. The predicted temperatures of the pan and background regions remain close to the ground-truth values throughout the cooking process. For the meat region, the generated thermal images generally follow the temperature evolution observed in the ground-truth data, although a tendency toward temperature underestimation can be observed during later cooking stages.

4. Conclusion

In this paper, we proposed a cooking-oriented RGB-to-thermal image translation framework using visual, temporal, and flipping-state conditions. Experimental results demonstrated that the proposed conditioning strategy improves thermal image generation. Future work will extend the framework to more diverse cooking scenarios and food categories.

References

[1] V. Kniaz, V. Knyaz, F. Remondino and W. G. Kropatsch, "ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset," ECCV Workshops, 2018.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, pp. 10684–10695, 2022.

[3] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," CVPR, 2017.

[4] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," ICCV, 2017.

[5] J. Ho, A. Jain and P. Abbeel, "Denoising Diffusion Probabilistic Models," NeurIPS, 2020.

[6] J. Song, C. Meng and S. Ermon, "Denoising Diffusion Implicit Models," ICLR, 2021.

[7] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," NeurIPS, 2021.

[8] C. Saharia, W. Chan, S. Saxena et al., "Palette: Image-to-Image Diffusion Models," SIGGRAPH, 2022.

[9] S. Hwang, J. Park, N. Kim et al., "Multispectral Pedestrian Detection: Benchmark Dataset and Baseline," CVPR, 2015.

[10] Y. Jiang, D. Wang, Z. Wang et al., "Recent Advances in Spectroscopy and Imaging Techniques for Meat Quality and Safety Detection: Principles, Applications and Future Perspectives," Food Science & Nutrition, vol. 13, no. 1, 2025.