Supplementary Material - ThermVision: Exploring FLUX for Synthesizing Hyper-Realistic Thermal Face Data and Animations via Image to Video Translation

Muhammad Ali Farooq*
muhammadali.farooq@universityofgalway.ie
School of Engineering
University of Galway
Galway, Ireland

Waseem Shariff
School of Engineering
University of Galway
Galway, Ireland
waseem.shariff@universityofgalway.ie

Peter Corcoran
peter.corcoran@universityofgalway.ie
School of Engineering
University of Galway
Galway, Ireland

ACM Reference Format:
Muhammad Ali Farooq, Waseem Shariff, and Peter Corcoran. 2025. Supplementary Material - ThermVision: Exploring FLUX for Synthesizing Hyper-Realistic Thermal Face Data and Animations via Image to Video Translation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3746027.3755448

1 Training Data Samples

This section presents the training data samples sourced from three distinct public datasets utilized during the training phase as shown in Figure 1. Notably, these datasets vary in image resolution and encompass a diverse range of head poses, genders, and facial expressions. This diversity is essential for ensuring comprehensive representation and enhancing the generative capability of the FLUX model across multiple datasets, leading to more effective training and improved performance [1].

2 Training Configuration Overview

Table 1 presents the key training parameters used for the Flux model. The configuration includes a learning rate of 0.0004, a network dimension of 64, and a total of 3000 training steps. Specific optimization techniques including `logit_normal` weighting, `shift` timestep sampling, and `sdpa` attention mode are employed to enhance model performance. Additionally, disk caching is enabled for both latent and text encoder outputs to optimize memory usage. The model utilizes `bf16` precision for both gradient and save data types, ensuring efficient computation. These settings collectively contribute to the effective training process which furthers contributes in generation of high-quality synthetic thermal images.

3 Additional Data Synthesis and Animation Results

Additional data synthesis along with face animations video results are provided in separate folders, (folder name: 'ThermVision Face Synthesis Results' and 'ThermVision Face Animations Results') showcasing 2D data synthesis with various smart transformations and animation results for both head pose and facial expression synthesis.

4 Thermal Facial Detection/ Localization Results

We have provided face localization results for both male and female subjects in the form of random individual frames, as well as complete video sequences (folder name: 'ThermVision Face Localization and Annotation Results'), demonstrating the consistency and accuracy of the face detector across different facial variations.

Parameter	Value
network_dim	64
network_alpha	64.00
learning_rate	0.000400
max_train_steps	3000
apply_t5_attn_mask	true
cache_latents	disk
cache_text_encoder_outputs	disk
blocks_to_swap	1
weighting_scheme	logit_normal
logit_mean	0.00
logit_std	1.00
mode_scale	1.29
timestep_sampling	shift
sigmoid_scale	1.0
model_prediction_type	raw
guidance_scale (Train and Val)	3.00
discrete_flow_shift	3.1582
highvram	false
fp8_base	true
gradient_dtype	bf16
save_dtype	bf16
attention_mode	sdpa

Table 1: Flux Model Training Parameters

*Corresponding author.

[1] https://mali-farooq.github.io/ThermVision/

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755448

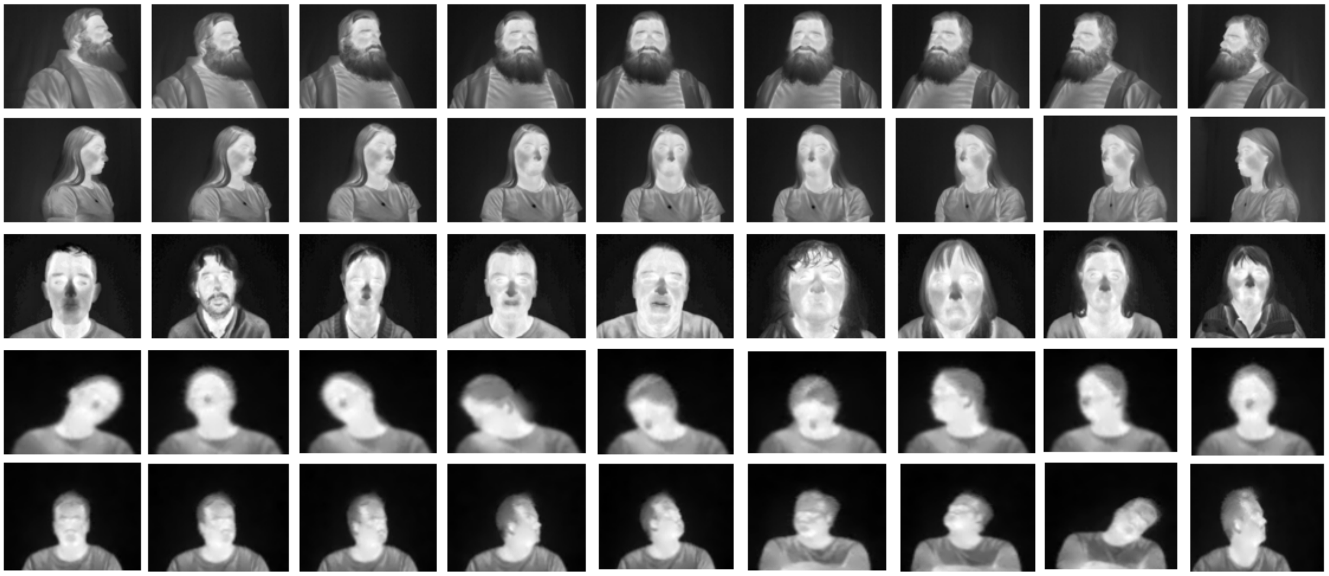

Figure 1: Training data samples acquired from Tufts, Carl and Charlotte ThermalFace database. The first and second row shows the data male and female data samples from Tufts thermal dataset, the third row shows male and female data samples from CARL thermal dataset where fourth and fifth row shows male and female data samples from Charlotte ThermalFace dataset.

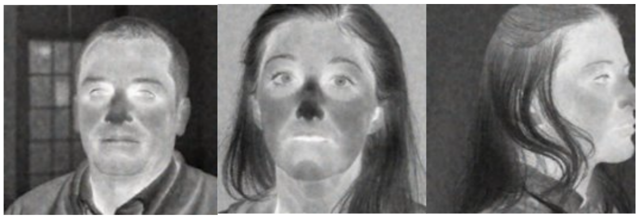

Figure 2: Thermal image generation issues: The left and middle images exhibit thermal noise hallucination and unrealistic temperature distribution, while the rightmost image shows a loss of facial structure under extreme yaw conditions.

5 Thermal Facial Landmark Detection Results

Additional facial landmark detection results generated using the pre-trained SF-TL54 landmark modelfeaturing 54 keypoints and face bounding boxes, are provided for male and female subjects across diverse head poses and facial expressions. These results are included in the supplementary material folder titled *'ThermVision Face Landmarks Results'*.

6 Limitations and Artifacts in Generated Data

Although diffusion-based models demonstrate promising capabilities in synthesizing thermal images, few limitations persist in the quality and semantic accuracy of the generated outputs. In particular, the generated images in few cases lack realistic heat distribution patterns that are characteristic of genuine thermal imagery. This absence of structured thermal gradients reduces the physiological plausibility of the outputs and may limit their applicability in tasks that rely on accurate thermal cues. Moreover, in some cases, essential facial features such as the eyes, nose, or mouth are either poorly defined or entirely missing. These omissions are especially problematic in scenarios involving human identification, facial analysis, or biometric and multimedia applications, where such details are critical. Some of the cases are refelcted in figure 2.