

# F-ViT: Foundation Model Guided Visible to Thermal 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  
vpatel136@jhu.edu

## Abstract

Thermal imaging is crucial for scene understanding, particularly in low-light and nighttime conditions. However, collecting large thermal datasets is costly and labor-intensive due to the specialized equipment required for infrared image capture. To address this challenge, researchers have explored visible-to-thermal image translation. Most existing methods rely on Generative Adversarial Networks (GANs) or Diffusion Models (DMs), treating the task as a style transfer problem. As a result, these approaches attempt to learn both the modality distribution shift and underlying physical principles from limited training data. In this paper, we propose F-ViT, a novel approach that leverages the general world knowledge embedded in foundation models to guide the diffusion process for improved translation. Specifically, we condition an InstructPix2Pix Diffusion Model with zero-shot masks and labels from foundation models such as SAM and Grounded DINO. This allows the model to learn meaningful correlations between scene objects and their thermal signatures in infrared imagery. Extensive experiments on five public datasets demonstrate that F-ViT outperforms state-of-the-art (SOTA) methods. Furthermore, our model generalizes well to out-of-distribution (OOD) scenarios and can generate Long-Wave Infrared (LWIR), Mid-Wave Infrared (MWIR), and Near-Infrared (NIR) translations from the same visible image. Code: <https://github.com/JayParanjape/F-ViT/tree/master>.

## 1. Introduction

Many critical applications in autonomous driving, robotics, and surveillance require a precise understanding of natural scenes that remains robust in low-light and nighttime conditions. Thermal imaging has been widely explored for scene understanding and object recognition in such environments, where external illumination is impractical due to various constraints.

The infrared spectrum is typically divided into two regions: a reflection-dominated region, which includes the

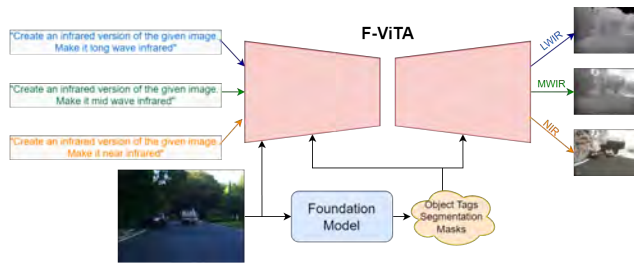


Figure 1. Our model, F-ViT, leverages pretrained foundation models to extract object tags and segmentation masks from visible images in a zero-shot manner, using this information to enhance translation to the thermal domain. Additionally, F-ViT enables user-guided infrared image generation through text prompts, allowing for the synthesis of specific infrared types—an ability not explored in existing methods.

Near-Infrared (NIR) and Shortwave Infrared (SWIR) bands, and an emission-dominated thermal region, comprising the Midwave Infrared (MWIR) and Longwave Infrared (LWIR) bands. Given its resilience to challenging lighting conditions, thermal imagery provides a valuable source of information for these applications [3, 4, 7, 14, 20, 29].

Recent studies have explored RGB-IR fusion to enhance tasks such as pedestrian detection [6, 21], segmentation [27, 30], and object detection [17, 33]. These approaches typically rely on paired RGB and thermal images from the same scene. However, capturing thermal images requires specialized hardware, including infrared cameras and sensors, making large-scale dataset collection both costly and labor-intensive. As a result, thermal datasets remain relatively scarce in the literature—especially paired RGB-thermal data—driving research efforts toward automatically generating infrared images from visible images.

Various methods have been proposed for image translation using Generative Adversarial Networks (GANs) [13, 16, 35] and Diffusion Models (DMs) [19]. These approaches aim to learn the distribution shift between the two domains while preserving the physical properties of infrared images. However, in the low-data regime of RGB-to-thermal translation, such models are highly prone to overfitting, which significantly limits their generalizability.

Recently foundation models (FMs) have been introduced that are trained on large-scale natural data from the internet. These models learn highly representative features for visual data that are beneficial for various downstream tasks and promote generalized performance. Some examples include CLIP [23] for image-text alignment, Grounding DINO [18] for open-world object detection, Segment Anything (SAM) [12] for prompted image segmentation and SAM2 [24] for prompted video segmentation. Given the generalizable nature of such models, using the knowledge contained in them can help address the limited data issues with the current RGB-thermal translation methods. Motivated by this, we propose **F-ViT**, which uses implicit guidance from FMs to improve the translation quality. For a given natural image, we first pass it to the FM to extract labeled masks of all objects present in the scene. Then, we inject this information during the diffusion process during training the translation model, which is based on Instruct Pix2Pix [1]. This can encourage the model to correlate the output pixel intensity of the masked regions with the label of the object. For example, regions with the label “person” should be given a higher pixel intensity in the thermal image. Such guidance also allows our model to do well on datasets with limited training pairs. Thus, our contributions are as follows:

1. We present F-ViT, a Diffusion Model-based approach for visible-to-thermal image translation that utilizes outputs from pretrained FMs to guide the translation process.
2. We show that F-ViT outperforms existing SOTA methods on five public datasets from NIR, MWIR and LWIR categories on multiple metrics.
3. We also show that F-ViT is able to generalize well to OOD data and can also generate the translated IR image from any of the three bands (LWIR, MWIR, NIR) given the user intent through a text prompt, as shown in Fig. 1. This is one of the first methods that shows such a functionality for visible-to-thermal translation.

## 2. Related Work

Visible-to-thermal translation has been widely studied in the literature, with a majority methods attempting to solve this problem using GANs. Some early methods use Pix2Pix [10], an image-to-image translation model which uses conditional adversarial networks and cycle-GAN [34], which adds an additional consistency loss for stable training. These methods are not specific to visible-to-thermal translation. ThermalGAN [13] was introduced for this task, with a focus on person identification using the thermal signature. InfraGAN [35] is also a GAN-based approach that performs translation to the infrared domain for general scenes using pixel-wise loss functions and structural similarity. Recently, EGGAN [16] was introduced that uses information from the edge map of the natural scene to guide the translation pro-

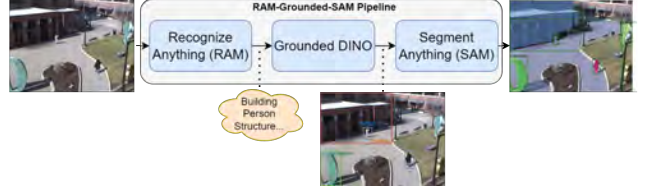


Figure 2. An example usage of the RAM-Grounded-SAM pipeline. The dotted lines indicate the output at every step.

cess. For this, they add an additional constraint to preserve the edges in the image.

As an alternative to the above GAN-based approaches, PID [19] was recently introduced, which uses DMs to perform the translation. In addition, this paper introduces physics-motivated loss functions that preserve the physical characteristics of the generated IR image. However, PID requires two-stage training, where the dataset-specific physics parameters are learnt first and then used as ground truths during the translation training. Furthermore, this work uses a common set of parameters for the dataset as a whole, rather than have a physical parameter for every object in the scene and has only been proposed for LWIR. Our method also uses a DM for translation, but it differs from PID in being a single-stage end-to-end approach. In addition, unlike PID, which explicitly learns dataset-level parameters, we implicitly try to learn object-level correlations by passing labeled maps during the diffusion process. Moreover, we show that our method produces good results for LWIR, MWIR as well as NIR images.

## 3. Methodology

### 3.1. Preliminaries: Grounded-SAM

Given a visible image, we use Grounded-SAM [25] to produce labeled masks. Grounded-SAM assembles Grounding-DINO [18], which is an open-world prompted object detector, and SAM [12], which is a promptable image-segmentor. Given a text prompt and an image, Grounding DINO can detect the object corresponding to the text input and shows a strong zero-shot performance thanks to its training data. Similarly, SAM was trained on 11 million images and 1.1 billion masks and exhibits a strong zero-shot performance for segmenting out an object corresponding to a point, box or mask prompt. Thus, given a text descriptor of an object, Grounded-SAM first employs Grounding DINO to extract a bounding box for that object and passes it to SAM as a prompt, thus generating the labeled mask, where the input text is the label.

**RAM-Grounded-SAM:** Grounding-SAM can be used for automatic dense image annotation [25], where given an RGB image, the output is a list of masks for the objects in the scene, which are annotated with their corresponding text labels. For this, a foundational image tagging model

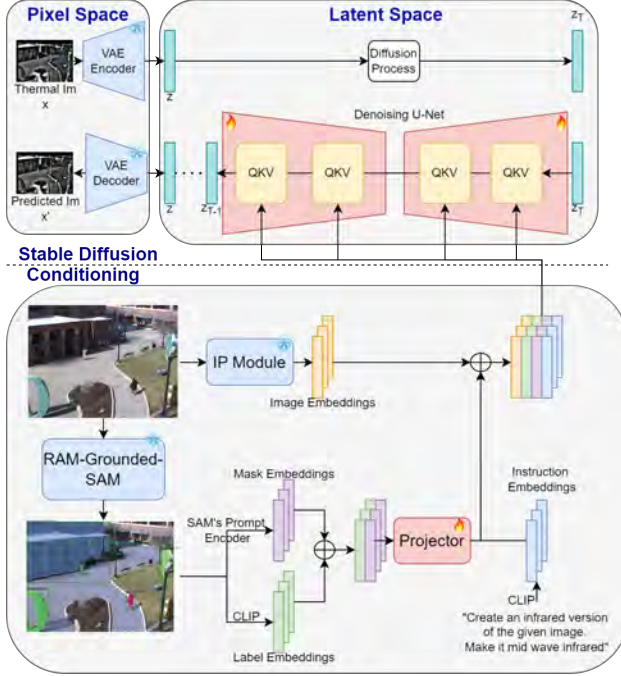


Figure 3. Training pipeline. The stable diffusion part is responsible for learning the distribution of thermal images. The conditioning part provides the visible image to enforce structural similarity as well as guidance from the foundation model and text instructions for improved translation. Only the Denoising UNet and the Projector module are trainable in the pipeline.

like Recognize-Anything (RAM) [32] is used. RAM is trained to detect all objects in a scene and generate tags for them. It is trained on a large corpus of image-caption pairs from the internet, which are parsed to obtain the image with its corresponding tags. Given these pretrained foundation models, the RAM-Grounded-SAM pipeline passes an image through the RAM model to generate tags for all the present objects. These tags are passed to the Grounding-DINO model to generate corresponding bounding boxes, which are further passed to SAM to generate masks. Note that this entire pipeline does not involve any trainable parameters. Fig. 2 shows an example of the pipeline output. These foundation models are pretrained on a vast pool of diverse datapoints and thus are able to produce valuable priors for tasks like visible-to-infrared translation that are based in a low-data regime.

### 3.2. Model Architecture

We use Stable Diffusion [26] for the translation process, similar to Instruct-Pix2Pix [1]. As shown in Fig. 3, we use a pretrained encoder to convert the thermal image to the latent space, where it undergoes the forward diffusion process and the backward denoising process. The denoising is done using a UNet conditioned with additional prompts to control

the generation process. The conditioning involves attention layers for each U-Net block, where the keys and the values are generated using the visible image and the text instructions and the queries are latent vectors at different timesteps. Finally, the pretrained decoder converts the generated latent back to the pixel space. In Instruct-Pix2Pix, the conditioning described before is done using an original image and a detailed text instruction, which describes the user-intended change to the original image. In contrast, F-ViT also adds two additional conditioning embeddings derived from foundation model outputs.

We pass the visible image through the frozen RAM-Grounded-SAM pipeline, which generates a number of text-annotated masks for the objects present in the image. We pass these masks through the prompt-encoder provided by SAM to generate discriminative embeddings. In addition, we pass the text-labels for every object through the pretrained CLIP encoder to generate text embeddings. These are concatenated along feature axis and passed through a trainable projector. This module is responsible for learning a transform that converts the FM embeddings to the space expected by Stable Diffusion. These are then concatenated to the image embeddings, obtained by passing the visible image through Instruct-Pix2Pix’s IP module, and instruction embeddings, which are CLIP embeddings of the text instruction. The latter two embeddings are unique for a given image-instruction pair, while the former two vary for every object. Hence, we repeat the latter two embeddings to be equal to the number of detected objects, for concatenation. For the text instruction, we generate 50 synonymous sentences for the text: **“Create an infrared version of the given image. Make it *< type >* wave infrared.”** Here, *< type >* can take the values in {near, mid, long}. Thus, we generate a conditioning tensor using the image, information from the foundation models, and a text instruction.

We train the entire pipeline on visible-thermal image pairs. Only the projector module and the denoising U-Net are the trainable modules in the pipeline. Since the conditioning tensors contain information about an object and its location in the image, we expect the model to learn physical correlations between the heat signatures of the object and the pixel intensity in the generated thermal image at the location of the object, for example, generating higher intensity pixels at the location of the object “person”. While existing methods expect the model to learn everything about the translation from the data, we utilize the guidance provided by FMs to reduce the load by providing localization and object information. Hence, our method is able to outperform existing methods in low-data regimes.

Method	FID ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	PSNR ( $\uparrow$ )
Pix2Pix [10]	199.06	0.38	0.30	16.34
ThermalGAN [13]	407.01	0.60	0.15	11.59
InfraGAN [35]	140.01	0.50	<u>0.41</u>	16.95
EKGAN-M [16]	139.11	0.43	0.29	10.32
EKGAN-U [16]	140.92	0.55	0.37	14.3
PID [19]	<b>84.26</b>	<u>0.36</u>	0.40	<u>17.26</u>
F-ViT (Ours)	<u>133.30</u>	<b>0.23</b>	<b>0.73</b>	<b>18.73</b>

Table 1. Results on FLIR-ADAS [8] dataset. Bold - best, underline - second best

## 4. Experiments and Results

### 4.1. Datasets

**Long-Wave Infrared (LWIR):** We use the FLIR-ADAS [8] and KAIST [9] 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 [22], 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 [5] 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 [15] 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.

### 4.2. 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  $5e-5$ .

### 4.3. Image Translation Experiments

We compare our method with existing SOTA visible-to-thermal translators on five datasets. The results on FLIR-ADAS are tabulated in Tab. 1. We find that our method is able to outperform existing methods on multiple metrics, with a significant rise observed in SSIM. However, we see a lower FID score than the current SOTA. While FID compares the distributions between the ground truth and the

Method	FID ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	PSNR ( $\uparrow$ )
Pix2Pix [10]	132.04	0.22	0.69	21.25
ThermalGAN [13]	277.85	0.24	0.66	19.74
InfraGAN [35]	222.96	0.159	0.76	<u>22.97</u>
EKGAN-M [16]	79.45	0.37	0.48	10.65
EKGAN-U [16]	76.27	0.27	0.63	17.07
PID [19]	<b>51.69</b>	<b>0.14</b>	<u>0.79</u>	<b>23.6</b>
F-ViT (Ours)	<u>76.18</u>	<u>0.21</u>	<b>0.88</b>	18.96

Table 2. Results on KAIST dataset [9]. Bold - best, underline - second best

Method	FID ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	PSNR ( $\uparrow$ )
EKGAN-M [16]	542.51	0.47	0.55	12.95
EKGAN-U [16]	75.76	0.18	<b>0.86</b>	<u>20.46</u>
PID [19]	85.58	0.25	0.58	14.1
F-ViT (Ours)	<b>61.97</b>	<b>0.15</b>	<u>0.78</u>	<b>20.92</b>

Table 3. Results on the OSU dataset [5]. Bold - best, underline - second best

Method	FID ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	PSNR ( $\uparrow$ )
InfraGAN [35]	241.42	0.36	0.87	17.15
EKGAN-M [16]	238.41	0.35	0.93	12.81
EKGAN-U [16]	<u>152.09</u>	0.23	0.95	18.45
PID [19]	179.65	<u>0.20</u>	<u>0.95</u>	<u>19.22</u>
F-ViT (Ours)	<b>118.52</b>	<b>0.16</b>	<b>0.96</b>	<b>27.55</b>

Table 4. Results on the LiTiV2012 dataset [15]. Bold - best, underline - second best

Method	FID ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	PSNR ( $\uparrow$ )
EKGAN-M [16]	126.2	0.18	0.83	15.25
EKGAN-U [16]	125.6	0.23	0.79	18.45
PID [19]	166.67	0.31	0.62	14.21
F-ViT (Ours)	<b>63.67</b>	<b>0.11</b>	<b>0.85</b>	<b>19.23</b>

Table 5. Results on the NIRScene dataset [2]. Bold - best, underline - second best

generated thermal image, SSIM measures the structural and visual soundness of the generated image. Hence, both are important to gauge the quality of generation. On the KAIST dataset, as can be seen in Tab. 2, we see a similar trend where the SSIM for our method is better but other metrics are comparable to PID. This is not surprising, given that PID models the physical quantities behind LWIR images and constrains the model accordingly. In contrast, our model captures the structure of objects better due to localization information present through the masks (causing improved SSIM), but at the same time has a slight tendency to miss



the differences in output intensities for different parts of the same object (causing increased FID). This can happen since there are multiple similar signals from the prompts for different parts of the same object. In addition, note that FLIR is a comparatively smaller dataset than KAIST, which is why we see an overall better performance with our method.

Tab. 3 shows the results on the OSU dataset, where we see a significant improvement in all metrics, while SSIM is comparable. Similarly, for LiTiV, as shown in Tab. 4, and for NIRScene, as seen in Tab. 5, we find that our method outperforms all other methods across all metrics. Note that these are cases of low-data regimes, where the advantages of our model are clearly demonstrated. We present qualitative results in Fig. 4 across all five datasets used for comparison, selecting examples from both indoor and outdoor scenes. As shown in the figure, our method (third column) performs consistently well across all datasets. PID excels in LWIR but struggles to generate accurate translations for MWIR and shows an overall increased intensity in the NIR case. This is expected, as the theoretical analysis and formulation of PID are primarily designed for LWIR translations. In contrast, EGGAN-M and EGGAN-U slightly distort object structures, as observed in the figure. In comparison, our method, F-ViT, generates thermal images that closely resemble the ground truth.

#### 4.4. Out-of-Distribution (OOD) Experiments

We use the MFNet [28] dataset to evaluate the generalizability of our method on out-of-distribution (OOD) data. MFNet contains visible images and their thermal counterparts in the LWIR range. As shown in Tab. 6, F-ViT, trained on the KAIST dataset, performs significantly better on MFNet compared to PID trained on KAIST. Similarly, for models tuned using only the FLIR dataset, our method outperforms PID trained on the same data. PID first learns estimates for physical quantities related to the thermal dataset as a whole, while our method implicitly learns object-wise estimates. As a result, our approach exhibits greater robustness when transferred to a new dataset, as reflected in the table. Additionally, we conduct an experiment where we train F-ViT using one dataset from each modality (KAIST for LWIR, LiTiV for MWIR, and NIRScene or NIR for NIR) and then test its performance on MFNet with the instruction to convert to long-wave infrared. We find that even with the inclusion of additional modalities, the performance on MFNet remains comparable to when trained using only the LWIR KAIST dataset.

#### 4.5. Text-Prompted Translation Experiments

F-ViT uses text instructions for guiding the diffusion model to perform the image translation. Hence, it is possible for it to learn to generate images for specific infrared spectra by specifying it in the instructions. To explore

Method	FID (↓)	LPIPS (↓)	SSIM (↑)	PSNR (↑)
PID [19] (tuned on KAIST)	196.10	0.39	0.85	13.53
Ours (tuned on KAIST)	<b>70.98</b>	<b>0.30</b>	<b>0.87</b>	<b>16.33</b>
PID [19] (tuned on FLIR-ADAS)	142.60	0.47	0.70	12.02
Ours (tuned on FLIR-ADAS)	<b>100.54</b>	<b>0.35</b>	<b>0.81</b>	<b>12.85</b>
Ours (tuned on all datasets from all wavelengths)	94.5	0.37	0.83	13.72

Table 6. OOD Results on the MFNet dataset [28]. Bold - best, underline - second best

% original data used	MIoU (↑)	Mean Pixel Accuracy (↑)
0	0.501	55.58
10	0.503	56.01
25	0.520	59.22
50	0.526	59.1
100	0.527	59.6

Table 7. Results on MFNet [28] using Sigma [30] as the RGB-T segmentation pipeline. The rows indicate the percentage of real thermal data used in the training data, with the rest of the data being synthetic. We see only a small drop in the metrics as we go up the table (more synthetic data used), towards the applicability of our method.

this, we train F-ViT on KAIST (LWIR), OSU (MWIR), LiTiV (MWIR) and NIRSCENE (NIR) together with the type of image in the instruction. For instance, for KAIST, we ask the DM to generate a long wave infrared image. Once trained, we test the model’s ability to generate all three types of IR images from a given RGB image from the test data. Some examples of such translations can be seen in Fig. 5. From the figure, we observe that F-ViT produces reasonable predictions for LWIR, MWIR, and NIR given a single RGB image. The GT column represents the ground truth thermal image from the dataset-specific wavelength range. The first five rows display images from the test sets of the datasets on which the model was trained. The last three rows show predictions on MFNet, which was not used during training, demonstrating strong out-of-distribution (OOD) performance with F-ViT. The final row illustrates a failure case where the model misses pedestrians in the MWIR and NIR predictions, indicating potential areas for improvement. Notably, no existing methods demonstrate such multi-spectral translation capability. We are the first to explore this direction with F-ViT and encourage further research in this area.

#### 4.6. Downstream Applications

**RGB-T Fusion for Segmentation:** Thermal imagery is often used for multi-modal segmentation, where the RGB image and its paired thermal image are processed and fused to produce the output segmentation map. We conduct experiments to show an application of our pipeline for thermal image generation. We use the F-ViT model, tuned on the

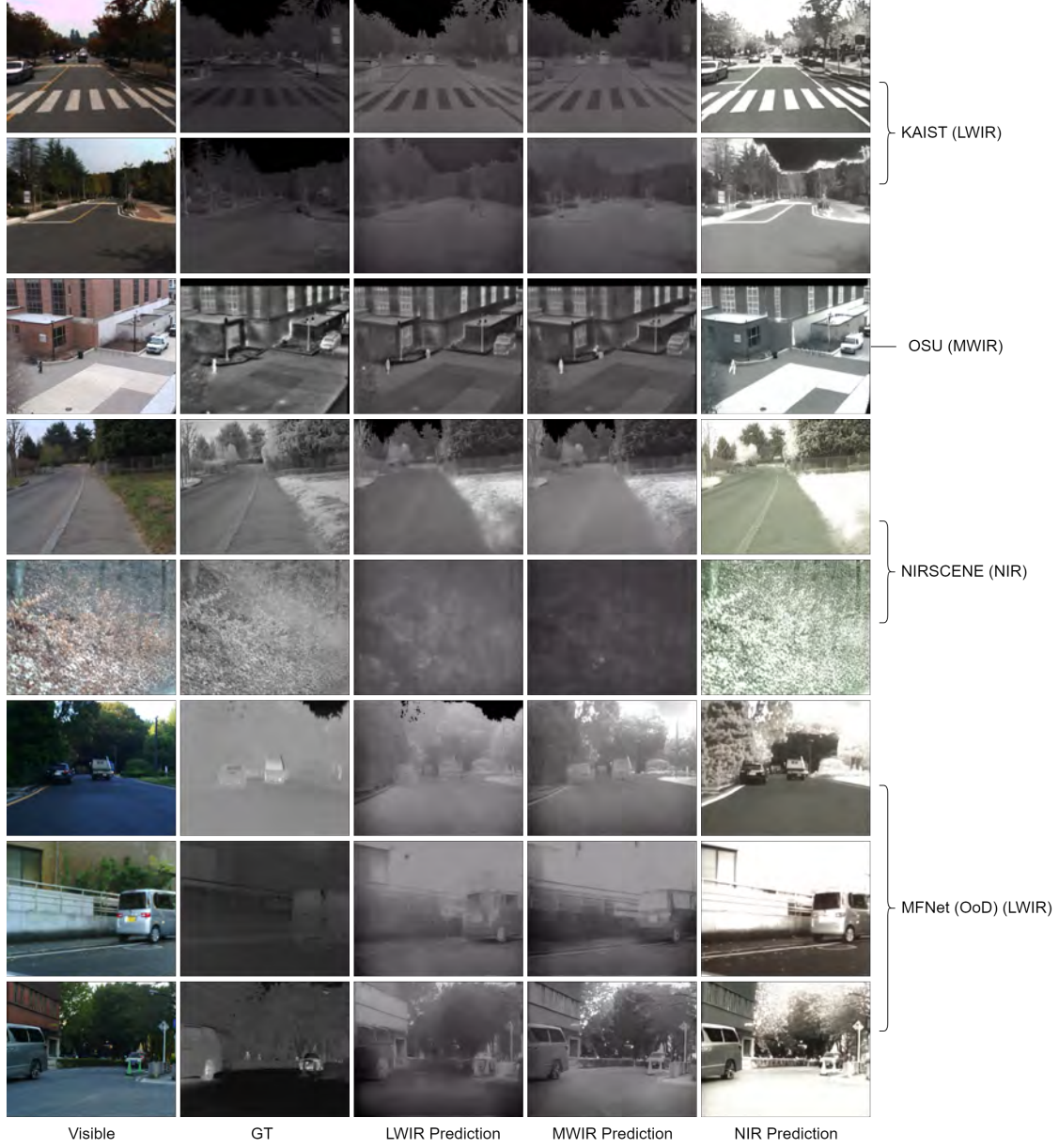


Figure 4. Qualitative Results over datasets from three different wavelength spectra. Our method (third column) is able to generate images more similar to the ground truth as compared to existing methods (fourth, fifth and sixth columns)

KAIST dataset, to generate thermal images for the MFNet [28] dataset. Next, we generate five types of data for training a RGB-T segmentation model, using 0%, 10%, 25%, 50%, and 100% of the actual thermal images respectively. The rest of the images are taken from the generated predictions from our model. We use the original thermal images for the testing data. We use the SOTA RGB-T fusion model, called Sigma [30] for the segmentation task. Sigma uses the recently introduced V-Mamba framework along with a concatenation and cross-attention mechanism to effectively fuse the two modalities. With this setup, we train Sigma

with the five datasets and test its performance, as tabulated in Tab. 7.

We observe that when only the original thermal data is used, Sigma gives a MIoU of 52.7 and mean pixel accuracy of 59.6. When we replace 50% of the data with the synthetic data generated with our method, the MIoU drops to 52.6 and accuracy to 59.1, which is not a significant drop compared to the reduction in original training data. With an all-synthetic training set, these numbers only reduce by 4 points. Note that the thermal data was created using F-ViT trained on KAIST only, thus being OOD for the generation

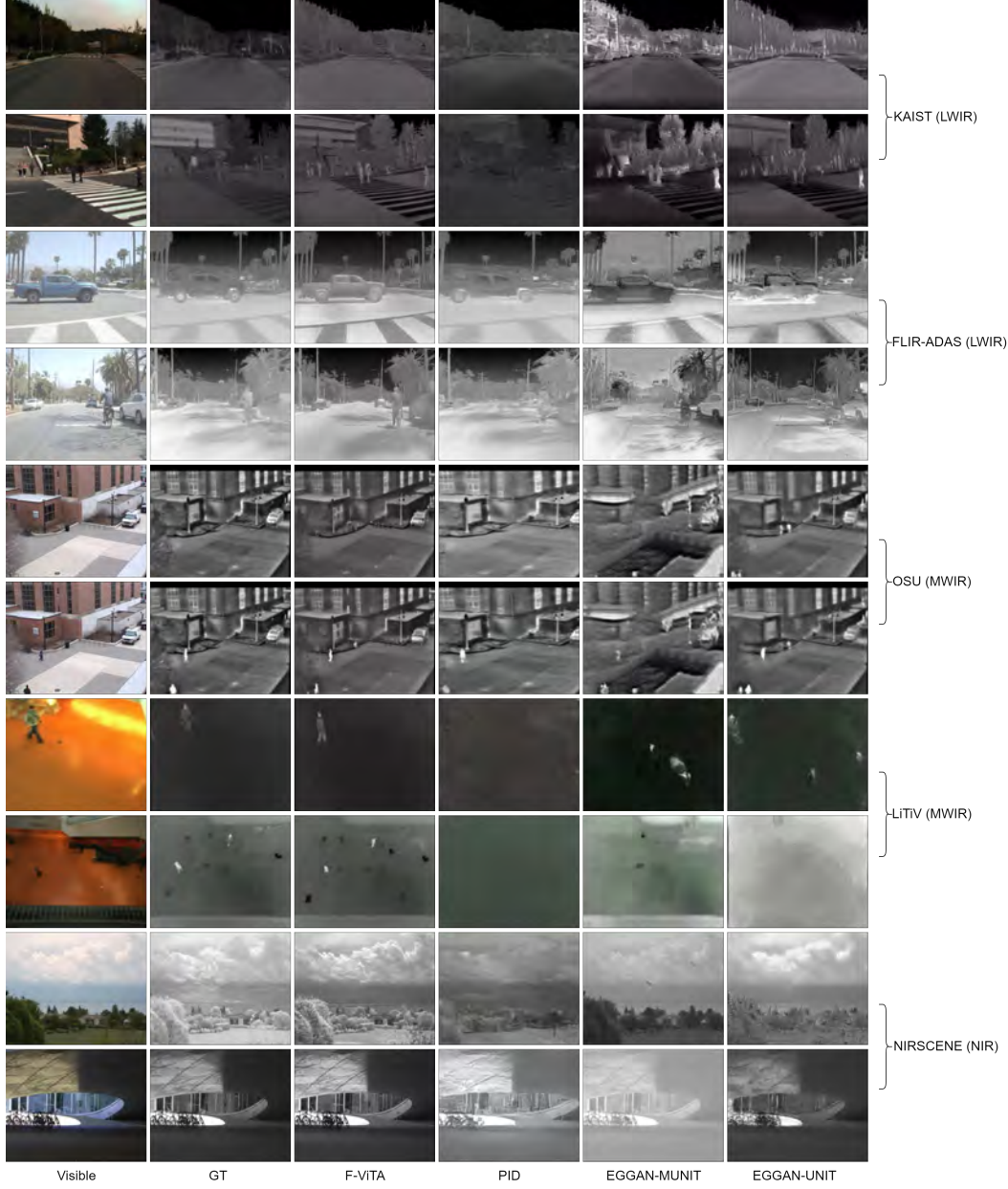


Figure 5. Text prompted translation capability of F-ViT. Our method is able to generate LWIR, MWIR or NIR images based on the text instruction. The second column shows the ground truth which is from the wavelength range specified next to the dataset name on the right.

model. Thus, these results also denote the quality of our predictions.

**Thermal Image Segmentation:** In this experiment, we train a Segformer [31] on the thermal training data using the same mixtures as in the previous experiment, and then evaluate its performance on the test data containing original thermal images. Note that we do not use RGB images here, making this a purely semantic segmentation task based on thermal images. As shown in Tab. 8, with 100% original training data, we achieve a MIOU of 0.38 and a mean ac-

curacy of 54%. Even when replacing 50% of the data with synthetic images, we observe comparable performance. Below this threshold, we see a noticeable decrease in accuracy, suggesting areas for further exploration. Nevertheless, the predictions from our method offer a viable alternative to capturing infrared images using specialized hardware.

**Pedestrian Detection in Thermal Imagery:** In this experiment, we use the thermal images generated by F-ViT for the KAIST test dataset as synthetic data to generate five splits, with 0%, 10%, 25%, 50% and 100% being the real



% original data used	MIoU ( $\uparrow$ )	Mean Pixel Accuracy ( $\uparrow$ )
0	0.30	43.38
10	0.30	45.24
25	0.30	40.01
50	0.38	51.29
100	0.38	54.12

Table 8. Results on MFNet [28] using Segformer [31] for thermal segmentation for 10 classes. The rows indicate the percentage of real thermal data used in the training data, with the rest of the data being synthetic. We see only a slight drop in the metrics as we go up the table (more synthetic data used), which suggests a use case of F-ViT for synthetic data generation.

% original data used	P ( $\uparrow$ )	R ( $\uparrow$ )	MAP@50 ( $\uparrow$ )
0	0.63	0.45	0.48
10	0.64	0.56	0.57
25	0.64	0.60	0.60
50	0.67	0.62	0.62
100	0.67	0.59	0.62

Table 9. Results on KAIST [9] using YOLOv8 [11] for pedestrian detection. The rows indicate the percentage of real thermal data used in the training data, with the rest of the data being synthetic. We see only a slight drop in the metrics as we go up the table (more synthetic data used), which suggests a use case of F-ViT for synthetic data generation.

data percentage and the rest filled using synthetic data as done previously. Then, we train YOLOv8 [11] on these splits and test the performance on real thermal data, which is not used during training. As shown in Tab. 9, we calculate the precision (P), recall (R) and the mean average precision at 0.5 threshold (mAP@50) for all cases. We observe that even after replacing 50% of the real data, there is on par performance and surprisingly, a higher recall. As we go up the table, we see that the drop in performance is quite small compared to the amount of real data replaced, thus signifying the usefulness of the generated thermal images for such downstream tasks. Similar to the segmentation task, while we see that completely using the thermal data still has room for improvement, F-ViT offers a viable solution for semi-supervised methods as seen from the table.

#### 4.7. Ablation Studies

We perform an ablation study over the components extracted from the foundation models that are passed to the diffusion process. As seen in Tab. 10, when neither of text, bounding boxes or mask embeddings are used (first row), we see a lower performance in all metrics. Adding just the text tags identified from the image (second row) improves these metrics. The text tags are extracted using the RAM model, and converted into embeddings using CLIP before passing to the DM. Next, we add only the embed-

Text Labels	Bounding Boxes	Masks	FID ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	PSNR ( $\uparrow$ )
			80.03	0.18	0.72	19.64
✓			76.43	0.20	0.74	20.5
	✓		76.01	0.20	0.75	20.01
		✓	69.81	0.19	0.76	20.7
✓	✓		75.18	0.17	0.76	20.24
✓		✓	<b>61.97</b>	<b>0.15</b>	<b>0.78</b>	<b>20.92</b>

Table 10. Ablation over the components passed to the diffusion process using the FM. ✓ indicates the component being used, with the first row indicating neither text, boxes or masks being used and the last row indicating F-ViT where text and masks are used.

dings from the bounding boxes for the objects detected by Grounding DINO (third row) and see on par metric values. However, adding mask embeddings from Grounded SAM (fourth row) gives us a slight improvement. This is expected since masks provide more precise localization than bounding boxes. In addition, unlike masks, a bounding box around bigger objects like buildings or trees may also encompass other objects like vehicles or people. Hence, we see an advantage when using masks. In the fifth row, we use both the text tags and the bounding boxes, which slightly improves the performance over second and third rows, where they are passed separately. Finally, the last row denotes the case used in F-ViT, with both the text embeddings and mask embeddings being provided to the DM. We see that this case improves over all other rows, since text and masks provide the localization and semantic information about the objects to the DM. We do not include cases where box and mask embeddings are used together, since adding boxes is superfluous in the presence of masks.

## 5. Conclusion

In this work, we introduce F-ViT, a diffusion model-based visible-to-thermal image translator that leverages pretrained foundation models to guide the diffusion process. We demonstrate that incorporating object identification, location, and semantic information via segmentation masks and text labels significantly enhances translation, particularly in low-data regimes. Extensive experimentation on five public datasets across LWIR, MWIR, and NIR spectra shows that F-ViT outperforms or matches existing methods on multiple metrics, with improved generalization on out-of-distribution datasets. Notably, F-ViT is the first method to enable conversion of the same RGB image into LWIR, MWIR, or NIR based on user-provided text instructions, opening new avenues for future research. We also highlight the potential applications of visible-to-thermal translation in downstream tasks like RGB-T fusion and thermal segmentation, demonstrating its value for semi-supervised learning in data-limited scenarios. Future work should further explore text-prompted inter-modality transfer, with applications in biometrics, robotics, and surveillance.



## 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. 2, 3
- [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. 4
- [3] Lidice E Castro Jimenez and Edgar A Martínez-García. Thermal image sensing model for robotic planning and search. *Sensors*, 16(8):1253, 2016. 1
- [4] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261, 2021. 1
- [5] 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. 4
- [6] Qing Deng, Wei Tian, Yuyao Huang, Lu Xiong, and Xin Bi. Pedestrian detection by fusion of rgb and infrared images in low-light environment. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2021. 1
- [7] Chiara Filippini, David Perpetuini, Daniela Cardone, Antonio Maria Chiarelli, and Arcangelo Merla. Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review. *Applied Sciences*, 10(8):2924, 2020. 1
- [8] FLIR Systems. Flir adas dataset. <https://www.flir.com/oem/adas/adas-dataset-form/>, 2025. Accessed: 2025-02-20. 4
- [9] 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. 4, 8
- [10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 4
- [11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 8
- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2
- [13] Vladimir V. Kniaz, Vladimir A. Knyaz, Jiří Hladůvka, Walter G. Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Computer Vision – ECCV 2018 Workshops*, pages 606–624, Cham, 2019. Springer International Publishing. 1, 2, 4
- [14] Mate Krišto. Review of methods for the surveillance and access control using the thermal imaging system. *Review of Innovation and Competitiveness: A Journal of Economic and Social Research*, 2(4):71–91, 2016. 1
- [15] Ichraf Lahouli, Rob Haelterman, Zied Chtourou, Geert De Cubber, and Rabah Attia. Pedestrian detection and tracking in thermal images from aerial mpeg videos. 2018. 4
- [16] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8291–8298, 2023. 1, 2, 4
- [17] Guibiao Liao, Wei Gao, Ge Li, Junle Wang, and Sam Kwong. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7646–7661, 2022. 1
- [18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 2
- [19] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. Pid: Physics-informed diffusion model for infrared image generation, 2024. 1, 2, 4, 5
- [20] Ben Miethig, Ash Liu, Saeid Habibi, and Martin v. Mohren-schildt. Leveraging thermal imaging for autonomous driving. In *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*, pages 1–5, 2019. 1
- [21] Ahmet Ozcan and Omer Cetin. A novel fusion method with thermal and rgb-d sensor data for human detection. *IEEE Access*, 10:66831–66843, 2022. 1
- [22] Zona Qiu. Flir-align: A repository for flir thermal and rgb image alignment. <https://github.com/zonaqiu/FLIR-align>, 2025. Accessed: 2025-02-20. 4
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2
- [24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2
- [25] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kun-chang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [27] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes

based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 18(3):1000–1011, 2020. [1](#)

- [28] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, page 35–43, New York, NY, USA, 2017. Association for Computing Machinery. [5](#), [6](#), [8](#)
- [29] Helene Torresan, Benoit Turgeon, Clemente Ibarra-Castanedo, Patrick Hebert, and Xavier P Maldague. Advanced surveillance systems: combining video and thermal imagery for pedestrian detection. In *Thermosense XXVI*, pages 506–515. SPIE, 2004. [1](#)
- [30] Zifu Wan, Yuhao Wang, Silong Yong, Pingping Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*, 2024. [1](#), [5](#), [6](#)
- [31] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: simple and efficient design for semantic segmentation with transformers. Red Hook, NY, USA, 2021. Curran Associates Inc. [7](#), [8](#)
- [32] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model, 2023. [3](#)
- [33] Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1224–1235, 2021. [1](#)
- [34] 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*, pages 2223–2232, 2017. [2](#)
- [35] Mehmet Akif Özkanoglu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. [1](#), [2](#), [4](#)