CSC413 Project Report

Mingshi Chi, Connor Lee, Martin Ffrench
Faculty of Applied Science
University of Toronto
{mingshi.chi, conn.lee, martin.ffrench}@mail.utoronto.ca

Abstract

Within the field of autonomous driving, thermal imaging has recently been utilized within a number of ADAS pipelines. However, due to limitations with the development of thermal technology, resolution of thermal images remains relatively small. Furthermore, large resolution thermal cameras have become expensive. Hence, we propose the use of techniques for the generative upsampling of RGB images in the context of ADAS Thermal image upsampling. The main motivation for this approach is to allow users to purchase cheaper low-resolution thermal cameras then up-sample their output for detection tasks. Our upsampling approach compares different methods including: Residual Encoder-Decoder Network and Super Resolution ResNet to create a baseline. Finally, Pix2Pix Conditional Adversarial Network will be utilized in two ways to take the low resolution images as input and generate high resolution output. We augment the UNet variant as well as ResNet to compare to the deterministic SRResNet. We will evaluate which approach produces better results empirically and through the use of a thermal detection model trained on the FLIR Dataset images.

1 Introduction

Within the field autonomous driving there is a large importance in designing multi-redundant systems to ensure the limitations of different sensor modalities are addressed. Thermal imaging has seen great success in aerospace and defence applications and is more recently seeing applications for ADAS through thermal camera vendors such as Telodyne FLIR. The primary motivation for the use of thermal image based perception is for use within anti-glare and night time systems. Thermal images do not suffer from image quality issues with glare and can be utilized to conduct course object detection. The major limitation however is that modern thermal imaging sensors are relatively low resolution. ADAS level thermal cameras have resolutions of 640x512 and can be extremely expensive. The goal for this work is to develop a generative model (or some method) which is able to reliably upscale low-resolution thermal images to the same resolution as that utilized for Autonomous Driving. The overall goal is to develop a model which gains a strong notion of how to upscale low-resolution thermal images of driving scenes (note that driving scenes are a prior here, allowing this model to be more feasible) to make thermal technology more accessible to autonomous driving. The remainder of this document will outline the related work pertaining to general image upscaling and our approach and experiments with our upsampling networks.

2 Related work

There are many works regarding the common task of image up-scaling and super resolution (SR). Often we see the use of auto-encoders to perform SR tasks. Various approaches include deep convolutional auto-encoders [1] and architecture such as RedNet [2], while more generative architecture exists as well in variational auto-decoder [3] and Super Resolution General Adversarial Networks (SRGAN) [4]. The image SR task has only recently been explored in the thermal image space as [5], though G. Batchuluun et al. propose methods using 3-channel thermal images and GANs.

3 Methods and Algorithms

3.1 Baseline Networks:

3.1.1 Residual Encoder-Decoder Network

Image restoration such as de-noising and super-resolution can be done using a deep fully convolutional encoding-decoding architecture that learns mappings from corrupted images to the original ones. The method is termed as "RedNet" - very deep Residual Encoder-Decoder Networks. The convolutional layers are used as feature extractors that eliminate noise and capture abstract features of the image. The deconvolutional layers are used to recover image details that may be previously missing. A skip-layer technique is used to address the vanishing gradient problem and pass details between layers that are

beneficial in recovering the original image [2]. While training the model, the epoch model with the best validation accuracy was saved.

3.1.2 Super Resolution ResNet

SRResNet is a 16 ResNetBlock network which utilizes a MSE and VGG loss to generate up sampled images given target and input pairs. This model does not leverage a discriminator to calculate loss[6]. The implementation used can be found here [1].

3.2 Pix2Pix

Pix2Pix uses architecture based on conditional generative adversarial networks which learn a mapping from an observed input image with added noise to some output. The goal, as with other GANs, is to train a generator to produce output images that are indistinguishable from "real" input images to the adversarially trained discriminator. Within the original Pix2Pix paper, the authors utilze two generator architectures. The first is a ResNet backbone using either 6 or 9 ResNetBlocks. The second is a Unet backbone supporting images with dimensions of multiples of 128 or 256 pixels in width and height (due to channel concatenation skip connections). For our implementation we augment the Unet 128 variant through the addition of a visual attention layer[2] (see Algorithm 1) within the innermost layer of the generator. The Unet variant utilizes output volume concatenation based skip connections such that output of layer i is provided as an additional channel to the input of layer $n - i$. We also explore the ResNet 6 block variant of Pix2Pix in order to provide insight on the effect of the discriminator to an analogous ResNet based network (i.e. SRResNet). This generator only has skip connections between adjacent layers. The discriminator for both layers is a *patch*GAN discriminator which produces an output prediction map (1 for real, 0 for fake) for regions of the input image. We do not preprocess the images for two reasons, first, the entire image provides good relational information for elements within the driving scene. Second, we are still able to train with a batchsize of 12 and 24 for the ResNet and Unet models respectively. For both networks we utilize the Adam Optimizer along side a Least Squares (LS) GAN loss to improve training stability [7].

Algorithm 1 SelfAttentionBlock

X := input feature maps
c := channels of X
$\mathbf{Q}$: = Conv2d(in=c, out=c//8, kernel=1)
$\mathbf{K}$: = Conv2d(in=c, out=c//8, kernel=1)
$\mathbf{V}$: = Conv2d(in=c, out=c, kernel=1)
γ: = additive attention weight
$\mathbf{e} = \mathbf{Q}\mathbf{K}^\top$ ▷ dot product attention
α =Softmax($\mathbf{e}$)
return $\gamma\alpha\mathbf{V} + X$

4 Experiments and Results

4.1 RedNet Image Restoration

	RedNet10	RedNet20	RedNet30
Batch 32	0.022	0.0219	**0.0217**
Batch 64	0.0244	0.0229	0.0228
Batch 128	0.0262	0.0237	0.0237
Batch 256	**0.0305**	0.0245	N/A

Table 1: Validation results

3 different models were trained for 30 epochs for the RedNet architecture on the FLIR dataset: 10 layers, 20 layers, and 30 layers [3]. The model with 30 layers preformed the best with a final training loss of 0.0253 and a validation loss of 0.0217.

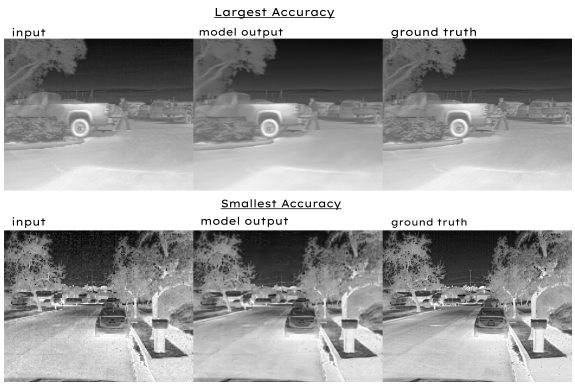


Figure 1: RedNet30 highest(99.9997) and lowest(99.9966) accuracy with MSE loss. Input, model output, and ground truth respectively.

Figure 1 shows the highest and lowest accuracy output by RedNet30 during testing. As seen in the figures, the model's up sampling results in a smoothed and fuzzy version of the input. This is due to using Mean Squared Error (MSE) loss:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{F}(X_i;\Theta) - Y_i\|_F^2 \qquad (1)$$

Given a collection of N training sample pairs X_i, Y_i, where X_i is a corrupted image and Y_i is the ground truth. Using this loss penalizes larger errors of a pixel from its ground truth which results in the model favouring smoother and blurrier pictures. Further examples are available in A.1.1.

[1]This implementation expands a 120x160 image to 512x640 https://github.com/Martin0xFF/pytorch-SRResNet-thermal
[2]trainable implementation can be found at https://github.com/Martin0xFF/pytorch-CycleGAN-and-pix2pix-thermal
[3]Training and testing code can be found at https://github.com/mingshi1214/RedNet

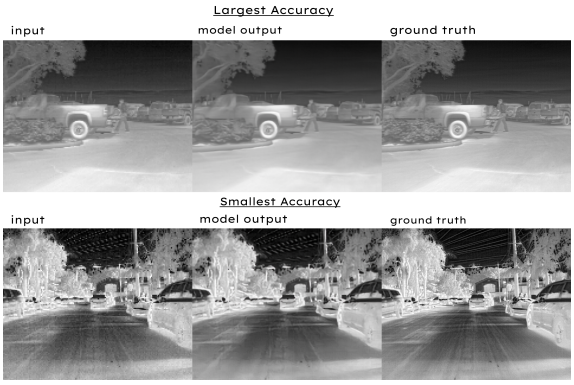


Figure 2: RedNet30 highest (99.9997) and lowest (99.9965) accuracy with L1 loss. Input, model output, and ground truth respectively.

When training with L1 loss instead, the images are less smooth and more geometric as can be seen in Figure 2 but have a higher loss in training and testing. Please refer to A.2.1 and A.2.2 for training losses. Viewing the results qualitatively, the MSE loss training results yielded pictures that more resembles the ground truth images. Additional images can be viewed at A.1.2.

4.2 Super Resolution ResNet

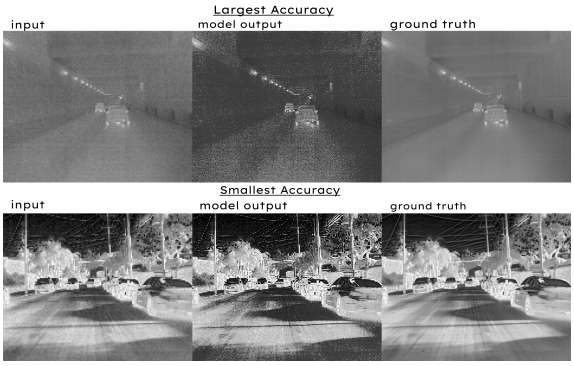


Figure 3: Pre-trained SRResNet highest (99.9986) and lowest (99.9850) accuracy. Input, model output, and ground truth respectively.

Initial experiments were made with the pre-trained weights from ImageNet [8]. Figure 3 show the best and worst test outputs using the pre-trained weights (99.9986 and 99.9850 respectively). Using the pre-trained weights produces more textures than in the ground truth image. This may be due to the difference in the training data used. ImageNet provides RGB images whereas FLIR is greyscale. The down-sampling techniques chosen were also different. The pre-trained weights were trained using inputs that were down-sampled using bi-cubic interpolation whereas the down-sampled imaged from the FLIR dataset used nearest neighbour down-sampling. Additional images can be viewed in A.1.3.

After training the SRResNet on domain specific data, an upscaled version of the test set (from test inputs) was produced. The mean and standard deviation between the ground truth test set and the upsampled test input can be seen within Table 2 along side other models. Notably, we see that the RedNet30 produces the smallest MSE followed by SRResNet and Bi-cubic. With this metric the performance between RedNet30 and SRResNet appears similar with RedNet doing slightly better. The difference in performance between Bi-cubic and the generated methods is clear however. Interestingly, the standard deviation of the Pix2Pix Unet + Attention (P2P U+A) model is much lower than all other models. A large MSE for the generated models is logical; during training an L1 loss is used and Pixel difference MSE isn't used directly.

Model output can be seen within Figure 4. The smallest loss result in Figure 4 illustrates a better SRResNet reconstruction. The scene for this case is relatively simple (few cars or other higher frequency image features) and the model output in this case appears to be blurred. The performance of the model here is very similar to RedNet30. In the case of largest loss, the scene is a bit more complicated particularly due to the sudden harsh variation in temperature of the road and high frequency features in the distance (the tree and foliage). These results differ from the results presented within Figure 3 as it seems that RGB trained SRResNet dramatically changes the contrast of the image in an attempt to improve the output quality.

Model	MSE	stdev
RedNet30	0.0327	0.0105
SRResNet	0.0334	0.0103
Bi-cubic	0.0380	0.0111
P2P U+A	0.0542	0.0062
P2P 6R	0.0668	0.0143

Table 2: Average Per Image Pixel MSE between Model output and ground truth

Figure 4: SRResNet samples with largest loss (0.026) and smallest loss (0.0009). Input, model output, and ground truth respectively.

4.3 Pix2Pix 6 Layer ResNet

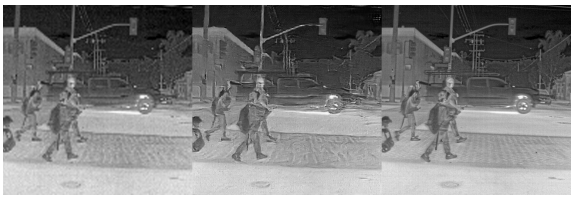


Figure 5: Pix2Pix ResNet6 - Notice the brush stroke like grouping of pixels within the model output

Within Figure 5, we see a quintessential sample of the model output. Here, it seems that the ResNet based generator tends to group collections of pixels together. This can be clearly seen within the contour of the car door within the intersection. It seems that the model has learned to produce a paint stroke like texture on output images. This may be due to the grouped nature of the *patch*GAN discriminator. Since the skip connections of the ResNet are strictly to adjacent layers, there will be some necessary combination of features producing the output.

4.4 Pix2Pix Unet+Attention

Figure 6: Pix2Pix Unet+Attention - Here we see better reconstruction particularly of the car, however the pedestrian's bicycle is lost.

Within Figure 6, we see that the image reconstruction is better than that of the ResNet based generator. In particular, the Unet+Attention generator does a better job of reconstructing the texture of the car within the scene, despite its distance to the thermal camera. However, it also seems that the pedestrian's bike was removed by the network, as it did not provide a strong enough signal and was presumed to be noise. There are other cases when these GAN based models either add features to a scene or remove features to a scene.

4.5 Downstream performance

Trained on	M mAP	L mAP	X mAP
P2P U+A	**0.211**	**0.217**	0.212
RedNet	0.210	0.214	**0.213**
Bi-cubic	0.182	0.185	0.182

Table 3: Downstream performance of upsampling modules - all classes mAP@0.5:0.95

Table 3 provides a summary of three different YOLO Models (M, L, X) trained on Pix2Pix Unet+Attention (P2P U+A), RedNet, or Bi-cubic interpolated samples. From the Table, we see that the use of a learned method provides immediate improvements in accuracy, while the use of a GAN method appears to provide marginally better performance.

5 Conclusion

We can see empirically the base models perform well in the super resolution recovery task. It is evident that information is lost in downscaling, and fine detail such as lettering and road markings are not recovered. Baseline performance was determined using various loss schemes on RedNet as well as SRResNet. RedNet with MSE loss produces smoothed fuzzy results with some retention of pixelization, whereas the L1 loss models result in less smooth, geometric images. Qualitatively, our SRResNet model outputs are worse than the results from RedNet, likely due to the pretraining of SRResNet being performed on visible light imaging (rather than thermal as present in the FLIR dataset). The discrepancy in performance highlights a novel characteristic of the thermal upscaling problem as we cannot simply apply visible light RGB trained models to the task. The generative models, Pix2Pix 6 Layer ResNet and Pix2Pix UNet with attention produce reconstructions that more closely resemble the high frequency ground truth images. The detail in both Pix2Pix outputs appear visually more similar to the ground truth than the RedNet outputs. The Pix2Pix UNet+Attention performs better than the ResNet based model likely due to the UNet architecture performing well in image segmentation. In the object detection task, it is evident that the generative method Pix2Pix (UNet+Attention) produces the best results, though only marginally superior to the other learned method implemented in RedNet. An important improvement to evaluating the quality and performance of the model outputs would be to derive a different quantifiable method to compare models than using MSE. As can be seen throughout Section 4. Experiments and Results, there is very little difference between the MSE values of the models explored. This is due to having very minimal difference between the pixel values of the grey-scale output image and the grey-scale ground truth image. For future exploration, Pix2Pix UNet+Attention provides promising performance and as such it would be interesting to see the results of additional attention blocks. Other methods that may be expanded upon in the future for the thermal image super resolution task include Variational Autodecoder reconstruction and CycleGAN super resolution as originally proposed and prototyped with.

References

[1] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, "Coupled deep autoencoder for single image super-resolution," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 27–37, 2017. DOI: 10.1109/TCYB.2015.2501373.

[2] X.-J. Mao, C. Shen, and Y. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *NIPS*, 2016.

[3] A. Zadeh, Y. C. Lim, P. P. Liang, and L. Morency, "Variational auto-decoder," *CoRR*, vol. abs/1903.00840, 2019. arXiv: 1903.00840. [Online]. Available: http://arxiv.org/abs/1903.00840.

[4] C. Ledig, L. Theis, F. Huszar, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016. arXiv: 1609.04802. [Online]. Available: http://arxiv.org/abs/1609.04802.

[5] G. Batchuluun, Y. W. Lee, D. T. Nguyen, T. D. Pham, and K. R. Park, "Thermal image reconstruction using deep learning," *IEEE Access*, vol. 8, pp. 126 839–126 858, 2020. DOI: 10.1109/ACCESS.2020.3007896.

[6] C. Ledig, L. Theis, F. Huszar, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016. arXiv: 1609.04802. [Online]. Available: http://arxiv.org/abs/1609.04802.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, 2016. DOI: 10.48550/ARXIV.1611.07004. [Online]. Available: https://arxiv.org/abs/1611.07004.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

6 Contributions

Briefly, each member contributed in the following ways:

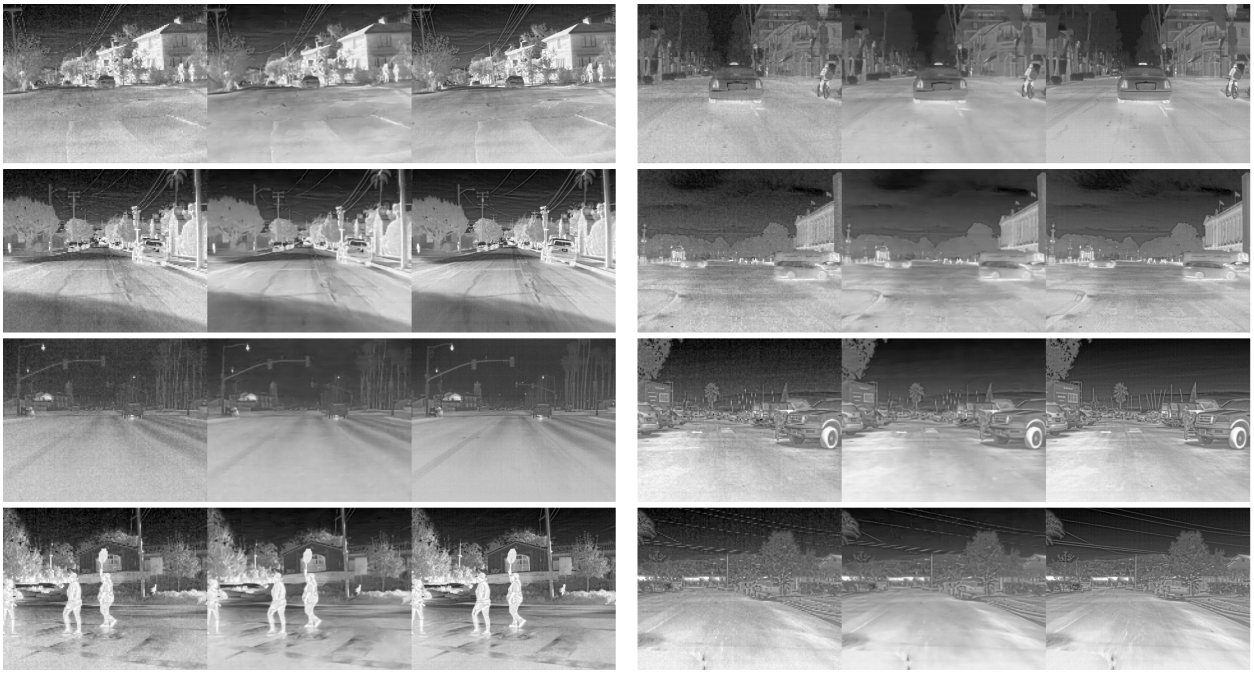
- Topic Exploration: Mingshi, Connor, Martin
- Design of models: Mingshi, Connor, Martin
- Training: Mingshi, Connor, Martin
- Data visualization: Mingshi, Martin
- Downstream Validation: Martin
- Report Writing: Mingshi, Connor, Martin
  - Abstract and Introduction: Connor, Martin
  - Related Works: Mingshi, Connor
  - Method/Algorithms: Mingshi, Connor, Martin
  - Experiment and Results: Mingshi, Connor, Martin
  - Conclusion: Mingshi, Connor, Martin
  - References: Mingshi, Connor, Martin
  - Appendix: Mingshi, Connor, Martin

A Appendix

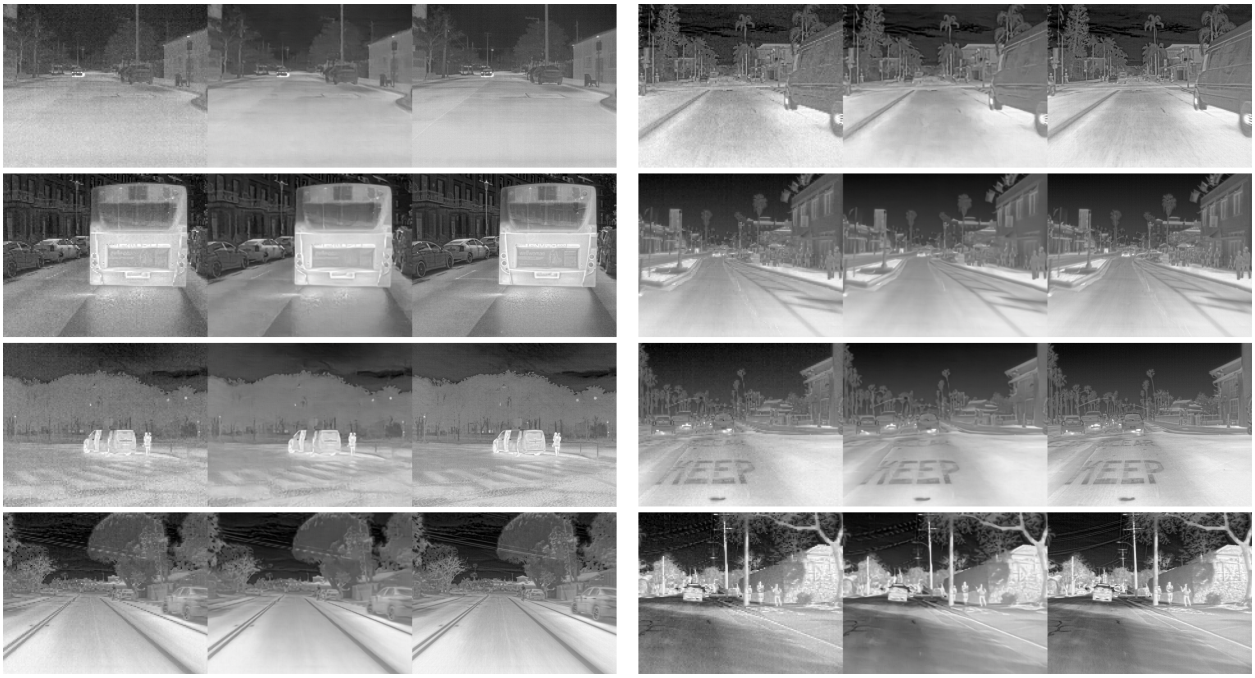
A.1 Samples of Baseline Outputs

To view the full images, please navigate to this folder:
https://drive.google.com/drive/folders/1tZKerfCfPZqfDa9mEiXuqf5Ec_d8vmyH?usp=sharing

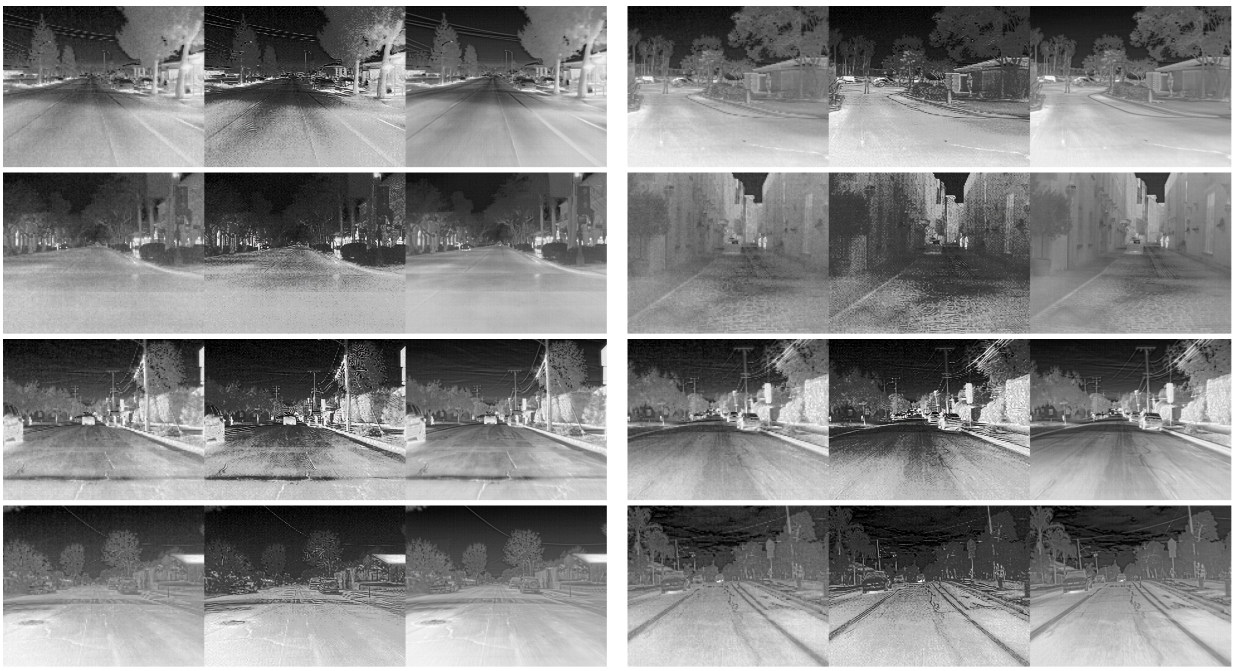
A.1.1 MSE training loss model outputs (RedNet30 for 32 batch size)



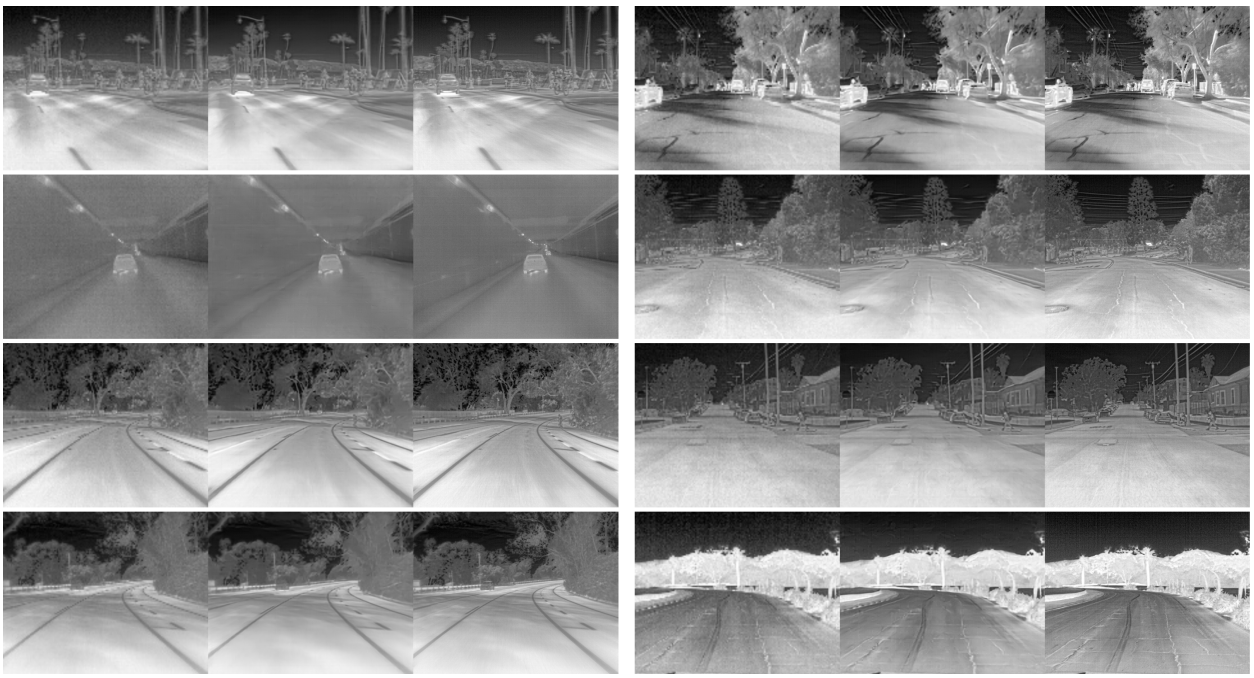
A.1.2 L1 training loss model outputs (RedNet30 for 32 batch size)



A.1.3 Pretrained weights from SRResNet results on FLIR Dataset



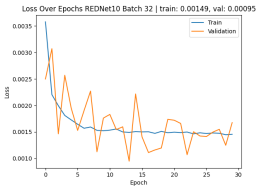
A.1.4 Trained weights of SRResNet on the FLIR dataset results

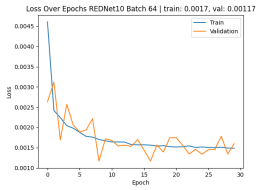


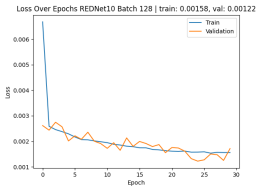
A.2 Loss Plots

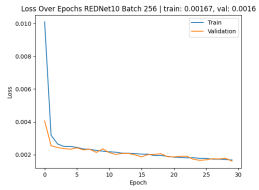
A.2.1 MSE Training loss for RedNet. Loss over epochs

RedNet10:

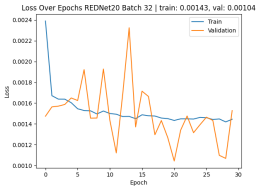


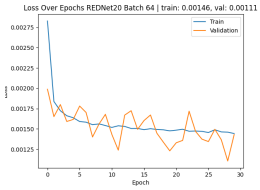


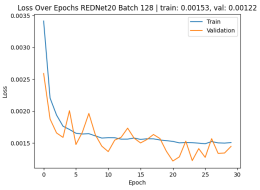


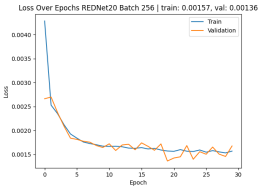


RedNet20:

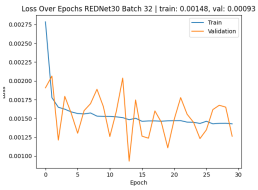


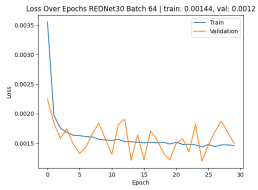


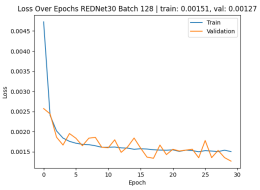




RedNet30:

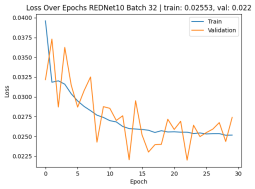


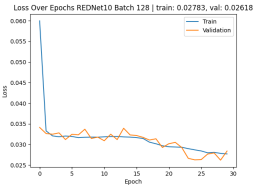


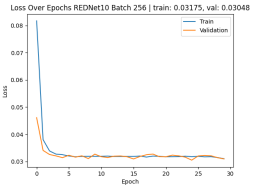


A.2.2 L1 training loss for RedNet. Loss over epochs

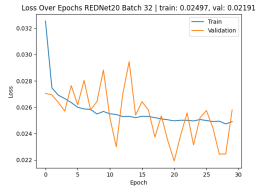
RedNet10:

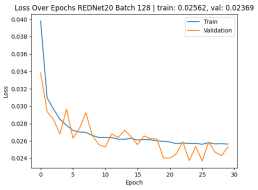


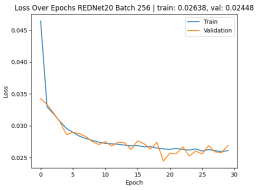




RedNet20:





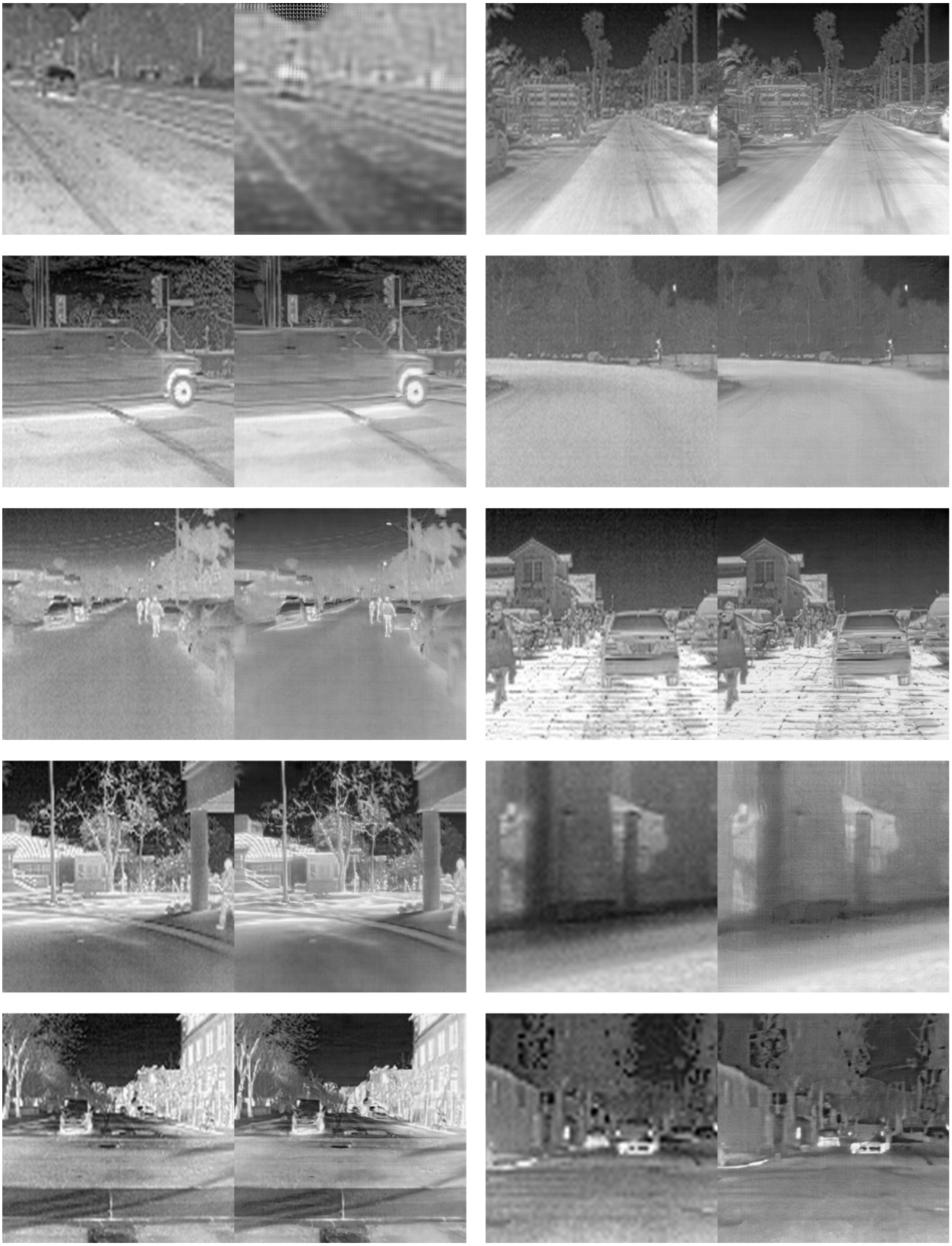


RedNet30:

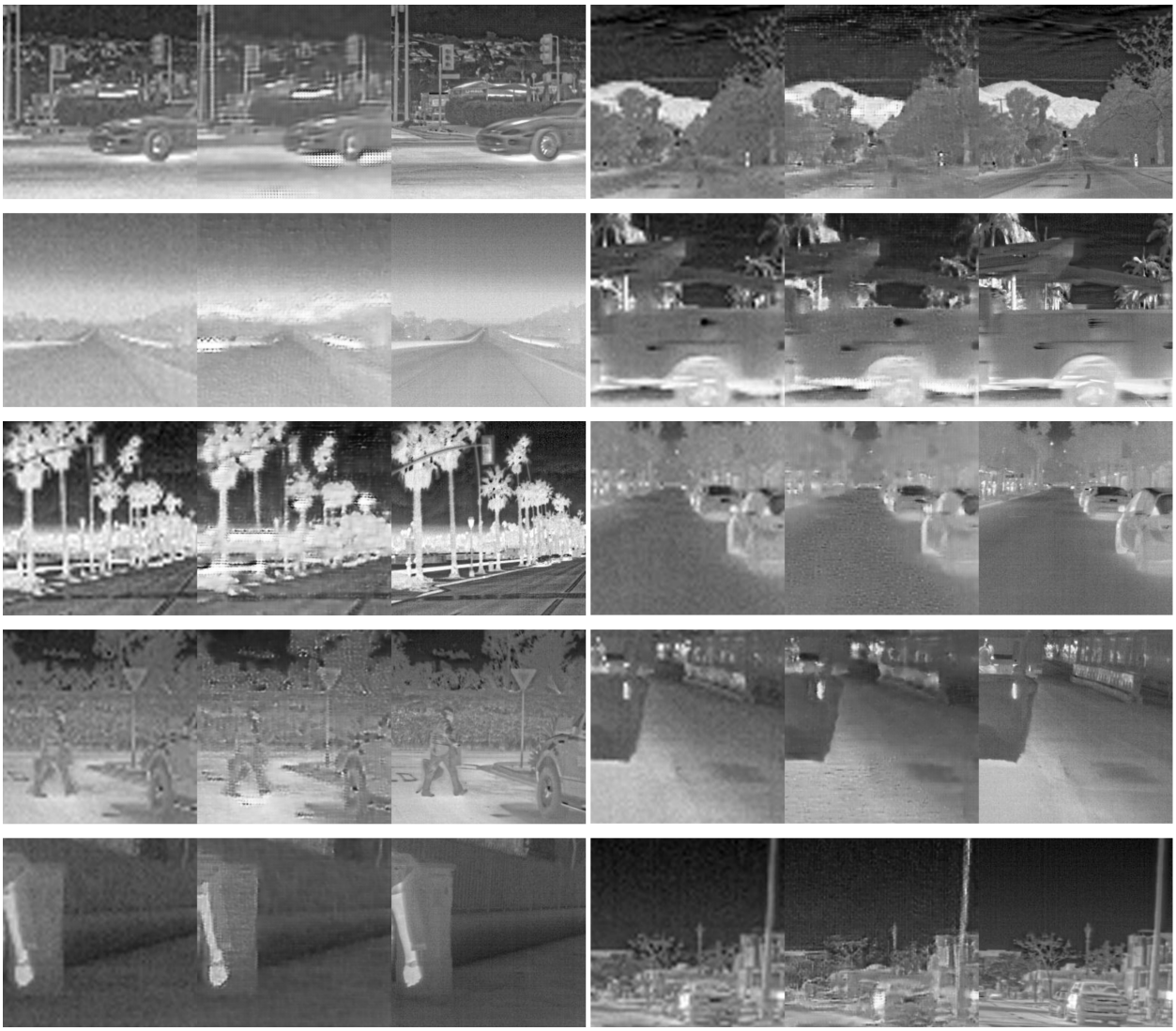
A.3 Samples of Generative Model Experiment Outputs

To view the full images, please navigate to this folder:
https://drive.google.com/drive/folders/1JTBxRf-Oi9eLt8CXEhOQ09R2LAUmVbXq?usp=sharing

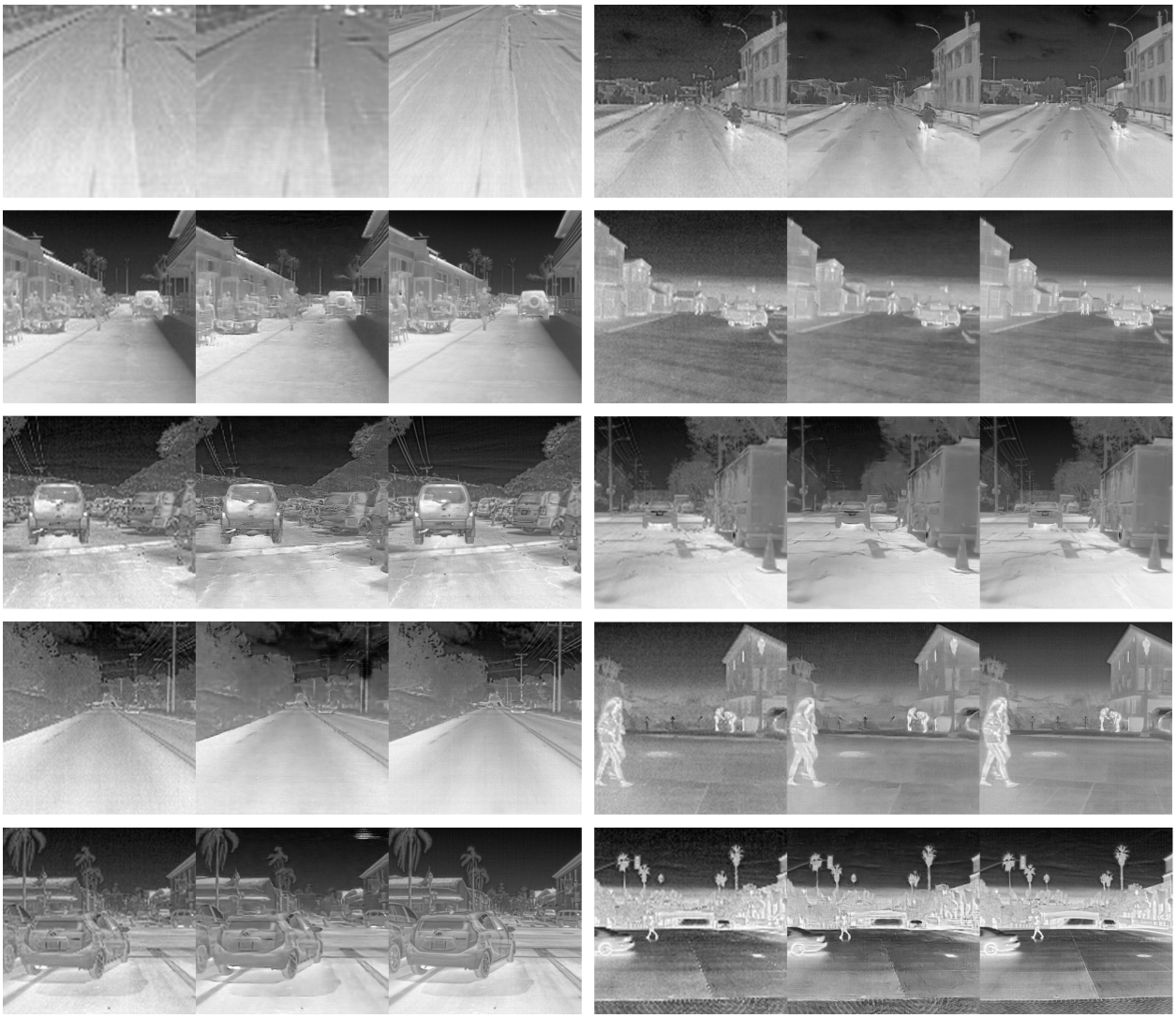
A.3.1 Cycle GAN RESNet 9 examples over epochs



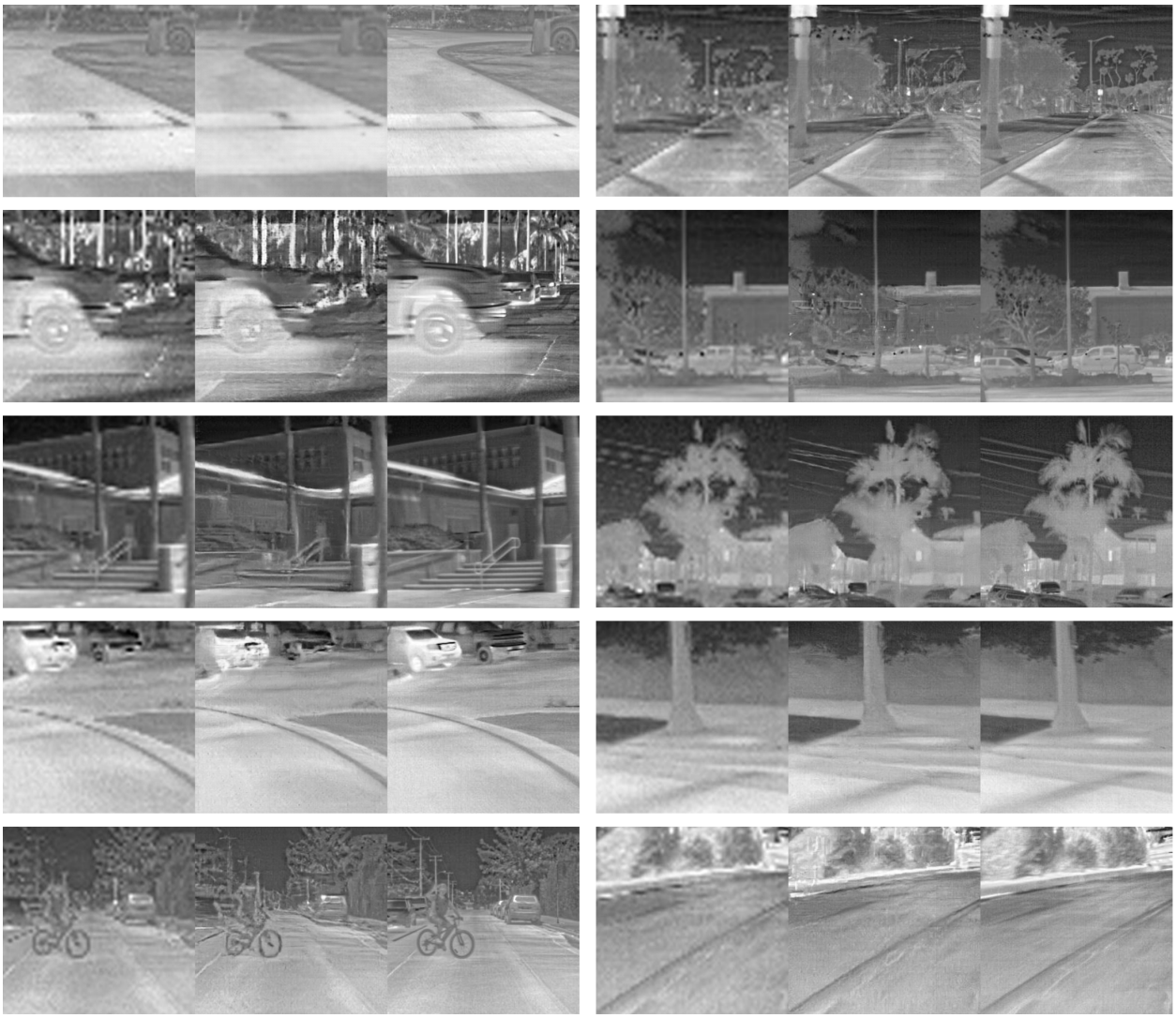
A.3.2 Pix2pix RESNet with Attention examples over epochs



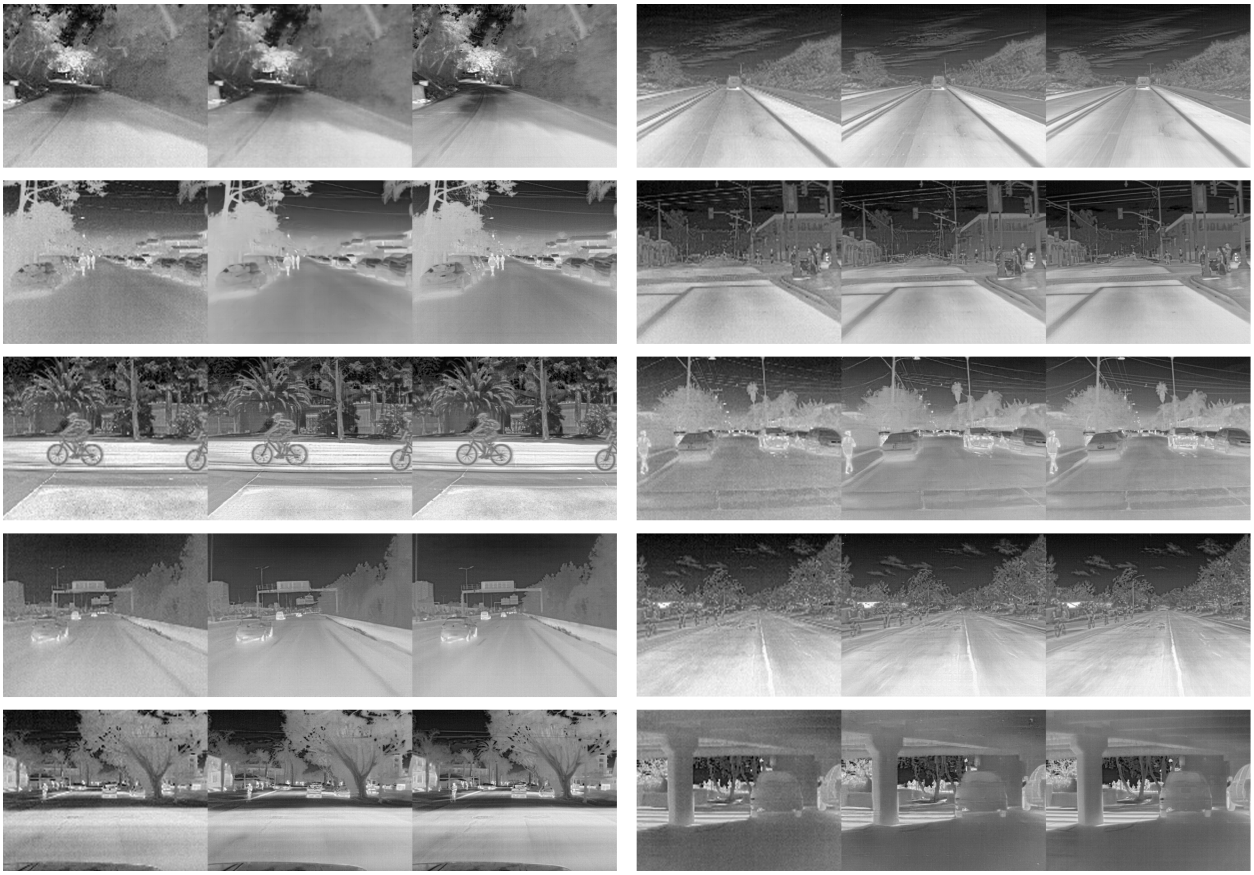
A.3.3 Pix2pix RESNet without Attention examples over epochs



A.3.4 Pix2pix Unet without Attention examples over epochs



A.3.5 Pix2pix Unet with Attention using Least Squares Generator Loss examples over epochs (128 pixels)



A.3.6 Pix2pix Unet without Attention using Least Squares Generator Loss examples over epochs (256 pixels)