Single Pair Cross-Modality Super Resolution
Supplementary Material

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

1. Additional Results

In Figure 1, additional results from our evaluation on the EPFL NIR dataset [4] are included. CMSR surpasses state-of-the-art cross-modal methods, despite the fact that those competing methods were pre-trained extensively on the full dataset.

2. Alternating Scales

2.1. Alternating Scales - Elaboration

Denoting our desired SR ratio (e.g. $2x, 4x$) by r, our network, CMSR, takes a target modality input of size $H\ x\ W$ alongside with an RGB input of size $rH\ x\ rW$, and produces a target modality output of size $rH\ x\ rW$. 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 2 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 1. 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.

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

Table 1: 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.

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.

2.2. 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 2 and Table 1 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 3 and 4, the best PSNR was obtained when $p = 0.3$, which starts decaying when $p > 0.3$. We notice that $0.5 > p > 0$ 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.

We have introduced CMSR, a method for cross-modality super-resolution.

Novelty CMSR presents a novel way to tackle cross-modality SR; it achieves state-of-the-art results, qualitatively (visually) and quantitatively, using a minimalistic,

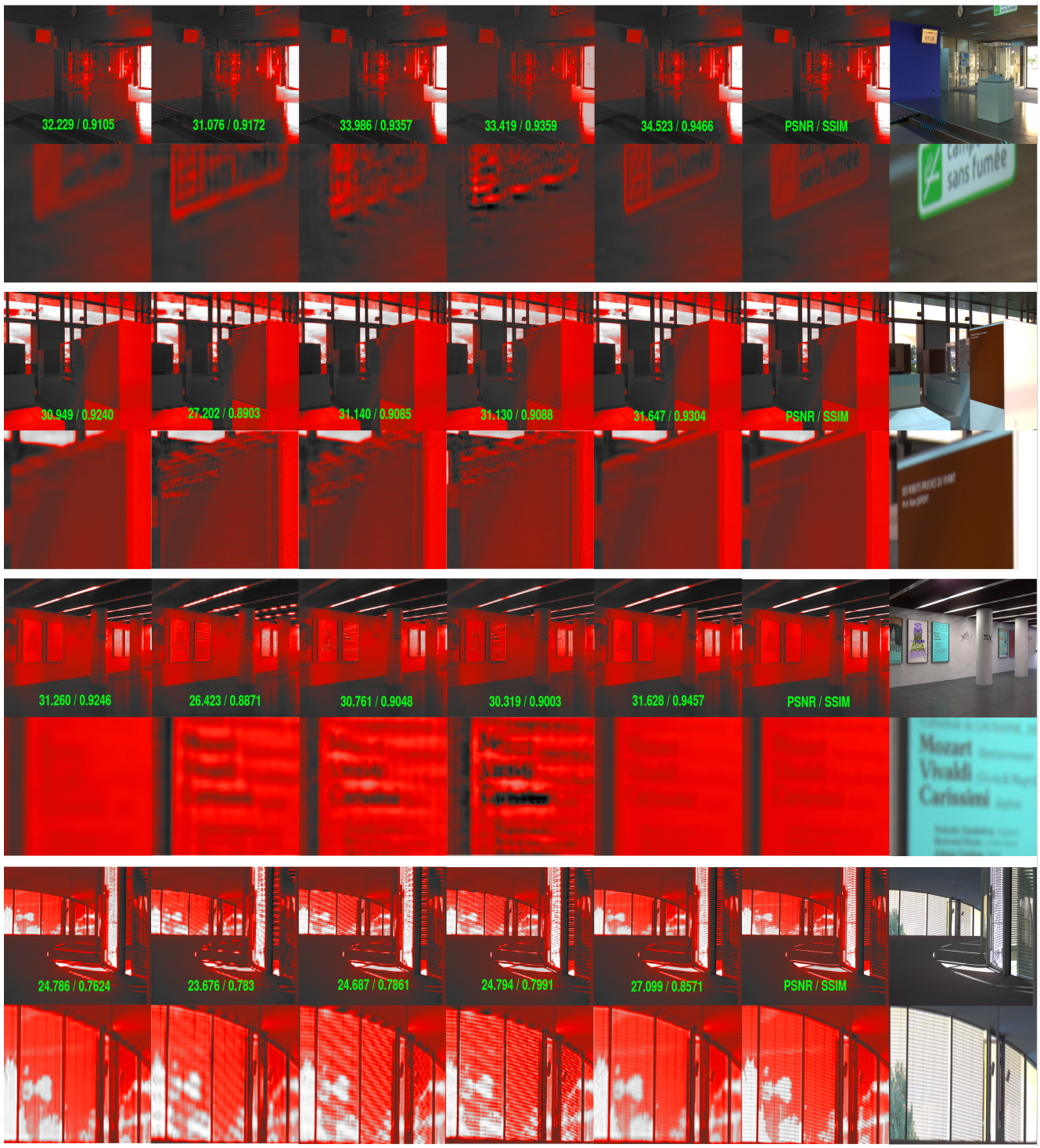

Figure 1: We compared CMSR both to its single-modality baseline, ZSSR, and to competing cross-modality methods VTSRCNN, VTSRGAN and Deep Joint Filtering, on the NIR modality, in the task of x 4 SR. Our method, CMSR, is able to produce better super-resolved images visually and numerically, despite not being previously trained.

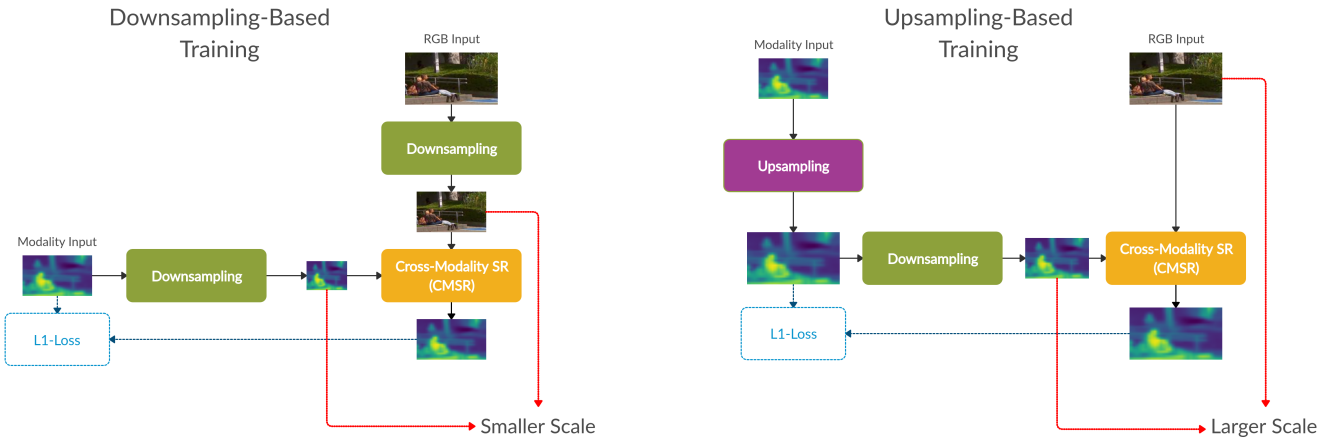

Figure 2: The difference between the two training schemes lies in the scale CMSR (in Orange) operates on. The two schemes start with the exact same input pair, but in the Upsampling-Based training scheme (right), CMSR is fed inputs of larger scale. This scale difference is also explained in Table 1.

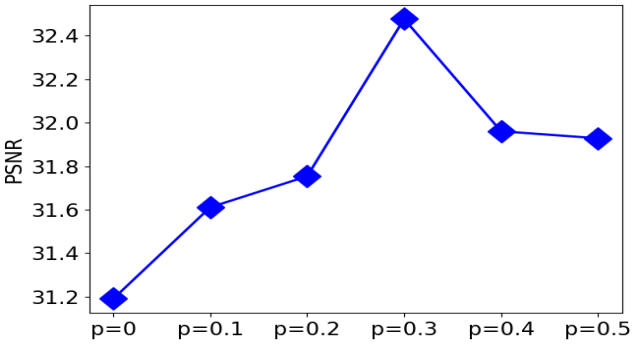

Figure 3: We evaluated CMSR using different alternation probabilities. Namely, we trained it using the Upsampling-Based training scheme (Figure 2) in fraction p iterations, and using the Downsampling-Based scheme in the remaining fraction $1-p$. We averaged this experiment across multiple runs. According to the results, $p = 0.3$ yields the best PSNR (32.476 *dB*). This can be also seen visually, in Figure 4

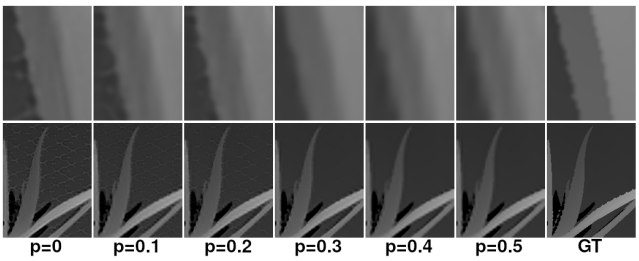

Figure 4: We compare two patches taken from the Alternating Scales ablation study results, summarized in Figure 3. According to our experiment, the best SR result is obtained when using $p = 0.3$ as the alternation probability.

easy to implement architecture, applied directly to any modality pair without pretraining.

Single Pair. As a *self-supervised* method, CMSR no training data, a prominent advantage when dealing with scarce and unique modalities. It adapts to the specifics of the given input pair, including 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.

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

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

[2] Michal Irani Assaf Shocher, Nadav Cohen. "zero-shot" super-resolution using deep internal learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Yijun Li, Jia-Bin Huang, Ahuja Narendra, and Ming-Hsuan Yang. Deep joint image filtering. In *European Conference on Computer Vision*, 2016.

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