Weakly Aligned Joint Cross-Modality Super Resolution

Guy Shacht       Sharon Fogel       Dov Danon       Daniel Cohen-Or
Tel-Aviv University

Abstract

Non-visual imaging sensors are widely used in the industry for different purposes. Those sensors are more expensive than visual (RGB) sensors, and usually produce images with lower resolution. To this end, Cross-Modality Super-Resolution methods were introduced, where an RGB image of a high-resolution assists in increasing the resolution of the low-resolution modality. However, fusing images from different modalities is not a trivial task; the output must be artifact-free and remain loyal to the characteristics of the target modality. Moreover, the input images are never perfectly aligned, which results in further artifacts during the fusion process.

We present CMSR, a deep network for Cross-Modality Super-Resolution, which unlike previous methods, is designed to deal with weakly aligned images. The network is trained on the two input images only, learns their internal statistics and correlations, and applies them to up-sample the target modality. CMSR contains an internal transformer that is trained on-the-fly together with the up-sampling process itself, without explicit supervision. We show that CMSR succeeds to increase the resolution of the input image, gaining valuable information from its RGB counterpart, yet in a conservative way, without introducing artifacts or irrelevant details.

1. Introduction

Super-Resolution (SR) methods are used to increase the spatial resolution and improve the level of detail of digital images, while preserving the image content. Such methods have important applications for multiple industries, such as health-care, agriculture, defense and film [28]. In recent years, more advanced methods of SR have been heavily based on Deep Learning [11, 23, 5] where one learns the mapping between Low-Resolution (LR) images and their High-Resolution (HR) counterparts, and applies the same mapping to an unseen low-resolution input, effectively performing super-resolution on that image.

The need for super-resolution becomes even more prominent when dealing with sensors other than the visible light, since those sensors typically produce images with lower resolution [20, 26]. For example, Infra-Red (IR) camera sensors are more expensive than classical camera sensors, and their output images commonly have much lower spatial resolution. While the aforementioned SR methods can still work on such images, there is still a big gap between the level of detail in the achieved results, and the one found in common RGB images. To bridge that gap, Joint Cross-Modality methods were developed. The idea is to use the higher-resolution RGB modality to guide the process of super-resolution on images taken by the lower resolution sensor, taking advantage of the finer details found in the RGB images. The challenge is to remain loyal to the target modality characteristics and to avoid adding redundant artifacts or textures from the RGB modality [2].

State-of-the-art Joint Cross-Modality SR methods rely on the assumption that their multiple inputs are well aligned [2, 3, 5, 7, 29]. Thus, they perform well only when the input images were captured by different sensors placed in the exact same position, and taken at the same exact time. In real-life scenarios, perfect alignment of multiple sensors is often hard to achieve. Aligning the images in a pre-process
typically yields only a weak alignment, dimming the effectiveness of joint cross-modality method.

In our work, we introduce a new method to perform joint cross-modality SR, where different modality images are allowed to be moderately misaligned, namely Weakly Aligned. We tackle the problem of misalignment using a learnable deformation that implicitly aligns the two images together. More specifically, our architecture includes a deformation model which aligns the RGB image to the target modality in a coarse-to-fine manner, before they are fused together. The network does not use any explicit supervision for the deformation subtask, but rather optimizes the deformation parameters to adhere to the super-resolution goal.

Furthermore, since most multi-modal pairs are not perfectly aligned, we are able to improve results even on supposedly well-aligned datasets, compared to previous methods (see Section 4). The SR module in our approach is based on ZSSR [31], and allows the network to perform SR using only the input pair without any training dataset. The network learns the internal statistics of the images by training on patches extracted from the input pair, and uses them to perform SR on the entire target modality image.

In addition to that, since over-transferal of information is an often arising problem in the world of multi-modal fusion [2, 3], our method is designed to transfer details from the RGB image carefully and conservatively; it avoids producing artifacts, or learning redundant details such as textures. It only learns the details that aid improving its super-resolution task. We show that our network achieves state-of-the-art results, while being generic in supporting any modality as input, requiring no training data and adjusting to any image size.

2. Related Works

Super-Resolution has been extensively studied throughout the last two decades. See [28] for a survey covering various SR techniques. Recent surveys [37, 5] cover more advanced methods, including Deep-Learning based methods. The first notable deep network-based method of SR method is SRCNN [11], a simple fully convolutional method that showed superior results to traditional methods. Like most methods, SRCNN uses external image datasets, like T91, Set5 and Set14 [22, 23] for training and evaluation.

However, it was claimed [16, 39, 41] that methods which rely on large external datasets do not learn the internal image-specific properties of the given input. In [16, 39], the subject of internal patch recurrence is investigated, leading to quantifiable results suggesting that patches of different scales tend to recur in the same image more than in external image datasets. This observation gave rise to powerful Zero-Shot methods [15, 31, 8], most notably ZSSR [31], which applies random cropping to its input image, effectively creating an internal image-specific dataset of patches taken solely from a single input. The method we present builds upon these ideas to deal with cross-modality, enjoying both the strong property of internal patch recurrence, together with the ability to transfer fine-modality details from our guiding modality input image to obtain super resolution images of even higher quality.

Other Modalities A straightforward generalization of SR performed on the visual modality is applying SR methods on varying modalities which are commonly acquired using low resolution sensors. Traditional SR methods for Thermal images (e.g., [25, 27]) have approached the problem by using signal reconstruction methodologies, whereas SR methods for depth-maps (e.g., [19, 36]) have been based on Markov random fields and coupled dictionary learning. Unlike the above methods, our method is generic in the sense that it can be applied to any given modality. In this paper, we evaluate our method on three modalities: Thermal (Infrared), NIR (Near-infrared), and depth-maps.

2.1. Joint Cross-Modality

In the Joint Cross-Modality setting the two different modalities are jointly analyzed to enhance one of them. As mentioned earlier, camera sensors capturing the RGB modality produce images with richer HR details than other modalities. Thus, a common setting is the usage of a visual HR version of the image, alongside with a LR version taken by the other modality sensor. This setting was adopted by all relevant joint cross-modality methods.

Visual-Depth In [29], a learning-based visual-depth method is presented. It is based on a CNN architecture operating on a LR depth-map and a sharp edge-map extracted from the HR visual modality. The network is trained on visual-depth aligned pairs from the Middlebury dataset [30]. In [38], a GAN-based method (CDeGAN) is presented. The method adds auxiliary losses that encourage keeping the resulting depth-maps smooth and texture-free, and is also trained on the Middlebury dataset.

Visual-Thermal (Infrared) In [7], a non learning-based joint visual-thermal method is presented. It is based on guided filtering of an up-sampled LR thermal input in areas that correlate well with the HR visual input. It is tested on visual-thermal pairs whose capturing sensors were manually calibrated to be aligned. Almastri et al. [2] introduced the learning-based visual-thermal SR methods VT-SRCNN and VTSRGAN, built on top of the existing SRCNN and SRGAN. They perform joint visual-thermal SR by concatenating feature maps extracted from each input

---

1In this paper, we use the terms RGB modality and visual modality interchangeably.
modality, and are trained and evaluated on the ULB17-VT [4] visual-thermal dataset consisting of well aligned pairs.

Cross-Modal Misalignment As noted by Almasri et al. [2], in the context of cross-modal super-resolution, misalignment is a major limitation in producing artifact-free SR results. In their paper, it is claimed that the artifacts added to the SR result appear where there is cross-modal displacements, and a better synchronized capturing device would likely solve that problem. Our method’s approach in handling cross-modal misalignment is to deform the RGB modality and align details that improve the SR objective to the target modality.

Our Method Our method differs from the aforementioned joint cross-modality techniques in two central aspects. First, it requires only weak alignment, as opposed to the aforementioned techniques which rely on well aligned pairs. Second, our network does not require any training data, and therefore avoids the need for a modal-specific dataset, relying on the internal image-specific statistics instead. This allows us to work on unseen modalities using a single architecture, and it is more suitable for cases where external modality image datasets are hard to obtain, making supervision practically impossible. Moreover, when facing unique modalities with high internal variance (i.e., the images look differently from one another), it is more feasible to rely on the internal image statistics, and not on a highly varied dataset, if one exists.

2.2. Image Registration

The subject of multi-modal image registration has been studied mainly in the context of medical imaging. Deep methods [32, 10, 9] have mostly based their architectures on a regressor, a spatial transformer and a re-sampler. They use supervision to optimize their regression and deformation models. It is also possible to use similarity metrics (like cross-correlation) [9] instead of training a regressor with supervision, and obtain an unsupervised registration framework.

In our work, multi-modal image registration is integrated into the main SR task. We use the same SR reconstruction loss to optimize our deformation parameters. Thus, we do not require aligned pairs for training. The deformation framework used in our method is divided into three steps in a coarse-to-fine manner [9, 12]. We first transform our image using global affine transformation for an initial rough approximation. Then, we further align our two modalities using CPAB [13, 33] transformation, which acts in a piecewise yet continuous manner. Finally, we use thin-plate spline (TPS) transformation for the final refinement of our alignment task.

3. Cross-Modality Super Resolution

The main motivation for our method is the ability to cope with pairs of images from different modalities which are only weakly aligned. To that end, our architecture includes a stage of local deformation which aligns objects in both images before they enter the SR network, as can be seen in Figures 4 and 5.

This concept can be used together with different super-resolution schemes. We chose to base our method on the ZSSR network of Shocher et al. [31] to enable our method to work on a single image, without pre-training. This has two key advantages: (i) it avoids the need to train on external image datasets, which are often scarce for various modalities, and (ii) it fully utilises the internal image statistics property, particularly relevant to non-standard capturing sensors.

Figure 4 describes the general architecture and the training process of our method, Cross-Modality Super-Resolution framework, CMSR. Our method includes three stages: a local deformation model to align the images of the different modalities, a patch selection phase which generates a training set out of a single pair of images, and a super-resolution network (CMSR). These components are introduced and described in Section 3.1. The way we incorporate those components into our training and inference schemes is covered by Sections 3.2 and 3.3.
3.1. Network Architecture

**Alignment using Learnable Deformation**  Our network corrects displacements between the two modalities on-the-fly, through a local deformation process applied to the RGB modality as a first gate to the network, optimized implicitly during training. In other words, instead of using explicit supervision to optimize the deformation parameters, they are trained with the super-resolution loss and therefore deform only parts which are relevant to this task. Our deformation process consists of three different transformation layers, performing the learned alignment in a coarse-to-fine manner.

The first layer of our deformation framework is the original Affine STN layer by Jaderberg et al. [18]. It captures a global affine transformation that is used to position the two modalities together as a rough initial approximation.

The second layer is a DDTN transformation layer (Deep Diffeomorphic Transformation Network, [33]), a variant of the original STN layer supporting more flexible and expressive transformations. Our chosen transformation model is CPAB (Continuous Piecewise-Affine Based, [13, 33]). It is based on the integration of Continuous Piecewise-Affine (CPA) velocity fields, and yields a transformation that is both differentiable and has a differentiable inverse. It is Continuous Piecewise-Affine w.r.t a tessellation of the image into cells. For this reason, it is well suited to our alignment task; each cell can be deformed differently, yet continuity is preserved between neighboring cells, yielding a deformation that can express local (per-cell) misalignments while preserving the image semantics.

The third and last layer of our deformation framework performs a TPS (Thin-plate spline) transformation, a technique that is widely used in computer vision and particularly in image registration tasks [6]. Our implementation (also taken from [33]) learns the displacements of uniformly-distributed keypoints in an arbitrary way, while each keypoint’s surrounding pixels are displaced in accordance to it, using interpolation [6]. Since TPS displaces its keypoints freely, the displacement is unconstrained to any image transformation model, and has the power to align the fine-grained objects of the scene, providing the final refinement of our alignment task.

**Patch Selection** Similarly to ZSSR [31] we produce our training set from a single pair of images by sampling...
patches using random augmentations. In our implementation we use scale, rotation, shear and translations. This random patch selection yields two patches that correspond to roughly the same area in the scene: one taken from the target modality and the second is taken from the deformed RGB modality which was previously aligned to the target modality.

**CMSR network** The CMSR network is the main component of our architecture as it is the component responsible for performing super-resolution. Namely, it produces a HR version of its target modality LR input image, guided by its HR RGB input. As Figure 4 and Figure 5 suggest, this component can be applied to varying image sizes, thanks to its fully convolutional nature.

The fully convolutional architecture of CMSR is based on the one from Shocher et al. [31]. However, a few changes have been made to better apply it to cross-modality SR (see Figure 6). The first gate to the network is up-sampling of the LR modality input to the size of the RGB input. This is done naively, using the Bi-cubic method, in case no specific kernels are given. From the up-sampled modality input we generate a feature map using a number of convolutional layers, denoted as Feature-Extractor 1 in Figure 6. From the RGB modality input that was previously aligned to target modality input, we generate a feature map using Feature-Extractor 2. We perform summation of the two resulting feature maps, one from each Feature-Extractor block, alongside with an up-sampled version of the LR target modality image, in a residual manner. This yields our HR super-resolved output.

**3.2. Training**

During each training iteration, we perform local deformation on the RGB modality input and produce a displaced version of it, aligned to the target modality image, as described in 3.1. Then, a random patch is selected from the input pair (illustrated in Figure 4), yielding two corresponding patches; one taken from the target modality, and the second from the displaced (aligned) RGB modality, as described in 3.1. The patch selection phase is an integral part of the network, and is done in a differentiable manner, so as to allow the gradients to backpropagate through it to the deformation model. This enables us to optimize the transformation on the entire RGB image despite using patches of the image during training.

In order to generate supervision for the training process, we down-sample both patches (Figure 4 in Green) by our desired SR ratio (e.g., 2x, 4x), denoted as $r$. The modality patch is down-sampled to allow training the network to reconstruct it with self-supervision, whereas the RGB patch is down-sampled accordingly, to keep the ratio between the two patches equal to $r$. Instead of down-sampling the RGB patch, it is also possible to naively up-sample the modality patch, and still preserve the same ratio, $r$, between patches. We found that by alternating between up-sampling and down-sampling of the aforementioned patches, we are able to significantly improve the results. More details regarding this technique can be found in the supplementary material.

**3.3. Inference**

At inference time, we use the trained CMSR network and deformation parameters, to perform SR on the entire target modality image guided by the RGB modality image (see Figure 5).

Since CMSR is fully convolutional, it can operate on any image size (e.g., both image patches of different scales, and full images) using the same network. We first apply the alignment dictated by the optimized deformation parame-
Figure 7: From left to right, respectively: ZSSR [31], CMSR (our method), Ground-Truth, and the RGB input. Here, CMSR succeeded to produce the windows despite never seeing their Thermal (IR) representation.

4. Results and Evaluation

4.1. Implementation Details

Our model is implemented in Tensorflow 1.11.0 and trained on a single GeForce GTX 1080 Ti GPU. The full code and datasets will be published upon acceptance in the project’s GitHub page. We typically start with a learning rate of 0.0001 and gradually decrease it to $10^{-6}$, depending on the slope of our reconstruction error line, whereas the learning rates of our transformation layers follow the same pattern, multiplied by constant factors. Those factors are treated as hyper-parameters, and should typically be larger when dealing with highly displaced input pairs, like in the case of Weakly Aligned modalities (Figure 3). Performing a 4x SR on an input of size 60x80 typically takes 30 to 60 seconds, depending on the desired number of iterations. To achieve SR of higher scales, we perform gradual SR with intermediate scales, as this further improves the results [21][35][31].

For Feature-Extractor 1 we use eight hidden layers, each containing 64 channels and a filter size of 3x3. We place a ReLU activation function after each layer except for the last one. The size of feature maps remains the same throughout all layers in the block. For Feature-Extractor 2 we typically use four to eight hidden layers with number of channels ranging from 4 to 128, a filter size of 3x3 and a ReLU activation function. The last layer has no activation and a filter size of 1x1. We find that highly detailed RGB inputs require Feature-Extractor 2 to have more channels. The hyper-parameters rarely require adjustments; they only require manual tuning when dealing with inputs that are unique, unusual, or ones that reflect very unusual displacements.

4.2. Evaluation with State-of-the-arts

**Strongly Aligned Modalities** We compared our method to cross-modal state-of-the-art SR methods on strongly aligned pairs. We used the ULB17-VT dataset [4], consisting of visual-thermal pairs that are mostly well aligned, as shown in Figure 2 (bottom row). This proves to be an easier case for joint cross-modal super-resolution, and typically requires only local understanding of the input pair. We have included the results of our evaluation in Table 1 showing that our method, despite not being previously trained, beats competing methods, averaged across the ULB17-VT dataset which was used by the said methods for evaluation in their original papers. Figures 8 and 11 include some visual results.

**Weakly Aligned Modalities** The Middlebury dataset [30] contains strongly aligned depth-visual pairs as shown in Figure 2 (top row). In that dataset, multiple angles and different sensor placements are included, for each pair. To obtain weakly-aligned pairs, we shuffled the pairs together such that the resulting pairs would correspond to a small sensor misplacement, shown in Figure 3 (left pair). We further increased the size of the dataset through random augmentations. We denote the new resulting dataset as Shuffled-Middlebury. CMSR surpasses competing cross-modal methods on those weakly aligned pairs by using a coarse-to-fine alignment approach, as summarized in Table 1.

**Single modality baseline model** We evaluated CMSR against the baseline state-of-the-art single modality method, ZSSR [31]. Our experiment shows that our method leverages the fine details in its RGB input and produces a SR output that is both appealing to the eye, and numerically closer to a Ground-Truth version, as shown in Figures 7, 8, 12 and 9.

4.3. RGB Artifacts

A fusion of multiple image sources, often causes the transfer of unnecessary artifacts from one modality to the other (e.g., [2]). Those artifacts not only sabotage the quality of the image, but harm the modality characteristics and could potentially make it unusable. Our method learns only the relevant RGB information that improves SR results; Figures 12 and 11 show cases where the RGB modality input contains a great amount of textural information, yet our SR output remains texture-free. In Figure 8 the learned RGB
Figure 8: We compare our method to its baseline method, ZSSR [31], as well as to another cross-modality method, VTSR-GAN [2]. On the right, the output of Feature-Extractor 2 (Figure 6) is given as the learned RGB residual which is added to our output. This RGB residual is artifact-free, contains no unwanted textures, and in fact, resembles an edge-map. For this reason, CMSR produces images that are visually pleasing, free of artifacts, and numerically better than competing methods.

Figure 9: Evaluation on the Near-infrared (NIR) modality, obtained from the RGB-NIR EPFL dataset [1]. This time, we let CMSR perform 4x SR in a single shot, without intermediate scales. The results surpass the baseline model by a margin, showing that CMSR learns fine details from its RGB input successfully.

residual is given; it contains no irrelevant textures and it resembles an edge-map, used to sharpen our output image.

4.4. Local Deformation

As shown in Figure [10] our method aligns the RGB modality input to the target modality input on-the-fly, to aid the joint cross-modal SR task. Although we have no aligned RGB ground-truth image, nor any target modality ground-truth image, we still correct those cross-modal misalignment successfully, thanks to an expressive deformation framework integrated into our architecture. The deformation parameters are optimized using the SR reconstruction loss; hence we learn only the deformations that are needed to minimize that loss and assist in the SR task.

5. Conclusions

We have introduced CMSR, a method for cross-modality super-resolution. Our method utilizes an associated high-resolution RGB image of the scene to boost its accuracy. The method presented is generic and yet outperforms state-of-the-art methods, even when its two modalities are misaligned, as elaborated below.

Generic. To the best of our knowledge, CMSR is the first self-supervised cross-modal SR method. It requires no training data, a prominent advantage when dealing with
scarce and unique modalities. It is trained on the target image only, and can thus, take any modality as input, and learns its internal, possibly unique, statistics, adapting to the unknown imaging conditions and down-scaling kernels. Furthermore, the method can be applied to any image sizes, and to any ratio between the two inputs. This is unlike other architectures that use strides for up-sampling [2], thus they are fixed to a specific image size and constant scale factor.

Figure 10: To demonstrate the alignment capabilities of CMSR, we evaluated it on a severely misaligned visual-thermal pair, (a) and (b), containing both global and local displacements. We overlaid the images with semi-transparency, once before evaluation (c), and again after training (d). CMSR deformed its RGB input and aligned it to the thermal modality in a coarse-to-fine manner, on-the-fly, and without supervision.

Figure 11: Our super-resolution method (CMSR) uses its RGB input, given in (d), conservatively. Here, we compare a patch, rich in RGB textures, taken out of the ULB17-VT evaluation results [1]. Compared to VTSRGAN (a), our result (b) is artifact-free. Ground-Truth is given in (c) as reference.

Figure 12: CMSR does not introduce irrelevant details. Note that despite the large amount of textural information in its RGB input (d), CMSR (b) ignores it and learns only the relevant information. The result is better than the baseline model, ZSSR (a), in comparison to a GT version (c).

Table 1: We compared CMSR to competing cross-modal SR methods, VTSRCNN and VTSRGAN [2], on the Strongly Aligned ULB17-VT dataset [4], as well as on the Weakly Aligned Shuffled-Middlebury dataset created by us. We have taken the mean PSNR / SSIM scores, measured against the modality 4x GT versions.

| Metric | Dataset | VTSRGAN | VTSRCNN | CMSR |
|--------|---------|---------|---------|------|
| PSNR   | U-VT    | 27.988  | 27.968  | 29.928 |
| SSIM   | U-VT    | 0.8202  | 0.8196  | 0.882  |
| PSNR   | SMB     | 27.925  | 28.189  | 28.652 |
| SSIM   | SMB     | 0.9547  | 0.9386  | 0.9341 |

Table: Performance. Our method is conservative, in the sense that it learns from its RGB features only when it contributes to the up-sampling process, without introducing outliers, ghosts, halos, or other artifacts.

We achieve state-of-the-art results, qualitatively (visually) and quantitatively, compared to competing cross-modal methods, as well as to our state-of-the-art single-modality baseline. Specifically, we show that the RGB modality indeed greatly contributes as a guide to the up-sampling process.

Misalignment: A unique property of our method is that it is robust to cross-modal misalignment. This property is imperative, since in real life conditions, sight misalignment is, more often than not, unavoidable. It should be emphasized that the alignment is done without pre-training or any supervision.

In the future we would like to further enhance our technique by applying the deformation in the feature space instead of the RGB pixel-space. The hope is that in this way, it would be possible to adopt a deformation-per-feature scheme that would reflect different displacements for different scene objects, possibly using segmentation.
6. Additional Results

In Figure 13, additional results from our evaluation on the EPFL NIR dataset are included. This dataset was originally used in Figure 9 of the original paper. The results indicate that our method avoids transferring unnecessary RGB textures to its output; it only learns from its RGB input when it contributes to the results. This conservative approach enables CMSR to surpass state-of-the-art cross-modal methods, despite the fact that those competing methods were pre-trained extensively on the full dataset, whereas our method operates on its single input pair, without pre-training, in a Zero-Shot manner.

7. Alternating Scales

In Section 3.2 of the submitted paper, the Alternating Scales technique is briefly discussed. It corresponds to training CMSR using two different scales, alternating between them across iterations. Here, we wish to further elaborate on this technique.

7.1. Alternating Scales - Elaboration

Denoting our desired SR ratio (e.g. 2x, 4x) by $r$, our network, CMSR, takes a target modality input of size $H \times W$ alongside with an RGB input of size $rH \times rW$, and produces a target modality output of size $rH \times rW$. Hence, by design, a ratio of $r$ must be preserved between CMSR’s two inputs (The architecture of CMSR is given in the original paper, Figure 6). Since CMSR is trained to reconstruct a random patch taken from its modality input (Figure 4 of the original paper, Training process), this random patch is down-sampled, by ratio $r$, before it is reconstructed by the CMSR network. However, since the ratio between CMSR’s two inputs must remain $r$, the corresponding RGB patch is also down-sampled accordingly, by ratio $r$. This way, we preserve the same ratio between CMSR’s two input patches, as needed.

Nonetheless, instead of down-sampling the RGB patch to match this required ratio, it is also possible to naively up-sample the modality patch by ratio $r$. Clearly, this has the same effect on the ratio between the two patches, which yet again remains $r$. However, this way, we obtain a different training scheme. Figure 14 compares the two different schemes, corresponding to the two different scales CMSR operates on.

We found that by alternating between the two schemes during training, we are able to significantly improve our results. We name this combination of training schemes as the Alternating Scales technique. It allows our network to be optimized using patches of their original scale, as explained in Table 2. We observe that training our network on patches of their original scale improves its generalization capabilities, since during the inference stage, the network operates on the full input pair, at its original scale.

7.2. Alternating Scales - Ablation Study

We have conducted an experiment to show the improvement obtained by the Alternating Scale technique. We trained CMSR using the two schemes (see Figure 14 and Table 2 for information on the schemes), alternating between them randomly. We used the Upsampling-Based scheme with probability $p$ and the Downsampling-Based with probability $1 - p$.

According to the results, summarized in Figures 15 and 16, the best PSNR was obtained when $p = 0.3$, which starts decaying when $p > 0.3$. We notice that $p > 0$ always yields better results than $p = 0$. This observation is important, since the risk of using sub-optimal $p$ values on new, unseen input pairs is minimal; using this technique is always better than not using it, regardless of $p$.

8. Alignment using Learnable Deformation - Ablation Study

To show the necessity of each layer of our coarse-to-fine deformation framework (Section 3.1 of the original paper), we evaluated CMSR on a Weakly Aligned pair, adding one layer at a time, averaged across multiple runs. The results indicate that each layer is necessary and plays a different role, which can be seen visually in Figure 18, and numerically in Figure 17.

Two additional points should be mentioned; First, we remind that our goal is not to perform image registration. Hence, we measure the quality of alignment through the quality of the yielded SR result, and not by conventional image registration metrics. Second, when CMSR is evaluated

| Training Scheme | Modality Scale | RGB Scale |
|-----------------|----------------|-----------|
| Down-sampling  | Original       | Down-scaled|
| Up-sampling    | Up-scaled      | Original  |

Table 2: In the Downsampling-Based training scheme, CMSR takes a down-sampled RGB input patch, but its modality input patch is reconstructed at its true, original scale. However, in the Upsampling-Based scheme, CMSR takes an original RGB input patch, at its true scale, but reconstructs a modality patch that was up-sampled beforehand.
Figure 13: We compared CMSR both to its single-modality baseline, ZSSR [31], and to a competing cross-modality method, VTSRCNN [2], on the NIR modality [1]. Our method, CMSR, is able to produce super-resolved images that are both visually pleasing, and numerically closer to a Ground-Truth version.
Figure 14: The difference between the two training schemes lies in the scale CMSR (in Orange) operates on. The two schemes start with the exact same input pair, but in the Upsampling-Based training scheme (right), CMSR is fed inputs of larger scale. This scale difference is also explained in Table 2.

Figure 15: We evaluated CMSR using different alternation probabilities. Namely, we trained it using the Upsampling-Based training scheme (Figure 14) in fraction $p$ iterations, and using the Downsampling-Based scheme in the remaining fraction $1 - p$. We averaged this experiment across multiple runs. According to the results, $p = 0.3$ yields the best PSNR (32.476 dB). This can be also seen visually, in Figure 16.

with no transformation layers on a severely misaligned pair (like the one in Figure 18), its RGB input remains mostly unused, enabling CMSR to produce a result that is comparably worse (as shown in 17), but does not reflect the failed fusion of misaligned RGB objects. This conservative approach allows our method to surpass competing cross-modal SR methods. CMSR leverages its RGB modality input only when it contributes to the final SR result; when CMSR has no transformation layers, a severely misaligned RGB input will mostly be ignored.

Figure 16: We compare two patches taken from the Alternating Scales ablation study results, summarized in Figure 15. According to our experiment, the best SR result is obtained when using $p = 0.3$ as the alternation probability.

Figure 17: We let CMSR perform 4x SR on a Weakly Aligned visual-thermal pair, with different transformation layers, averaged across 5 runs. The results indicate that each layer contributes to the final PSNR, which can also be seen visually in Figure 18.
Figure 18: We evaluated CMSR using different transformation layers. In the leftmost column, the resulting deformed RGB image is given. In the other columns we show the resulting alignment, visualized through blending of the R-G (Red-Green) channels of the aforementioned deformed RGB image, together with the Ground-Truth thermal image (which is unavailable to CMSR).
References

[1] RGB-NIR scene dataset.

[2] Feras Almasri and Olivier Debeir. Multimodal sensor fusion in single thermal image super-resolution. arXiv preprint arXiv:1812.09276, 2018.

[3] Feras Almasri and Olivier Debeir. RGB guided thermal super-resolution enhancement. In 2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech), pages 1–5. IEEE, 2018.

[4] Feras Almasri and Olivier Debeir. ULB17-VT. 2 2019.

[5] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. CoRR, abs/1904.07523, 2019.

[6] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell., 11(6):567–585, 1989.

[7] Xiaohui Chen, Guangtao Zhai, Jia Wang, Chunjia Hu, and Yuanchun Chen. Color guided thermal image super resolution. In 2016 Visual Communications and Image Processing (VCIP), pages 1–4. IEEE, 2016.

[8] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen. Deep network cascade for image super-resolution. In European Conference on Computer Vision, pages 49–64. Springer, 2014.

[9] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hesam Sokooti, Marius Staring, and Ivana Isgum. A deep learning framework for unsupervised affine and deformable image registration. CoRR, abs/1809.06130, 2018.

[10] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Marius Staring, and Ivana Isgum. End-to-end unsupervised deformable image registration with a convolutional neural network. CoRR, abs/1704.06065, 2017.

[11] Chao Dong, Chen-Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. CoRR, abs/1501.00092, 2015.

[12] Tobias Fechter and Dimos Baltas. One shot learning for deformable medical image registration and periodic motion tracking. CoRR, abs/1907.04641, 2019.

[13] O. Freifeld, S. Hauberg, K. Batmanghelich, and J. Fisher III. Transformations based on continuous piecewise-affine velocity fields. In IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017.

[14] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In ICCV, 2009.

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[16] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, pages 349–356, 2009.

[17] Michal Irani and Shmuel Peleg. Improving resolution by image registration. CVGIP: Graph. Models Image Process., 53(3):231–239, Apr. 1991.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. CoRR, abs/1506.02025, 2015.

[19] Daeyeong Kim and Kuk-jin Yoon. High-quality depth map up-sampling robust to edge noise of range sensors. In 2012 19th IEEE International Conference on Image Processing, pages 553–556. IEEE, 2012.

[20] Y. Kiran, V Shrinishiti, W Lino Hans, and N Venkateswaran. A single-image super-resolution algorithm for infrared thermal images. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 17(10):256–261, 2017.

[21] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. CoRR, abs/1704.03915, 2017.

[22] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[23] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. CoRR, abs/1609.04802, 2016.

[24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. CoRR, abs/1707.02921, 2017.

[25] Hai-cang Liu, Shu-tao Li, and Hai-tao Yin. Infrared surveillance image super resolution via group sparse representation. Optics Communications, 289:45–52, 2013.

[26] Emanuele Mandanici, Luca Tavasci, Francesco Corsini, and Stefano Gandolfi. A multi-image super-resolution algorithm applied to thermal imagery. Applied Geomatics, Feb 2019.

[27] Yuxing Mao, Yan Wang, Jintao Zhou, and Haiwei Jia. An infrared image super-resolution reconstruction method based on compressive sensing. Infrared Physics & Technology, 76:735–739, 2016.
[28] Kamal Nasrollahi and Thomas B. Moeslund. Super-resolution: A comprehensive survey. *Mach. Vision Appl.*, 25(6):1423–1468, Aug. 2014.

[29] Min Ni, Jianjun Lei, Runmin Cong, Kaifu Zheng, Bo Peng, and Xiaoting Fan. Color-guided depth map super resolution using convolutional neural network. *IEEE Access*, 5:26666–26672, 2017.

[30] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, Apr. 2002.

[31] Assaf Shoher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. *CoRR*, abs/1712.06087, 2017.

[32] Martin Simonovsky, Benjamin Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. *CoRR*, abs/1609.05396, 2016.

[33] Nicki Skafte Detlefsen, Oren Freifeld, and Soren Hauberg. Deep diffeomorphic transformer networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4403–4412, 2018.

[34] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. *CoRR*, abs/1511.02228, 2015.

[35] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. *CoRR*, abs/1804.02900, 2018.

[36] Jun Xie, Cheng-Chuan Chou, Rogerio Feris, and Ming-Ting Sun. Single depth image super resolution and denoising via coupled dictionary learning with local constraints and shock filtering. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[37] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, and Jing-Hao Xue. Deep learning for single image super-resolution: A brief review. *CoRR*, abs/1808.03344, 2018.

[38] Lijun Zhao, Jie Liang, Huihui Bai, Anhong Wang, and Yao Zhao. Simultaneously color-depth super-resolution with conditional generative adversarial network. *CoRR*, abs/1708.09105, 2017.

[39] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984. IEEE, 2011.