SINGLE PAIR CROSS-MODALITY SUPER RESOLUTION

Anonymous authors
Paper under double-blind review

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 a low-resolution modality. However, fusing images from different modalities is not a trivial task, since each multi-modal pair varies greatly in its internal correlations. For this reason, traditional state-of-the-arts which are trained on external datasets often struggle with yielding an artifact-free result that is still loyal to the target modality characteristics.

We present CMSR, a single-pair approach for Cross-Modality Super-Resolution. The network is internally trained on the two input images only, in a *self-supervised* manner, learns their internal statistics and correlations, and applies them to up-sample the target modality. CMSR contains an internal transformer which is trained on-the-fly together with the up-sampling process itself and without supervision, to allow dealing with pairs that are only weakly aligned. We show that CMSR produces state-of-the-art super resolved images, yet without introducing artifacts or irrelevant details that originate from the RGB image only.

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. (Nasrollahi et al. (2014)) In recent years, more advanced methods of SR have been heavily based on Deep Learning. (Dong et al. (2015); Ledig et al. (2016); Anwar et al. (2019))

The need for super-resolution becomes even more prominent when dealing with sensors that capture other modalities, different than the visible light spectrum, since those sensors typically produce images with substantially lower resolution. (Kiran et al. (2017); Mandanici et al. (2019)) For example, Infra-Red (IR) camera sensors are more expensive than classical camera sensors, and their output images commonly have much lower spatial resolution. To bridge that gap in level-of-detail, 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 that may be present only in the RGB modality.

In this work, learning is performed internally, relying solely on the input pair of images. This approach does not require any training data, and therefore avoids the need for a modal-specific dataset, relying solely on the internal image-specific statistics instead. (Shocher et al. (2017)) Using an internal super-resolution method is *particularly* strong in the context of cross-modality, since it allows the network to fit to the unique properties and the modality characteristics of the specific input pair. This feature stands in contrast to the somewhat impractical task of generalizing to a large cross-modal image dataset; each multi-modal pair is inherently unique in its internal correlations, and therefore must be treated differently.

State-of-the-art Joint Cross-Modality SR methods rely on the assumption that their multiple inputs are well aligned. (Almasri et al. (2018a;b); Zhao et al. (2017); Chen et al. (2016); Ni et al. (2017); de Lutio et al. (2019); Li et al. (2017)) Thus, they perform well only when the input images were

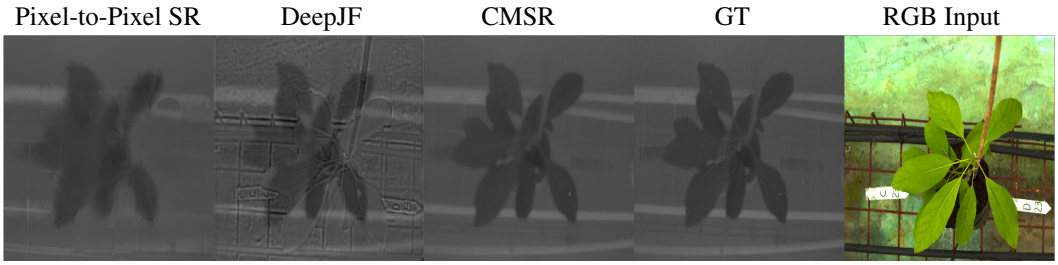

Figure 1: In this **visual-thermal** pair, the two inputs are displaced (visualised in Figure 9), a phenomenon which often occurs in real-world scenarios. This is a failure case even for state-of-the-arts (Pixel-to-Pixel SR de Lutio et al. (2019) and DeepJF Li et al. (2017)), as they are based on perfect alignment. CMSR adapts to the given misalignment, and corrects it as a part of the SR process, producing a sharp result (41.069 dB PSNR). The deformed RGB image is presented in Figure 9

captured by different sensors placed in the exact same position, and taken at the exact same time. As to be shown, in real-life scenarios perfect alignment of multiple sensors is often hard to achieve. In our work, we present new means to allow the two modalities to be moderately misaligned, namely weakly aligned. Our network contains a learnable deformation component that implicitly aligns details in the two images together. More specifically, our architecture includes a deformation model that aligns details from 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 sub-task, but rather optimizes the deformation parameters to adhere to the super-resolution goal. Figures 1 and 7 present cases where a weakly aligned pair causes state-of-the-arts methods to fail, whereas our method produces high-quality super-resolved output.

Another notable advantage of our single-pair approach is the avoidance of over-transferal of information. Previous approaches which train on external cross-modal datasets are often limited in their ability to adjust to the specific nature of the input pair. For this reason, they often struggle in cases where the guiding modality should be only minimally used, or even completely ignored; they tend to fuse redundant details from the guiding modality anyway, resulting in the addition of textures and artifacts to the lower resolution modality. Our method is designed to adjust to the specific input pair and therefore transfers details from the higher resolution image carefully and conservatively, learning only the details which aid improving the super-resolution task. Figure 2 presents an example with cross-modality ambiguity. Namely, the RGB modality contains an object which does not exist in the target modality; this object should ideally be ignored. Our network successfully avoids transferring it, whereas a competing cross-modality method results in unwanted artifacts and textures. We show that our network achieves state-of-the-art results on different modalities (Thermal, NIR, Depth), while being generic in supporting any modality as input and requiring no pre-training (and thus, no training data).

2 Related Works

Super-Resolution has been extensively studied throughout the last two decades. See Nasrollahi et al. (2014) for a survey covering various SR techniques. Recent surveys (Yang et al. (2018); Anwar et al. (2019)) cover more advanced methods, including Deep-Learning based methods. The first notable deep network-based method of SR method is SRCNN, (Dong et al. (2015)) 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 (Lai et al. (2018); Ledig et al. (2016)) for training and evaluation.

However, it was claimed (Irani & Michal (2009); Zontak et al. (2011); Shocher et al. (2017)) that methods which rely on large external datasets do not learn the internal image-specific properties of the given input. In Irani & Michal (2009); Zontak et al. (2011), the subject of internal patch recurrence is investigated, and the benefits of a single-image approach were shown. This strong observation gave rise to powerful Zero-Shot methods, (Huang et al. (2015); Shocher et al. (2017); Cui et al. (2014)) and most notably ZSSR. (Shocher et al. (2017)) Our work extends the concept of Zero-Shot to cross-modality. This way, we not only enjoy the advantage of being dataset indepen-

dent and adjusting to any modality, but we also enable our network to adapt to the specific properties (which are to be discussed) of the specific input pair.

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.

In Ni et al. (2017), 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. (Scharstein et al. (2002)) The method Deep Joint Image Filtering (Li et al. (2017)) presented a framework for denoising and upsampling depth-maps. Their network performs concatenation of features extracted from the guiding image and features extracted from the target modality image. It was evaluated on the Middlebury dataset and has shown promising results for its task. It is however designed to be pretrained on a full multi-modal datasets of perfectly aligned pairs. Almastri et al. (Almasri et al. (2018a)) introduced the learning-based visual-thermal SR methods VTSRCNN 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 modality, and are trained and evaluated on the ULB17-VT (Almasri et al. (2019)) visual-thermal dataset consisting of well aligned pairs. In Guided Super-Resolution as Pixel-to-Pixel Transformation (de Lutio et al. (2019)), the problem of guided depth-maps SR was posed as a pixel-to-pixel translation of the HR guiding modality to a newly predicted HR depth-map, constrained by the intensities of the matching regions from the LR depth-map input. This method has shown to produce sharp HR depth-maps, but is based on perfect alignment between the input and the guiding modality.

Cross-Modal Misalignment In the context of cross-modality super-resolution, misalignment is a major limitation in producing artifact-free SR results. This was previously discussed in related works, (Almasri et al. (2018a); Li et al. (2017)) and shown explicitly in this paper. 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 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 feature is especially attractive for cross-modality super-resolution; learning from the single input pair encourages the network to adapt to the specific cross-modal properties existing in that particular pair, which may be unique. This is unlike previous state-of-the-arts who are trained on external datasets and are often limited in their capability to adapt to a specific pair, and therefore result in over-transferal of information (such as in Figures 2, 10 and 14). Second, it requires only *weak* alignment, as opposed to the aforementioned techniques which rely on well aligned pairs. This attribute is critical when operating in real-life scenarios. For example, state-of-the-arts like Pixel-to-Pixel SR (de Lutio et al. (2019)) and DeepJF (Li et al. (2017)) struggle when applied to pairs that were captured in less-than-optimal imaging conditions, as shown most notably in Figures 1 and 7.

2.2 MULTI-MODAL ALIGNMENT

The subject of multi-modal image registration has been studied mainly in the context of medical imaging. Deep methods (Simonovsky et al. (2016); de Vos et al. (2017; 2018)) 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 (de Vos et al. (2018)) instead, and obtain an unsupervised registration framework. In Arar et al. (2020), unsupervised registration was performed through an image-to-image translation objective. Namely, the better the network translates one modality to the second modality (which is the modality being deformed), the better the deformation is assumed to be done.

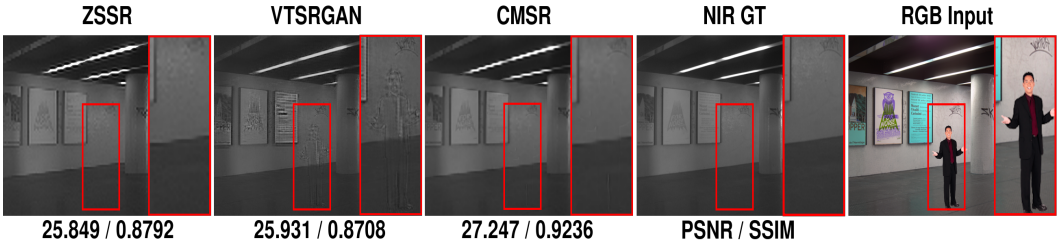

Figure 2: CMSR successfully ignores details that appear only in the RGB image (a standing man). It does not add ghosts, while VTSRGAN does. CMSR also surpasses the baseline single-modality method, ZSSR, which only operates on the NIR input.

In our work, multi-modal image registration is **not** performed per se. Our goal is not to register the two input modalities together completely, but to give the network enough freedom to align only the details that assist and adhere to the super-resolution task. For this reason, it is possible that the network chooses to only partly align the guiding modality. The alignment phase 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 consists of three steps performed in a coarse-to-fine manner, with the help of affine, CPAB and TPS layers. (O. et al. (2017); Skafte Detlefsen et al. (2018)) More specific details are found in later chapters.

3 Single-Pair Cross-Modality Super Resolution

One of the fundamental problems of cross-modality super resolution is that it is hard to transfer only the relevant details from the higher resolution image to the lower resolution one while ignoring unnecessary details, which often cause ghosts and unwanted artifacts (such as those in Figure 2 and Figure 14 found in the Appendix). When training on a large dataset of cross-modality image pairs, it is hard (and often impossible) for a network to learn which details exactly should be transferred and which should not be, for each given cross-modality pair. This is mostly because similar objects might be present (and thus, should be transferred) in some pairs in the dataset, and not in others. To avoid this problem, we opted to use a super resolution method trained on a *single* input pair, and enable the network to adapt to it specifically.

3.1 Network Architecture

Our network includes a patch selection component which generates a training set out of a single pair of images, and a super-resolution network. Our method enables dealing with misaligned pairs by including a deformation phase, done internally, which aligns objects in both images right before they enter the SR network (see Figures 4 and 5). We hereby introduce and describe the components of our network, which are incorporated into our training and inference schemes as covered in Sections 3.2 and 3.3.

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. This is done by three transformation layers operating in a coarse-to-fine-manner; a TPS layer, an affine layer and a CPAB layer. Further details are found in Appendix A.2.

Patch Selection 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.

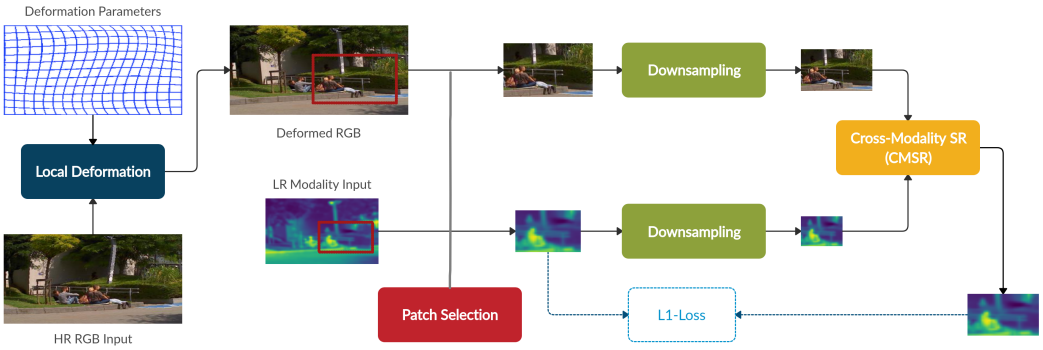

Figure 4: *Training process.* The RGB image first goes through a deformation step which aligns it to the target modality (in blue). Then, random patches are selected by an augmentation step (in Red) and down-sampled (in green). The patches are used to train the CMSR network (in orange) and the deformation parameters. The loss function is measured between the super-resolved output and the input target modality images.

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

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. [1] From the up-sampled modality input we generate a feature map using a number of convolutional layers, denoted as ***Feature-Extractor 1*** in Figure 3. 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.

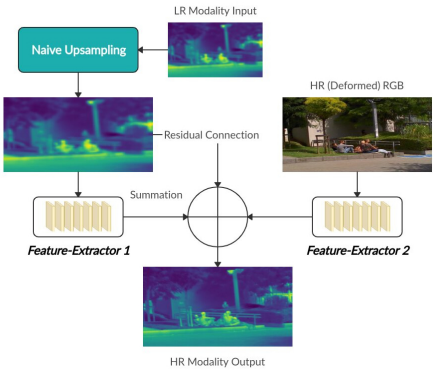

Figure 3: CMSR performs three-way summation; two of the resulting feature maps, one from each modality, are summed together with the original modality input that is naively up-sampled, in a residual manner.

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 A.2. 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 the two patches and use the original target modality patch as ground-truth. We use L_1 reconstruction loss between the reconstructed patch and original input target modality patch. Note that there is no ground truth for a perfectly aligned RGB modality. Instead, the deformation parameters are optimized using the same L_1 reconstruction loss as an integral part of the SR task.

[1] Optimal blur kernels can be directly estimated as shown in Irani & Michal (2009), and are fully supported by our method as an additional input to the network.

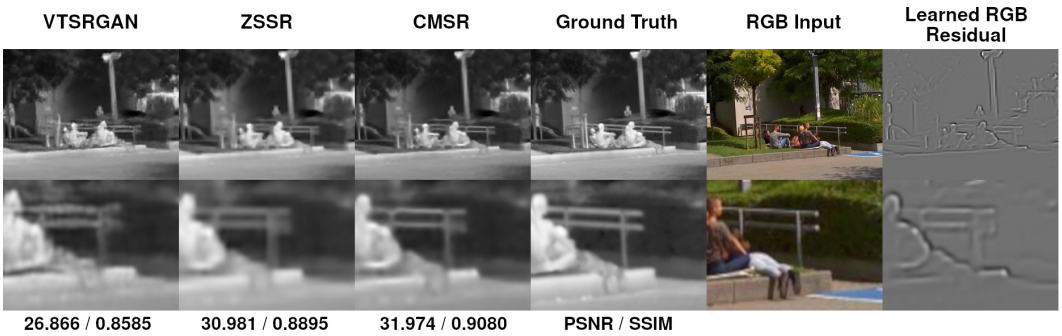

Figure 6: We compare our method to its baseline method, ZSSR (Shocher et al. (2017)), as well as to another cross-modality method, VTSRGAN (Almasri et al. (2018a)) on a **visual-thermal** pair from the ULB17-VT evaluation. On the right, the output of ***Feature-Extractor 2*** (Figure 3) is given as the learned RGB residual which is added to our output. This RGB residual resembles an edge-map; it is artifact-free and contains no unwanted textures

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 parameters, and then feed the LR target modality image and the aligned HR RGB image to the SR network which outputs a HR version of the target modality image.

After the HR target modality image is obtained, we perform two additional refinement operators aimed to further improve our SR results. The first operator, **Geometric Self-Ensemble**, is an averaging technique shown to improve SR results. (Lim et al. (2017); Timofte et al. (2015); Shocher et al. (2017)) The second operator, **Iterative Back-Projection**, is an error-correcting technique that was used successfully in the context of SR. (Glasner et al. (2009); Irani et al. (1991); Shocher et al. (2017))

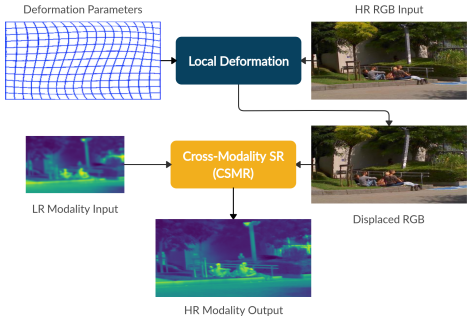

Figure 5: *Inference.* During inference, the learned deformation parameters and the CMSR component are used to upsample the original LR modality input image, guided by the HR RGB input image.

4 Results and Evaluation

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. The full implementation details are found in Appendix A.1.

4.1 Evaluation with State-of-the-arts

Thermal (Infrared). We compared our method to cross-modal state-of-the-art SR methods on visual-thermal pairs. We used the ULB17-VT dataset (Almasri et al. (2019)), consisting of pairs (two examples are shown in Figure 11 in the Appendix, bottom row) that are mostly well aligned. We have included the results of our evaluation in Table 1, showing that our method, despite not being previously trained, beats competing methods. In Figure 6 a visual result from that evaluation is included. In Figure 1 another visual-thermal example is given, taken from a visual-thermal agricultural dataset.

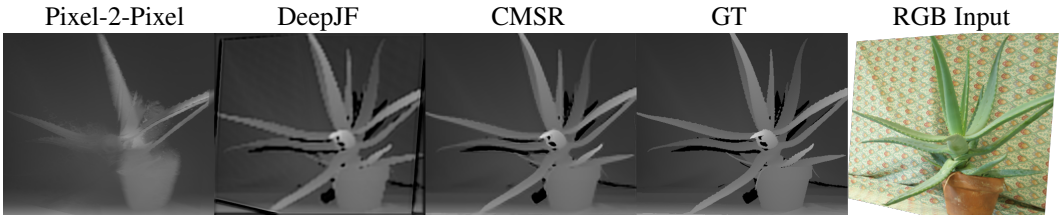

Figure 7: This misaligned **visual-depth** pair is taken from our *Shuffled*-Middlebury dataset evaluation. Compared to cross-modality state-of-the-arts Pixel-to-Pixel SR (de Lutio et al. (2019)) and Deep Joint Filtering (Li et al. (2017)), who struggle to produce a clear result, our method succeeds in this task (32.239 dB PSNR / 0.9403 SSIM) thanks to its alignment capabilities.

NIR (NEAR-INFRARED). In Figure 2 and in the Appendix Figures 10 and 14 we include visual results from our evaluation on the EPFL NIR dataset. (Pierre-francois Laquerre & Duplain (2011)) The conservative approach of our method enables it to surpass state-of-the-art methods (Table 1), even though the competing methods were pre-trained, whereas our method operates on a single input pair without pretraining.

DEPTH. The Middlebury dataset (Scharstein et al. (2002)) contains strongly aligned depth-visual pairs as shown in the Appendix Figure 11 (top row). Multiple angles from different sensor placements are included. To obtain weakly-aligned pairs, we shuffled the pairs together such that the resulting pairs would correspond to sensor misplacements. An example is given in the Appendix Figure 12 (left pair). 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 and presented in Figure 7.

SINGLE MODALITY. We evaluated CMSR against the baseline state-of-the art single modality method, ZSSR. (Shocher et al. (2017)) Our experiment shows that our method leverages the fine details in its RGB input and produces a SR outputs that are closer to a Ground-Truth version, as shown numerically in Table 1 and visually in Figures 6 and in the Appendix. (Figure 10)

Metric	Dataset	ZSSR	VTSRGAN	VTSRCNN	DeepJF	CMSR
PSNR	U-VT	26.789	27.988	27.968	27.036	**29.928**
SSIM	U-VT	0.8567	0.8202	0.8196	0.8410	**0.882**
PSNR	SMB	27.784	27.925	28.189	26.124	**28.652**
SSIM	SMB	0.9140	**0.9547**	0.9386	0.8865	0.9341
PSNR	NIR	28.807	30.665	30.143	27.094	**31.201**
SSIM	NIR	0.8931	0.9005	0.8837	0.8694	**0.9200**

Table 1: We compared CMSR to competing cross-modal SR methods, VTSRCNN and VTSRGAN (Almasri et al. (2018a)) and Deep Joint Filtering (Li et al. (2017)), on the various datasets. (ULB17-VT, *Shuffled*-Middlebury, EPFL NIR) and have taken the mean PSNR / SSIM scores, measured against the modality $4x$ GT versions.

4.2 ANALYSIS

RGB ARTIFACTS. A fusion of multiple image sources, often causes the transfer of unnecessary artifacts. Those artifacts sabotage the image and harm its characteristics. Our method learns only the relevant RGB information that improves SR results; Figures 1, 2, 7 and 14 (Appendix) show cases where the RGB input contains a great amount of textural information, yet our SR output remains texture-free.

LOCAL DEFORMATION ABLATION. As shown in Figure 9, our method deforms the RGB modality, for better alignment, without using any aligned RGB ground-truth images. Our deformation component provides the network the ability to align only details that assist and adhere to the super-resolution task, rather than committing to an image-to-image alignment per se.

To show the necessity of each layer of our coarse-to-fine deformation framework, we evaluated CMSR on a weakly aligned pair, adding one layer at a time and averaging across multiple runs. The

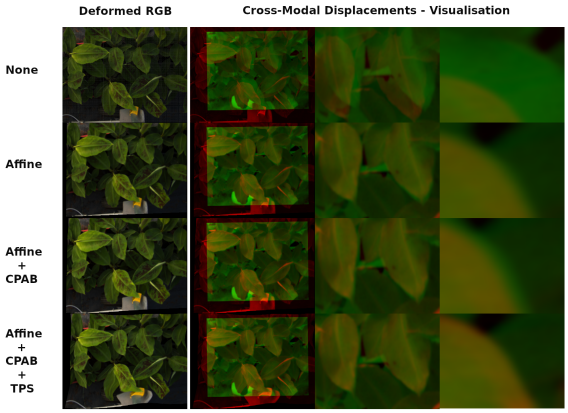

Figure 8: We evaluated CMSR using different transformation layers. In the leftmost column, the resulting deformed RGB image is given. In the other columns we show the resulting alignment, visualized through blending of the R-G (Red-Green) channels of the aforementioned deformed RGB image, together with the Ground-Truth thermal image.

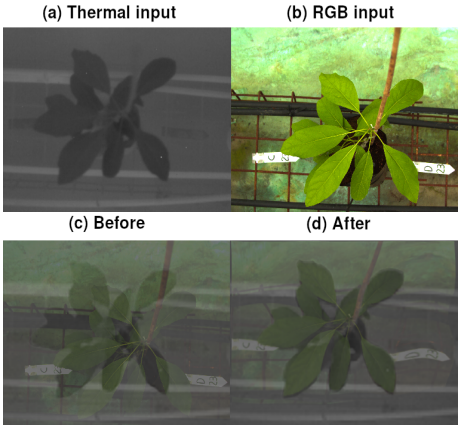

Figure 9: We evaluated CMSR on a severely misaligned visual-thermal pair, (a) and (b), with both global and local displacements. We overlaid the images, once before training (c), and once after training the network (d). CMSR deformed its RGB input on-the-fly to better alignment, relying solely on our SR loss.

results indicate that each layer is necessary and plays a different role in the alignment process as can be seen visually in Figure 8, and numerically in Figure 13 (Appendix). Our goal is not to perform perfect registration between the images, but rather to align only the necessary details to improve the quality of the higher resolution output. Hence, we measure the quality of the alignment through the generated SR result (e.g. in Figures 1 and 7), and not by conventional image registration metrics.

5 Conclusions

We have introduced CMSR, a method for cross-modality super-resolution. Our method is applied on the single input pair and yet outperforms state-of-the-art methods.

Single Pair. As a *self-supervised* method, CMSR requires no pretraining and therefore no training data, a prominent advantage when dealing with scarce and unique modalities. It is operated on the target pair only, and can thus adapt to the specific properties on the given pair. Those possibly unique properties include, among others: (i) the specific cross-modal misalignment that exists within the input pair and (ii) the degree and the manner in which the guiding modality should be incorporated.

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.

Applicability. Since CMSR requires no pretraining, it can be applicable to many different unique pairs, without the need to extensively train a network for each new pair which is unique (in the same sense that was hereby explained).

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, instead of deforming the entire RGB image once, we would like to deform different RGB objects differently, possibly using semantic segmentation, for further enhancement.

REFERENCES

Almasri, Feras, and Debeir. Multimodal sensor fusion in single thermal image super-resolution. *arXiv preprint arXiv:1812.09276*, 2018a.

Almasri, Feras, and Debeir. Rgb guided thermal super-resolution enhancement. In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, pp. 1–5. IEEE, 2018b.

Almasri, Feras, and Olivier Debeir. ULB17-VT. 2 2019. doi: 10.5281/zenodo.2557535. URL https://zenodo.figshare.com/articles/ULB17-VT/7679588.

Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *CoRR*, abs/1904.07523, 2019. URL http://arxiv.org/abs/1904.07523.

Moab Arar, Yiftach Ginger, Dov Danon, Ilya Leizerson, Amit Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. *CoRR*, abs/2003.08073, 2020. URL https://arxiv.org/abs/2003.08073.

Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, 1989. doi: 10.1109/34.24792. URL https://doi.org/10.1109/34.24792.

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)*, pp. 1–4. IEEE, 2016.

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

Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as a learned pixel-to-pixel transformation. *CoRR*, abs/1904.01501, 2019. URL http://arxiv.org/abs/1904.01501.

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. URL http://arxiv.org/abs/1704.06065.

Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Isgum. A deep learning framework for unsupervised affine and deformable image registration. *CoRR*, abs/1809.06130, 2018. URL http://arxiv.org/abs/1809.06130.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. URL http://arxiv.org/abs/1501.00092.

Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *ICCV*, 2009. URL http://www.wisdom.weizmann.ac.il/~vision/SingleImageSR.html.

Huang, Jia-Bin, 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.

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

Irani, Michal, and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graph. Models Image Process.*, 53(3):231–239, April 1991. ISSN 1049-9652. doi: 10.1016/1049-9652(91)90045-L. URL http://dx.doi.org/10.1016/1049-9652(91)90045-L.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. URL http://arxiv.org/abs/1506.02025.

Y Kiran, V Shrinidhi, W Jino 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.

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. URL http://arxiv.org/abs/1704.03915.

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.

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. URL http://arxiv.org/abs/1609.04802.

Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *CoRR*, abs/1710.04200, 2017. URL http://arxiv.org/abs/1710.04200.

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. URL http://arxiv.org/abs/1707.02921.

Emanuele Mandanici, Luca Tavasci, Francesco Corsini, and Stefano Gandolfi. A multi-image super-resolution algorithm applied to thermal imagery. *Applied Geomatics*, Feb 2019. ISSN 1866-928X. doi: 10.1007/s12518-019-00253-y. URL https://doi.org/10.1007/s12518-019-00253-y.

Nasrollahi, Kamal, and Thomas B. Moeslund. Super-resolution: A comprehensive survey. *Mach. Vision Appl.*, 25(6):1423–1468, August 2014. ISSN 0932-8092. doi: 10.1007/s00138-014-0623-4. URL http://dx.doi.org/10.1007/s00138-014-0623-4.

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.

Freifeld O., Batmanghelich K. Hauberg S., and Fisher III. Transformations based on continuous piecewise-affine velocity fields. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7814343.

Noemie Vetterli Pierre-francois Laquerre, Nicolas Etienne and Caroline Duplain. Rgb-nir data. https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html, 2011.

Scharstein, Daniel, and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, April 2002. ISSN 0920-5691. doi: 10.1023/A:1014573219977. URL https://doi.org/10.1023/A:1014573219977.

Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. *CoRR*, abs/1712.06087, 2017. URL http://arxiv.org/abs/1712.06087.

Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. *CoRR*, abs/1609.05396, 2016. URL http://arxiv.org/abs/1609.05396.

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

Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. *CoRR*, abs/1511.02228, 2015. URL http://arxiv.org/abs/1511.02228.

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. URL http://arxiv.org/abs/1804.02900.

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. URL http://arxiv.org/abs/1808.03344.

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. URL http://arxiv.org/abs/1708.09105.

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

A Appendix

A.1 Implementation Details

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 12). 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. (Lai et al. (2017); Wang et al. (2018); Shocher et al. (2017))

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.

A.2 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. To that end, 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. Hence, the goal of the deformation step is not to form a perfect alignment between the images, but rather to allow partial alignment to boost the super-resolution task, where needed. 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. (Jaderberg et al. (2015)) 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, Skafte Detlefsen et al. (2018)), a variant of the original STN layer supporting more flexible and expressive transformations. Our chosen transformation model is **CPAB** (Continuous Piecewise-Affine Based, O. et al. (2017); Skafte Detlefsen et al. (2018)). 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,

Figure 10: We compared CMSR both to its single-modality baseline, ZSSR, (Shocher et al. (2017)) and to competing cross-modality methods, VTSRCNN, VTSRGAN (Almasri et al. (2018a)) and Deep Joint Filtering, (Li et al. (2017)) on the NIR modality, in the task of x 4 SR. Our method, CMSR, is able to produce better super-resolved images visually and numerically, despite not being previously trained.

Figure 11: The visual-depth pairs from the Middlebury dataset (top row) and the visual-thermal pairs from the ULB17-VT dataset (bottom row) show strong multi-modal registration. Under less than optimal imaging conditions, such alignment is hard to achieve.

Figure 12: Two examples of Weakly Aligned modality pairs. To visualize the misalignment, we overlaid them with semi-transparency. Note, the ghosting effect where cross-modal misalignment occurs.

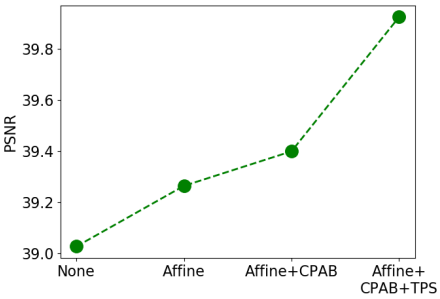

Figure 13: We let CMSR perform 4x SR on a Weakly Aligned visual-thermal pair, with different transformation layers, averaged across 5 runs. The results indicate that each layer contributes to the final PSNR, which can also be seen visually in Figure 8

.

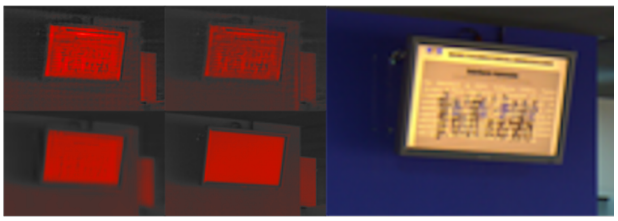

Figure 14: CMSR uses its RGB input conservatively. Compared to VTSRCNN (top left) and VTSRGAN (top right), CMSR avoids introducing noticeable redundant artifacts and textures induced by RGB modality. Ground-Truth (bottom right) is given as reference.

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. (Bookstein (1989)) Our implementation (also taken from Skafte Detlefsen et al. (2018)) 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. (Bookstein (1989)) 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.