# Exploring the usage of diffusion models for thermal image super-resolution: a generic, uncertainty-aware approach for guided and non-guided schemes

Carlos Cortés-Mendez, Jean-Bernard Hayet
Centro de Investigación en Matemáticas, AC
Guanajuato, Gto., México
{carlos.mendez,jbhayet}@cimat.mx

## Abstract

*In this paper, we explore the use of diffusion models for the thermal imaging super-resolution problem, with the PBVS workshop Thermal Image Super-Resolution Challenge (TISR) as an application context. In addition of adapting the recently proposed Resshift diffusion approach to the problem of SR for thermal imaging, we show how this diffusion model can be also used nearly effortless in both the guided and non-guided TISR tasks, where the guidance comes from the visible image. More crucially, we show that a natural and often under-leveraged output from this diffusion approach is the quantification of the aleatoric uncertainty on the resulting HR prediction. By using this property, we empirically show that per-pixel standard deviation of the samples produced by a super-resolution diffusion model are a good estimator for the per-pixel absolute error in scenarios where the HR ground truth is not available.*

## 1. Introduction

In image processing, Super Resolution (SR) is one of the most common tasks, that consists in producing a high-resolution (HR) image starting from a low-resolution (LR) one. It has numerous applications, ranging from biomedical imaging [15] to surveillance [10] or remote sensing [18], to give a few examples. By definition, SR is difficult to solve as it is an inverse, ill-posed problem, where the degradation operator for the forward problem typically involves a down-sampling step and the addition of noise.

Solutions to SR may involve the use of more sensing modalities or more image acquisitions from the same modality. However, in some cases, getting these additional data is impossible and the problem has to be solved with only one LR image as an input. Because of the strongly ill-posed nature of the SR problem in this case, solving it often implies the construction of strong probabilistic priors on the potential HR solutions, conditioned to the LR images. These priors may be built from datasets of examples of LR/HR pairs taken from the domain of interest. This explains why most of the recent literature in SR involves machine learning and in particular deep learning models, which have shown they are capable of learning complex mappings between the LR and HR image spaces. In this setup, SR can be cast as a regression problem.

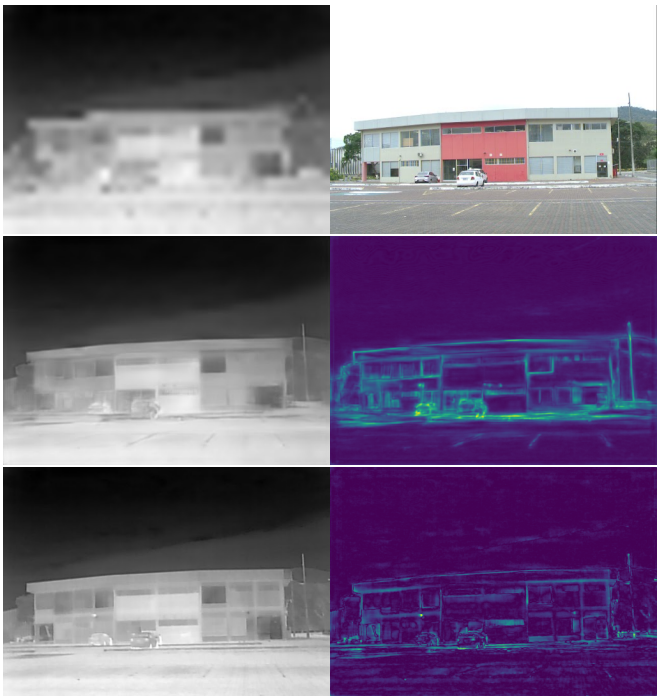

Figure 1. Our proposed diffusion-based approach, in the case of $\times 16$ SR with a visible image as a guide (Track 2 of TISR Challenge). On the first row, the model inputs are the LR image and the visible domain one; on the second row, the outputs are the HR image (left) and an estimate of the uncertain areas of the predicted HR, based on the generation of 32 samples from the diffusion model; on the third row, we depict the ground truth HR (left) and the errors between the estimate and GT (right).

On the other hand, in the last years, generative modeling has been used as a tool of choice to solve such inverse problems (e.g. image colorization, image de-compression, and SR). The main motivation behind it is that generative models (and more particularly, in this case, conditional generative models) provide a natural way to approximate the *full distribution* of potential solutions, instead of producing only the likeliest one. For inverse problems with a high level of degradation such as SR, i.e. with a many-to-one forward process, this may be important, since it provides a way to quantify the resulting *uncertainties* in the solution.

Although SR has been studied in many domains, mostly with natural images from the visible spectrum, little work has been done on proposing and evaluating algorithms for thermal imagery. Thermal images offer a wide range of advantages over visible images in particular areas such as inspection, security, rescue, for example. For a few years now, the PBVS workshop incentivizes the research on SR algorithms for thermal imaging through the organization of the Thermal Image Super-Resolution Challenge (TISR) [11].

In the work we present here, we explore the use of a particular kind of generative models for this specific problem of thermal image super-resolution and this specific TISR challenge, namely diffusion models [7, 12]. Diffusion models are having a great impact for many computer vision tasks involving the generation of images. We leverage a recently proposed diffusion architecture [22] that we adapt to the context of thermal imaging and we show, as illustrated through an example from Track 2, in Fig. 1, (1) that it can be used in a very versatile way, both for the non-guided and guided tasks (1 and 2) of the TISR challenge and (2) that our diffusion model output provides an estimate of the uncertainty on the super-resolution image, which can be useful for applications making use of the SR algorithm.

In summary, our contributions are the following:

- We adapt the recently proposed Resshift diffusion approach to the problem of SR for thermal imaging;
- We show how this diffusion model can be used nearly effortlessly in the guided and non-guided tasks of the PBVS Thermal Image Super-Resolution Challenge (TISR);
- We finally show that a natural output from this diffusion approach is the quantification of the aleatoric uncertainty on the resulting HR output.

## 2. Related work

As for many inverse problems, a key to the resolution of SR problems is the introduction of strong priors on the possible resulting HR images and one traditional way to do it is to cast the problem as a regularized optimization problem. The regularization scheme may involve total variation-based objective functions [1], low-rank representations [14], sparse representations [6], to give a few examples.

Starting from the mid 2010s, deep convolutional neural networks have been used to learn end-to-end mapping from low resolution images to high resolution images. As an example, [5] introduced SRCNN, one of the first fully convolutional aarchitecture applied to SR. In [9], the authors propose an architecture coined as deep Laplacian Pyramid Network to reconstruct iteratively the residuals of high-resolution images, leveraging the pyramidal structure and transposed convolutions for generating the consecutive upsampled versions of the HR images.

Generative modeling have been used in the SR area relatively early with the advent of deep generative modeling methods. One of the main motivation has been its use as a deep prior. For example, the GLEAN method [2] uses a pre-trained GAN as a deep prior for an encoder-bank-decoder architecture. In [21], an explicit kernel prior is used on a patch basis in combination to a Monte-Carlo EM algorithm to determine a Maximum A Posteriori solution.

As the latest successful instance of generative models, diffusion techniques [7, 12] have been used in the context of SR. In [13], the DDIM diffusion scheme is used with a few technical modifications to produce the SR3 architecture. To generate HR images, the process starts with pure Gaussian noise and iteratively refines the image with a U-Net trained on denoising at various noise levels, and conditioned on the LR image. In [3], Inversion by Direct Iteration (InDI) is proposed as a general image restoration framework, that shares a lot of properties with diffusion models.

One of the main drawbacks of diffusion models such as DDIM, as used in [13], is that inference times may be quite heavy when using the standard diffusion processes, involving typically hundreds of denoising steps. In [22], a work upon which we based ours, the authors propose an interesting way to recast the diffusion model, with the forward noising process starting at the HR image and leading to a noisy version of the LR image; this way, the number of denoising steps is heavily reduced, by an order of magnitude. We build upon the former model to propose a solution to the thermal image super resolution problem, show that it can easily be modified to include guidance from other modalities, and leverage the resulting samples to produce an estimate of the aleatoric uncertainty.

## 3. Proposed approach

Our proposed method is based on the ResShift diffusion approach recently proposed in [22]. As commented above, this approach allows to accelerate the inference (forward) process by an order of magnitude by refactoring the forward and backward processes. The results from this super-resolution diffusion model is then further refined using a small UNet to work on the details. We empirically find that our approach has the following advantages:

- With a relatively cheap computational overhead, a single model can be pre-trained and later fine-tuned to work in a

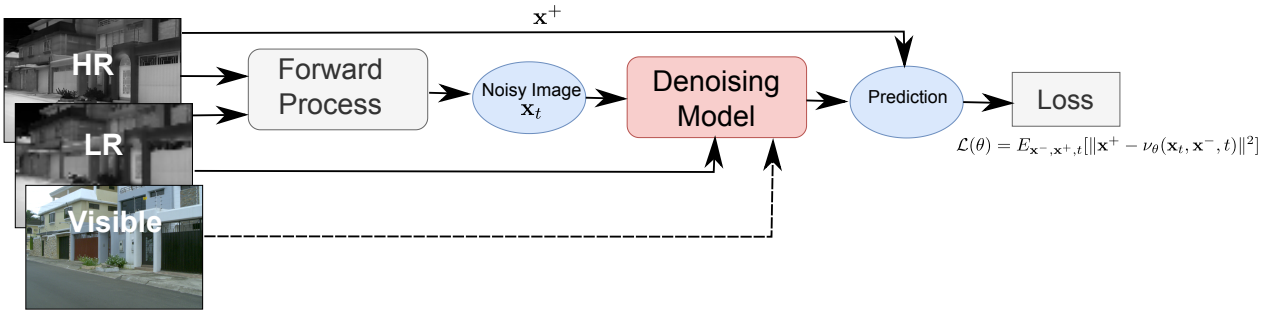


Figure 2. Our proposed architecture, based on ResShift, which is finetuned om the TISR Challenge dataset. The basic diffusion model uses both upsampled LR and HR images in the forward process and the denoiser. With the diffusion scheme it is possible to use the visible image as condition for the denoising model instead of the LR image thus keeping the same architecture for both tasks.

specific domain (such as thermal images) or at a specific scale factor ($\times 8$ or $\times 16$).
- Its two-model approach allows each model to focus on the overall image (diffusion model) and the details (UNet).
- It can be very easily used either in a non-guided or in a guided SR scheme (i.e. by using an image from the visible domain).

### 3.1. ResShift Diffusion Model

We recall here the principles developed in ResShift and the reader is invited to consult [22] for more details. Let $\mathbf{x}^+, \mathbf{x}^-$ a pair of HR and LR images from the training dataset. The idea of ResShift is to re-state the diffusion forward process so that, instead of starting from the HR image $\mathbf{x}^+$ and ending up with pure Gaussian noise, we end it up with a noisy version of $\mathbf{x}^-$. The reverse process is adapted in consequence. The forward process can be written as the following Markov chain, for $t > 0$ and a shifting sequence $\{\eta_t\}_{t=1}^T$ and $\alpha_t \triangleq \eta_t - \eta_{t-1}$ such that $\eta_1 \approx 0$ and $\eta_T \approx 1$:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}^-) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1} + \alpha_t(\mathbf{x}^- - \mathbf{x}^+), \kappa^2\alpha_t\mathbf{I}) \quad (1)$$

with $\mathbf{x}_0 \triangleq \mathbf{x}^+$. It results in the closed-form marginal

$$p(\mathbf{x}_t|\mathbf{x}^+, \mathbf{x}^-) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}^+ + \eta_t(\mathbf{x}^- - \mathbf{x}^+), \kappa^2\eta_t\mathbf{I}) \quad (2)$$

with the aforementioned effect to end up with a noisy version of $\mathbf{x}^-$ ($\mathbf{x}_T$ is centered on $\mathbf{x}^-$ with variance $\kappa^2$). For the reverse process, we use a variational approximation of $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}^-)$ as a Gaussian with mean a function of $\mathbf{x}_t, \mathbf{x}^-, t$ and variance $\kappa^2\frac{\eta_{t-1}}{\eta_t}\alpha_t\mathbf{I}$. It results convenient to reparameterize this mean function as

$$\frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t}\nu_\theta(\mathbf{x}_t, \mathbf{x}^-, t) \quad (3)$$

and train the network $\nu_\theta$ under the loss

$$\mathcal{L}(\theta) = E_{\mathbf{x}^-,\mathbf{x}^+,t}[\|\mathbf{x}^+ - \nu_\theta(\mathbf{x}_t, \mathbf{x}^-, t)\|^2] \quad (4)$$

It is important to note that the denoising model $\nu_\theta$ takes both $\mathbf{x}_t$ and $\mathbf{x}^-$ as input, which are concatenated in the channel dimension, just before the first convolution. Doing this implies to use an upsampled version of the LR image $\mathbf{x}^-$, which we obtain by bicubic interpolation.

During inference, to reverse the forward process, refined versions of $\mathbf{x}_t$ are computed by applying Eq. 3, starting from an upscaled and noisy version of $\mathbf{x}^-$ instead of just noise as in DDPM [7]:

$$\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{x}^-, \kappa^2\mathbf{I}). \quad (5)$$

The intuition behind this is that the forward process with very few steps tends to keep the general shapes of the image while removing the details. This has the effect that the noisy versions of $\mathbf{x}^+$ and $\mathbf{x}^-$ are similar after $t$ steps and in turn allows to spare a lot of diffusion steps. As such, the ResShift scheme only uses $T = 15$ diffusion steps, instead of the usual hundreds, sparing a lot of computational time during inference. In this work, we train the denoising model $\nu_\theta$ under the ResShift scheme on the ImageNet-1k dataset [4] under the same noise scheduling parameters (in particular, $T = 15$).

Finally, note that this diffusion model is trained and used within a latent space produced by an autoencoder with a pair of encoder/decoder $\epsilon_\psi/\delta_\psi$ with parameters $\psi$. This means that the equations above should be rewritten in terms of the latent variable $\mathbf{z}$, related to the image through $\mathbf{z} = \epsilon_\psi(\mathbf{x})$ (encoder) and $\mathbf{x} = \delta_\psi(\mathbf{y})$ (decoder)

$$\mathbf{z}_{t-1} \sim \mathcal{N}(\frac{\eta_{t-1}}{\eta_t}\mathbf{z}_t + \frac{\alpha_t}{\eta_t}\nu_\theta(\mathbf{z}_t, \mathbf{z}^-, t), \kappa^2\frac{\eta_{t-1}}{\eta_t}\alpha_t\mathbf{I}). \quad (6)$$

In this work, we have employed a VQVAE [16] that has also been trained on the ImageNet-1k first and then frozen for the rest of the process.

### 3.2. Finetuning on the TISR Thermal Dataset

We use the checkpoint of the model that was pre-trained as described above, on the ImageNet-1k dataset, and fine-tune it in the thermal images dataset [11].

For this purpose, we fine-tune 3 separated instances of the base model trained in ImageNet-1k:

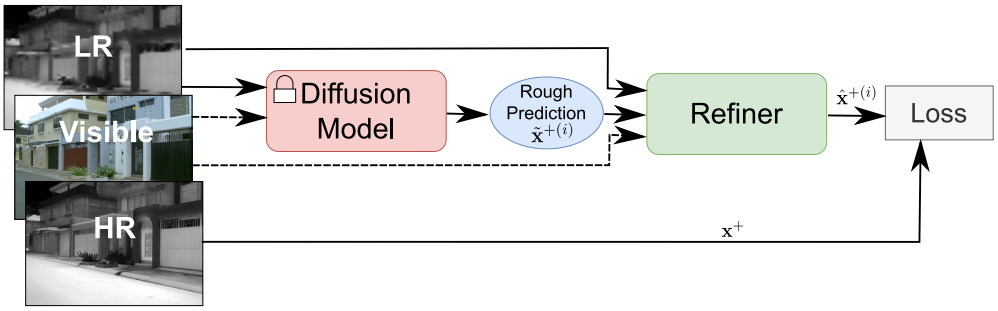


Figure 3. The proposed Refiner UNet, in which the diffusion model is frozen. During training

1. For Track 1 (unguided SR), we fine-tune the model using the LR image $\mathbf{x}^-$ as both the mean of the distribution of $\mathbf{x}_T$ and the condition for the denoising model (concatenated at the start), by using a $\times 8$ scale factor.
2. For Track 2 (SR guided through visible images), we fine-tune the model using the LR image $\mathbf{x}^-$ as the mean of the distribution of $\mathbf{x}_T$ while the visible image $\mathbf{x}^v$ is used as condition for the model (concatenated at the start).
3. We repeat the same process for the $\times 16$ part of Track 2.

In any case, the estimate produced by the diffusion model is denoted as $\tilde{\mathbf{x}}^+$. Note that, since the denoising model is fully convolutional, no additional adjustments are needed for the architecture to work in all the three cases. These principles are summed up in Figure 2.

### 3.3. Refiner network

After fine-tuning, we have observed that we obtain competitive results in the perceptual metrics (*AlexNet, SqueezeNet*) but rather poor performance on the more standard metrics such as PSNR or SSIM. It has been observed that diffusion models tend to hallucinate [22] and this manifests into the model generating details and textures that seem *real* (similar to those in the training dataset) but are not the exact same in the original image, hence the discrepancy between perceptual and traditional similarity measures. Furthermore, since diffusion models are inherently stochastic, we have leveraged this stochasticity to measure the variance in the reconstructed image by using multiple samples of images produced by the model. We have empirically found that the areas in which the model has the most uncertainty are indeed, the areas with the finest details. We further discuss the details of this uncertainty estimation in the next section.

As a proposed solution to deal with this imprecision for regions with finer details, we introduce a small Unet model (7M parameters) with only 2 skip connections, with the purpose of predicting the details missing in the prior result of the diffusion model. We refer to this model as the *Refiner* $\rho_\phi$ and we train it to predict the HR image using LR image and the prior result from the diffusion model as inputs, to produce an improved estimate $\hat{\mathbf{x}}^+$ of the HR image

$$\hat{\mathbf{x}}^+ = \rho_\phi(\tilde{\mathbf{x}}^+, \mathbf{x}^-). \tag{7}$$

Again, 3 different refiners have been trained for the three modalities of the TISR challenge:

- For Track 1, we have trained 1 refiner.
- For Track 2, we have trained 2 refiners for $\times 8$ and $\times 16$ scale factors, respectively. Each one of these receives an additional input in the form of the visible image.

A diagram of the Refiner architecture can be found in Figure 3. Note that, in the case the visible image is available, we use it too.

### 3.4. Producing aleatoric uncertainty estimates

As commented above, we take advantage of the stochastic nature of the proposed diffusion model to produce several HR samples $\tilde{\mathbf{x}}^{+(i)}$ of the predictive distribution $p(\mathbf{x}^+|\mathbf{x}^-)$, for $i = 1, ..., N$. Let us stress that this sample generation process avoids computational overheads by using mini-batches. Then, we compute a variance map based on the $N$ estimates values at each pixel. We will see in Section 4 that these variance maps do reflect the aleatoric uncertainties (i.e. those related to the data themselves) over the HR estimate $\tilde{\mathbf{x}}^+$, and correlates well with the observed errors. The same variances can be also estimated after the Refiner module on refined samples $\hat{\mathbf{x}}^{+(i)}$, as this network does not add more stochasticity. In fact, being able to produce these estimates of the variance is one of the most compelling reasons why we would use a generative model such as a diffusion model in the first place.

## 4. Experimental results

### 4.1. Implementation details

We have first trained the UNET denoiser used for the ResShift model from Section 3.1 on the ImageNet-1k dataset with a scale factor of x4 and image size of 256 for 70k iterations using Adam [8] optimizer with a learning rate of $5 \times 10^{-5}$ and using a batch size of 32. This UNet denoiser is composed of 8 residual layers in the encoder and 8 in the

Figure 4. Results on Track 1 validation set. From left to right, LR, prediction, standard deviation among multiple predictions.

Figure 5. Results on Track 2 ($\times 8$) validation set. From left to right, LR, prediction, standard deviation among multiple predictions.

decoder with a skip connection between each and an attention layer [17] in between each residual layer, reaching a total of 122 million parameters. For further details of the UNet denoiser refer to our code [1].

This model reached a PSNR score of 23.79, a SSIM [20] score of 0.683 and an LPIPS (VGG) [23] score of 0.2437 seconds over the ImageNet test dataset which proves to be competitive with respect to the state of the art [19, 21, 22]

We have then proceeded to finetune the UNet denoiser using the thermal images as described in section 3.2. Each model has been trained for 40k iterations with a learning rate of $5 \times 10^{-6}$ and using a batch size 32.

Finally, the refiner UNet presented in Section 3.3 has been trained, in its three sub-variants, for 10k using Adam [8] optimizer, with a learning rate of $1 \times 10^{-4}$ and using a batch size of 16. This UNet refiner is much smaller, composed of only 2 total residual layers in the encoder and 2 in the decoder with a skip connection each and no attention layers. This model is about only 7M parameters in size.

[1] A link to the code repository will be included in the final version, in case of acceptance.

### 4.2. Results on TISR

In the TISR challenge, our method has not achieved outstanding results in the PSNR and SSIM metrics but in Track 1, it has got the second place in the LPIPS Alex and LPIPS Squeeze metrics and 10th place in the LPIPS VGG metric.

In Figs 4 (Track 1), 5 (Track 2, $\times 8$) and 6 (Track 2, $\times 16$), we present a sample of qualitative results obtained on the different TISR tracks. The first two columns are the inputs and the last two columns are the outputs (HR image and uncertainty map). In Figs. 5 and 6, in particular, one can notice how the visible image is used to infer the very small details (bars of the balcony, poles,...) that are completely lost in the LR image. Those regions with high frequencies are, again, part of those with large standard deviations in the rightmost image, altogether with regions of high thermical signature. On the unguided samples (Fig. 4), some artifacts

Figure 6. Results on Track 2 ($\times 16$) validation set. From left to right, LR, visible condition, prediction, standard deviation among multiple predictions.

are visible, e.g. windows and balconies on the right side of the building, but most of the high frequency content is reconstructed in a visually pleasant way.

Finally, in terms of computational times, we have measured that the average time for inference of the diffusion model was 2.77 to generate a batch of size 4, with an image size of $(640, 448)$ and an *NVIDIA Titan RTX* GPU.

### 4.3. Quantification of the aleatoric uncertainty

As commented above, using a generative model such as a diffusion model opens the door to providing an estimate of the uncertainties over the output HR image, by using a set of samples output from the generative model. In the following, we produce these estimates by using $N = 32$ samples.

As an illustration, in Fig. 7, we take one example from the validation dataset and compare the standard deviations on the pixel values reported through the diffusion model (left side) vs. the absolute values of the errors with respect to the ground truth image (right side). As we can see, and as one would expect, the highest uncertainty values occur in the image regions with highest frequencies: Edges, foliage, for example, i.e. those regions that are harder to reconstruct. Also, qualitatively speaking, the estimated deviation maps and error maps seem to be positively correlated. Finally, while on the top row, we give the deviations and errors for the direct outputs $\tilde{\mathbf{x}}^{+(i)}$ from the ResShift model, on the bottom row, we give the same data for the refined images $\hat{\mathbf{x}}^{+(i)}$. We can see that the effect of the refiner is to diminish both the variances on the samples and the errors.

In addition, for a more quantitative view on this aspect, in Table 1, we exhibit the computed standard deviations among samples of the validation dataset, for each track. What we can say from this table is that, in each case (different tracks and with/without Refiner module) the per-pixel standard deviation is positively correlated with the absolute errors; we also see that the Refiner modules always reduces the standard deviations; it has also a positive effect on the correlation itself, although this effect is significant only in the case of Track 2. It should be noted that images are normalized in the range $[0, 1]$.

| | | Avg Std | Avg Abs Error | Corr |
|---|---|---|---|---|
| T1 | No Refiner | 0.0127 | 0.0293 | 0.6444 |
| | Refiner | 0.0090 | 0.0264 | 0.6472 |
| T2 x8 | No Refiner | 0.0332 | 0.0512 | 0.5844 |
| | Refiner | 0.0095 | 0.0249 | 0.6485 |
| T2 x16 | No Refiner | 0.0286 | 0.0547 | 0.5296 |
| | Refiner | 0.0141 | 0.0374 | 0.6074 |

Table 1. Average per-pixel standard deviation is computed by generating $N = 32$ HR samples for each LR image from the validation dataset. Correlations are computed by averaging the correlation of the per-pixel standard deviation and the per-pixel absolute error over validation dataset, again with $N = 32$ samples.

### 4.4. Ablation results

The metrics for the results both with and without the Refiner module are presented in Table 2. As it can be seen, the Refiner module induces significant improvements for all the metrics, in all the cases, and mainly for the non-perceptual metrics (PSNR and SSIM).

Furthermore, as already commented and presented in Table 1, using the refiner network reduces the standard deviation of samples from the diffusion model while increasing the correlation between standard deviation and absolute errors.

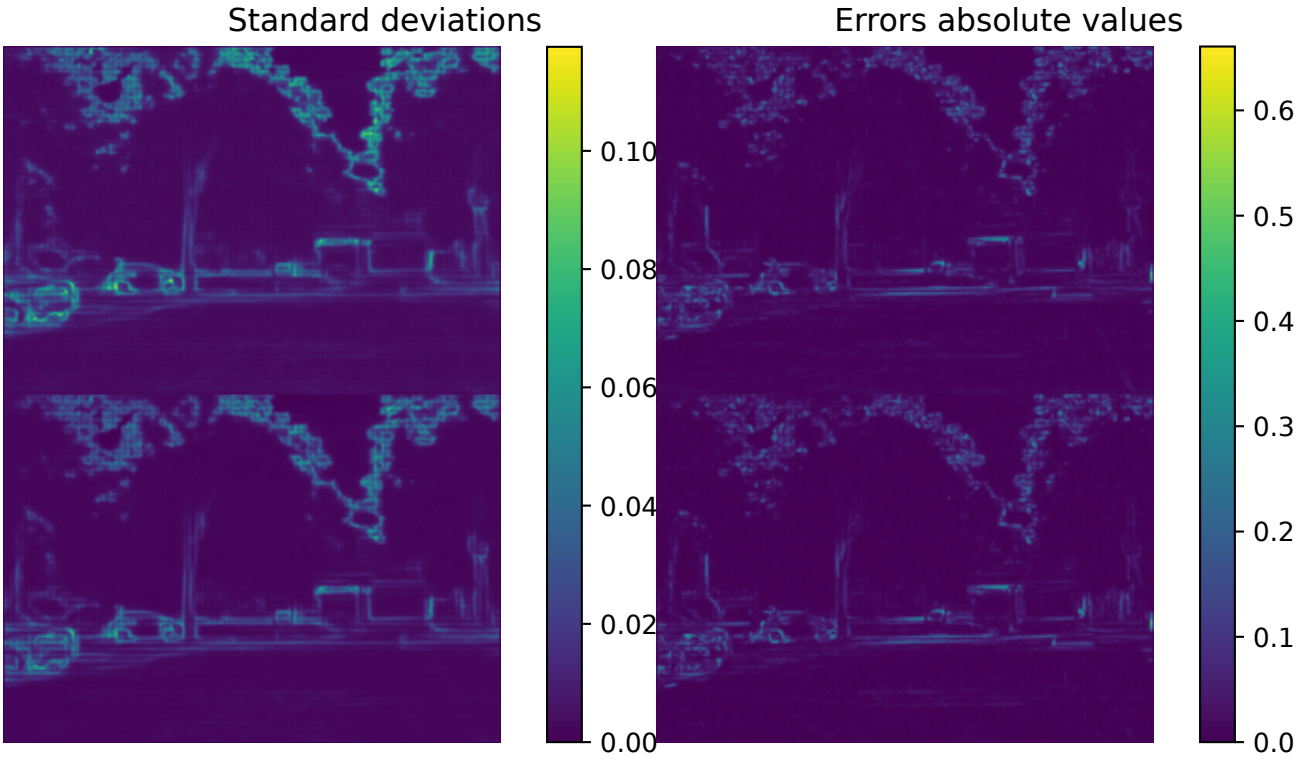


Figure 7. Aleatoric Uncertainty Quantification: For one example of the validation dataset, comparison of the standard deviations of the pixel values reported through the diffusion model (left) vs. the absolute values of the errors with respect to the ground truth image (right). On the top row, the deviations and errors are evaluated on the rough outputs $\tilde{\mathbf{x}}^{+(i)}$ from the diffusion model; on the bottom row, the results are provided for the final outputs $\hat{\mathbf{x}}^{+(i)}$ (after the refiner module).

| | | PSNR↑ | SSIM↑ | LPIPS Squeeze↓ | LPIPS Alex↓ |
|---|---|---|---|---|---|
| Track 1 | No Refiner | 25.52 | 0.7743 | 0.1492 | 0.2137 |
| | Refiner | 26.05 | 0.7920 | 0.1946 | 0.2466 |
| | Winner of Challenge | 27.52 | 0.8355 | 0.2149 | 0.2528 |
| Track 2 x8 | No Refiner | 19.30 | 0.6572 | 0.1916 | 0.2911 |
| | Refiner | 27.59 | 0.8288 | 0.1871 | 0.2348 |
| | Winner of Challenge | 31.52 | 0.9127 | 0.1270 | 0.1479 |
| Track 2 x16 | No Refiner | 18.48 | 0.6315 | 0.2092 | 0.3218 |
| | Refiner | 23.75 | 0.7470 | 0.2168 | 0.2987 |
| | Winner of Challenge | 25.99 | 0.8266 | 0.1786 | 0.2251 |

Table 2. Ablation results with respect to the proposed Refiner UNet, in which the diffusion model is frozen.

## 5. Conclusion

We have presented an exploration of conditional diffusion techniques for the problem of thermal imaging super-resolution, taking the PBVS TISR challenge as a playground. We have shown how the conditional diffusion model design allows us to train a single model on ImageNet and then fine-tune it for multiple tasks, including tasks with guidance from images of different modalities (such as from the visible domain).

A lightweight Refiner model has been proposed to leverage the generative power of diffusion while keeping artifacts under control. We have shown empirically that it helps reducing artifacts in images generated by a diffusion model.

Finally, we have shown that the diffusion model itself, through its generative capabilities, allows us to estimate pixel-wise uncertainties which loosely match the pixel-wise errors without using the original HR image. This pixel-wise uncertainty maps can be used to estimate where a super-resolution output may be wrong.

Among our planned future work, we will explore to diffusion guidance to use the images from other domains, i.e. by including in the diffusion sampling process gradient terms to enforce cross-modal similarity in the higher frequency content; we also plan to explore the uncertainty maps to adapt the refining efforts in the regions where we expect higher errors.

## References


[1] S. Derin Babacan, Rafael Molina, and Aggelos K. Katsaggelos. Total variation super resolution using a variational approach. In *2008 15th IEEE International Conference on Image Processing*, pages 641–644, 2008. 2

[2] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 2

[3] Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration, 2024. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[6] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5

[9] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843, 2017. 2

[10] Yanwei Pang, Jiale Cao, Jian Wang, and Jungong Han. Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. *IEEE Transactions on Information Forensics and Security*, 14:3322–3331, 2019. 1

[11] Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla, Dai Bin, Li Ruodi, Li Shengye, Zhiwei Zhong, Xianming Liu, Junjun Jiang, and Chenyang Wang. Thermal image super-resolution challenge results - pbvs 2023. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 470–478, 2023. 2, 3

[12] 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. 2

[13] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(04):4713–4726, 2023. 2

[14] Feng Shi, Jian Cheng, Li Wang, Pew-Thian Yap, and Dinggang Shen. Lrtv: Mr image super-resolution with low-rank and total variation regularizations. *IEEE Transactions on Medical Imaging*, 34(12):2459–2466, 2015. 2

[15] Jun Shi, Zheng Li, Shihui Ying, Chaofeng Wang, Qingping Liu, Qi Zhang, and Pingkun Yan. Mr image super-resolution via wide residual networks with fixed skip connection. *IEEE Journal of Biomedical and Health Informatics*, 23:1129–1140, 2019. 1

[16] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[18] Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 2022. 1

[19] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. 5

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[21] Z. Yue, Q. Zhao, J. Xie, L. Zhang, D. Meng, and K. K. Wong. Blind image super-resolution with elaborate degradation modeling on noise and kernel. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2118–2128, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2, 5

[22] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 4, 5

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5