Supplementary - F-ViTA: Foundation Model Guided Visible-to-Infrared Translation

Jay Nitin Paranjape
Johns Hopkins University
jparanj1@jhu.edu

Celso M de Melo
DEVCOM Army Research Laboratory
celso.m.demelo.civ@army.mil

Vishal M. Patel
Johns Hopkins University
vpatel36@jhu.edu

Figure 1. The rows represent the RGS output and our corresponding predictions on FLIR.

1. Discussion on RAM-Grounded SAM (RGS)

: Some examples of RGS outputs are shown in Fig. 1. We observe that not all objects are recognized by RGS for most of the frames. Hence, it is wasteful to disregard these examples. Instead, we would like to use the masks as weak signals to improve the model prediction for the masked regions. From this exercise, we observe that (a) the model performs better in regions recognized by RGS than other regions (woman in row2 vs other people, vehicles in row1 vs trees). However, it still learns to translate the other regions satisfactorily (people and vehicles in row2). Thus, improving RGS can improve the translation quality. (b) The diffusion model is not that affected by the labels themselves as long as the embeddings are similar. From the image row 2 (requires zooming), the predicted label "palm tree" gets similar treatment as label "tree". At the same time, rare concepts like "ponytail" in row 2 do not seem to affect the model even though they are predicted by RGS.

2. Failure Case Analysis

F-ViTA is fairly robust to the misses or over-labelling by RAM-Grounded SAM (RGS). For example, it sometimes labels the helmet and clothes separately but does not segment the person as a whole (Fig. 2 (a)), and in some cases, it completely misses the person if they appear small in the image (Fig. 2 (b,c)). Yet, the final prediction is able to generate a valid thermal image, since RGS-generated labels serve only as a weak signal to the diffusion process. That said, we generally observe better results for objects correctly predicted by RGS, showing that F-ViTA can be improved continuously as foundation models advance. In addition, there are limiting cases where both the diffusion model and RGS fail to capture certain objects (Fig. 2 (d)), which can be addressed by improving either or both models in future work.

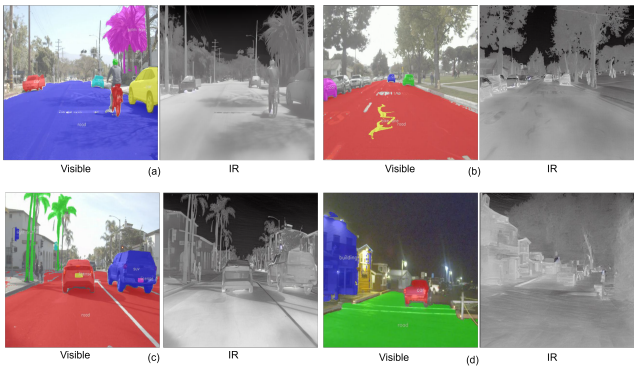

Figure 2. (a),(b),(c) denote the cases where RGS failed to detect pedestrians but F-ViTA was able to correctly depict them in the IR image, suggesting some robustness to RGS failures. (d) shows the case where the second car is not detected and F-ViTA also mixes up the two vehicles in the output. This shows a limiting case of our method.

Method	Inference time (s)	FID (↓)	LPIPS (↓)	SSIM (↑)	PSNR (↑)
InfraGAN	0.14	241.42	0.36	0.87	17.15
EGGAN-M	2.5	238.41	0.35	0.93	12.81
EGGAN-U	0.08	<u>152.09</u>	0.23	0.95	18.45
Instruct-Pix2Pix [1]	138.39	0.23	0.72	18.25	
PID	11.6	179.65	<u>0.20</u>	<u>0.95</u>	<u>19.22</u>
F-ViTA (Ours)	3.2	**118.52**	**0.16**	**0.96**	**27.55**

Table 1. Results on LiTiV dataset with inference time. Bold - best, underline - second best

3. Computational Efficiency

F-ViTA is a diffusion based method and hence requires multiple inference steps. For this paper, 100 steps have been used per inference, which translates to roughly 3.25 seconds per image, as tested on LiTiV dataset. Out of this, roughly 0.2s are taken up by running inference on GSAM. As shown in the Table 1, while the time is greater than GAN-based methods, the resulting improvements are significant for the amount of additional time. PID, in comparison is also a diffusion-based method but requires much more time than our method due to a more complex denoising model.

4. Datasets

Long-Wave Infrared (LWIR): We use the FLIR-ADAS [4] and KAIST [5] datasets within this wavelength range. FLIR-ADAS has thermal images in the range of $7.5 - 13.5\mu m$. However, there is a slight misalignment between the natural and thermal images. Therefore, we use the aligned version of this dataset [7], which includes 4,890 training and 126 testing visible-thermal pairs. The KAIST dataset also contains thermal images within the $7.5–13.5\mu m$ range, representing road scenes, with 12,538 training pairs and 2,252 testing pairs.

Mid-Wave Infrared (MWIR): We use two datasets to conduct experiments within this range. The OSU dataset [3] contains images from both the MWIR as well as the LWIR domain. There are 4,862 training pairs and 3,683 testing pairs. The LiTiV dataset [6] also consists of scenes involving roads and pedestrians. It contains a total of 4,564 training and 1,761 testing pairs.

Near Infrared (NIR): We use the NIRScene [2] dataset to represent this wavelength, which contains sceneries and landscapes. This dataset is split into 381 training pairs and 96 testing pairs.

5. Implementation Details

We use the SwinT-OGC checkpoint for Grounding DINO, swin large checkpoint for RAM and vit-base model for SAM. We retain the default settings of the foundation models during training as well as testing. The training was done on a single NVIDIA RTX6000 GPU with a batch size of 1. The base learning rate was set to 5e-5.

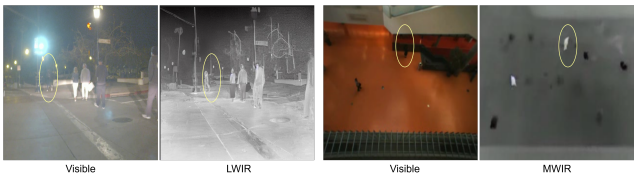

Figure 3. Pedestrians that are difficult to make out in RGB images (yellow ovals) can be easily identified in the generated IR images.

6. Pedestrian Detection Examples

One of the most important applications of thermal imagery across autonomous driving, surveillance systems, robotics etc. is person / pedestrian detection. In many cases, RGB images may not be sufficient to detect people as they may be obscured or not clearly visible due to various environmental conditions and lighting. In such cases, thermal images, which are mostly free of environmental disturbances, can be used to easily detect the people. In this regards, F-ViTA acts as a complement to the real data collected using IR sensors, which is both expensive and tedious, and hence may be limited in quantity. Some examples of F-ViTA outputs are shown in Figure 3. Here, the yellow ovals represent people who are not clearly visible in the RGB image but are easily distinguishable in the generated thermal image.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[2] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR11)*, pages 177–184, Colorado Springs, 2011. 2

[3] James W. Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2):162–182, 2007. Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum. 2

[4] FLIR Systems. Flir adas dataset. https://www.flir.com/oem/adas/adas-dataset-form/, 2025. Accessed: 2025-02-20. 2

[5] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 2

[6] Ichraf Lahouli, Rob Haelterman, Zied Chtourou, Geert De Cubber, and Rabah Attia. Pedestrian detection and tracking in thermal images from aerial mpeg videos. 2018. 2

[7] Zona Qiu. Flir-align: A repository for flir thermal and rgb image alignment. https://github.com/zonaqiu/FLIR-align, 2025. Accessed: 2025-02-20. 2