CIPPSRNet: A Camera Internal Parameters Perception Network Based Contrastive Learning for Thermal Image Super-Resolution

Kai Wang¹,², Qigong Sun¹*, Yicheng Wang¹, Huiyuan Wei², Chonghua Lv², Xiaolin Tian², Xu Liu²
¹SenseTime Research, ²Xidian University

kiwi0128@qq.com, sunqigong@sensetime.com, wang156757299@gmail.com

Abstract

Thermal Image Super-Resolution (TISR) is a technique for converting Low-Resolution (LR) thermal images to High-Resolution (HR) thermal images. This technique has recently become a research hotspot due to its ability to reduce sensor costs and improve visual perception. However, current research does not provide an effective solution for multi-sensor data training, possibly driven by pixel mismatch and simple degradation setting issues. In this paper, we proposed a Camera Internal Parameters Perception Network (CIPPSRNet) for LR thermal image enhancement. The camera internal parameters (CIP) were explicitly modeled as a feature representation, the LR features were transformed into the intermediate domain containing the internal parameters information by perceiving CIP representation. The mapping between the intermediate domain and the spatial domain of the HR features was learned by CIPPSRNet. In addition, we introduced contrastive learning to optimize the pretrained Camera Internal Parameters Representation Network and the feature encoders. Our proposed network is capable of achieving a more efficient transformation from the LR to the HR domains. Additionally, the use of contrastive learning can improve the network’s adaptability to misalignment data with insufficient pixel matching and its robustness. Experiments on PBVS2022 TISR Dataset show that our network has achieved state-of-the-art performance for the Thermal SR task.

1. Introduction

In the past, photographic products that were based on RGB technologies set up a milestone in the visual developments of humans. The ability of RGB cameras to capture vibrant colours and detailed information has propelled them to rapid popularity. However, it is virtually impossible for RGB cameras to obtain luminescent images that are dense in high-frequency information under unfavorable illumination conditions. In this case, thermal cameras can be used to detect thermal radiation emitted by objects in the long-wave infrared (LWIR) spectral range, and reveal objects that are invisible to human perception. Thermal cameras are passive sensors that detect infrared light at temperatures greater than absolute zero [8]. Therefore, they can be used in military [9], agriculture [33], maritime surveillance [13], medical [25], industry [24], urban development [7], etc. Especially, the recent outbreak of the COVID-19 epidemic has endowed thermal cameras with a new application - body temperature measurement [31]. However, the expensive materials limit the popularity of upmarket thermal cameras. Meanwhile, low-cost thermal cameras generally face the physical defect in accuracy and noise interference. It becomes very difficult to improve the resolution of thermal cameras from hardware-level. So, the software-level algorithm becomes a new option to improve the image resolu-
SISR is the task of upscaling a LR image with resolution enhancement. SRCNN [4] proposes a natural image super-resolution model based on convolutional neural network (CNN), which provides a new guidance for natural image super-resolution. Furthermore, to perform an advanced visual perception, recent works including [20] [18] [32] [16] [19] designed more refined network structures. Some work such as [1] [35] [34], estimate degradation patterns in real-world scenarios in order to solve complex degradation problems. Recently, some SISR methods for thermal images have also been explored. Some models [29] [3] [23] [21] are dedicated to designing special network structures according to the characteristics of thermal images. [23] solves the alignment problem of different source sensor data, which makes it more applicable to real-world scenarios. These Thermal Image Super-Resolution (TISR) methods not only improve the visualization of thermal images, but also are help for these tasks such as thermal-based objects detection [17], image segmentation [26], and recognition [38]. However, existing TISR methods still face many challenges due to the great differences in the imaging principle of different natural images.

Due to differences in imaging perspectives, the LR-HR data from different thermal cameras usually need to be aligned. As shown in Figure 2, the pixels of a real LR-HR thermal imaging data pair do not be matched one-to-one. Obviously, if we get the corresponding pixels pairs, the network can learn the mapping between LR images and HR images. However, the limited performance of the alignment algorithm could not always promise data adaptation. Because we usually cannot obtain pixel pairs with accurate correspondences, the modeling and training process of thermal image super-resolution models is quite complicated, and the performance of the models is not ideal. Moreover, the current industrial standards for thermal imaging cameras are not uniform. The internal parameters of thermal imaging cameras vary greatly due to the different thermal imaging materials and processing methods used by different manufacturers, which greatly affects the generalization performance of TISR algorithm. Inevitable parameter differences cause the imaging data to conform to different data distributions, the HR and LR data are not in the same data domain. Based on these analyses, for TISR, how to effectively realize domain transformation is an urgent problem. Therefore, we propose a new TISR model - Camera Internal Parameters Perception Super-resolution Network (CIPPSRNet) to solve the above problem. We introduce contrastive learning to improve the adaptability of the network to misaligned data, which reduces the distance between features of misaligned data in the feature space. At the same time, we design the Camera Internal Parameters Representation network (CIPRNet), which models the camera-internal representation as a feature vector which is incorporated into LR features. The CIPRNet learns a transform mapping from LR to HR domain with camera internal information which can be used for SR tasks with different thermal cameras. We validated the effectiveness of proposed model by training and testing on disjoint dataset. We compared the obtained SR results with other state-of-the-art visible SR algorithms [20] and thermal SR methods [28] [23]. The key contributions of this work can be summarized as follows:

- This work proposes a novel network for TISR task named Camera Internal Parameters Perception Super-resolution Network (CIPPSRNet), which is capable of achieving a more efficient transformation from the LR to the HR domains and adapting multi-sensor data.
- A novel network called Camera Internal Parameters Representation Network (CIPRNet) is proposed to achieve the CIP representation adapted to domain transformation.
- A contrastive learning-based method is proposed to train encoder that are unaffected by registration, which make this method more suitable for real scenarios.
- Our model participated in the PBSD2022 TISR challenge and achieved a relatively impressive result in Track 2. The PSNR/SSIM on the test dataset is 29.68/0.7886.

In the rest of the paper, Section 2 provides an overview of SR works on both thermal and visible domains. Section 3 describes the architecture and methodology of the proposed model. Section 4 discusses the experimental analysis, including details, dataset description, ablation studies, quantitative and qualitative analysis. Section 5 presents the conclusion.

2. Related Works

2.1. Visible image super-resolution

The study of visible image super-resolution (VISR) has been lasted for decades. The traditional VISR methods are generally model-based. In recent years, the learning-based methods, especially CNN-based methods, have become more recognized with their great and impressive performance. The first CNN-based method SRCNN [4] extracts features from the LR images. It learns the mapping between LR and HR features and reconstructs HR images from them. Inspired by [4], there were great quantity of work about CNN-based SR methods, such as FSRCNN [5] EDSR [20] VDSR [15], etc. Some of them have exploited lightweight framework for a faster inference, such as ECBSR [39], AdderSR [32] and ClassSR [16].
Figure 2. Alignment of Thermal HR-LR data pairs in real scenarios. Top left- HR Reference Image. Top right- Non-aligned LR Image. Bottom left- HR Reference Image. Bottom right- Aligned LR Image. The yellow dots in the figure represent the feature points that match the two images. The alignment allows the two images to become more pixel-level correspondent, which is very beneficial for the training of SR models.

Some of them based on the unsupervised learning and contrastive learning, such as DRN [10], DASR [34]. Towards real-world VISR, there were some works to deal with the complex degeneration problem such as [34]. Some recent work combined with vision transformer (ViT) [6], such as SwinIR [19].

2.2. Thermal image super-resolution

The success of deep learning for VISR has promoted the related researches on TISR. Some thermal image datasets [37] [14] [22] have been proposed in recent years, however, these datasets are dedicated to address objects detection and tracking problem, which cannot be reasonably applied to SR tasks. Thus, a dataset containing 101 thermal images was proposed in [30] for solving SR challenge. To solve the issue of insufficient data, [27] proposed a CycleGAN architecture for TISR, along with a large thermal image dataset. As a more efficient computing process, the TherISuRNet [3] introduced low-frequency feature extraction module and asymmetric high-frequency feature extraction module to TISR task. [23] introduced a Channel Splitting-based Convolutional Neural Network (ChasNet) for thermal image SR eliminating the redundant features in the network.

2.3. Contrastive learning

Contrastive learning has experienced considerable progress in self-supervised representation learning during recent years [11] [2]. The main idea is to view each sample as a distinct class to train the distinguishing ability of model. For the SR task, [34] handled various degradations of blind SR by distinguishing degraded representations between different images by contrastive learning. [36] extracts texture features of positive and negative samples by wavelet transform to train a mighty contrast discriminator, allowing contrastive learning to be better used for low-level vision tasks like SR. In this work, feature representations of random mismatched pixel were generated from a LR image and the corresponding HR image, and contrastive learning was performed to obtain an alignment-independent encoding feature. This allows the SR network to tolerate the deleterious effects of mis-alignment, as shown in Figure 1.

3. Proposed Method

We proposed a CNN-based model for PBVS2022 TISR track-2 competition. The challenge of this task requires conversion from LR to HR thermal images with a scale of ×2. The difficulty is that the LR-HR data come from a variety of thermal cameras. As shown in Figure 4, the SR framework we proposed consists of three parts:

- Camera Internal Parameters Representation Network (CIPRNet)
- U-shaped Perception Network (UPN)
- Cycle Perception Network (CPN)

Both the UPN and CPN contain CIPRNet. Through UPN, a LR image is translated to ×2 SR image. The generated SR image from the LR image sent to CPN is used to solve the ill-posed problem. As shown in Figure 4, in UPN, the LR image is first implemented a bicubic interpolation, and the output \( I_{LBIC} \) has the same size as the HR image. The equation is:

\[ I_{LBIC} = F_{BICUBIC}(I_{LR}), \]

where \( F_{BICUBIC}(I_{LR}) \) is the bicubic interpolation function, \( I_{LR} \) represents the input LR thermal image.

Then \( I_{LBIC} \) will get the LR CIP representation \( X_{LCIP} \) through CIPRNet, whose backbone network is ResNet18.
And the LR encoding feature $F_{EN}$ will be obtained by sending $I_{LBIC}$ to U-shaped encoding network (including two 3x3 convolutional layers, three DownSampling blocks (DSB)), respectively. The equation is:

$$
\begin{align*}
X_{LCIP} &= F_{CIPRNet}(I_{LBIC}) \\
F_{EN} &= F_{U-ENCODER}(I_{LBIC}),
\end{align*}
$$

where $F_{CIPRNet}$ is the CIPRNet function, $F_{U-ENCODER}$ is the function of U-shaped encoder.

To enable the $F_{EN}$, the LR encoding feature, contains the camera internal information, $F_{EN}$ and $X_{LCIP}$ are fed into the Representation Perception Module (RPM), which finally outputs the LR coding feature $F_{CEN}$ with camera internal information. The equation is:

$$
F_{CEN} = F_{CONV}(F_{DSB}(F_{EN}), F_{RESHAPE}(F_{FC}(X_{LCIP}))),
$$

where $F_{FC}$ is the function of fully connection layer. $F_{RESHAPE}$ reshape the 1-d vector to a convolutional kernel. $F_{DSB}$ is the function of DSB. $F_{CONV}(F, K)$ represents the convolution of feature map $F$ via a kernel $K$.

Finally, the LR coding feature $F_{CEN}$ is passed through the decoder of the U-shaped network (including one 1x1 convolutional layer 3 upsampling modules (USB)) and the reconstruction module (including two 3x3 convolutional layers), and then long-skip connections are employed to guarantee the basic performance of SR image $I_{SR}$. The equation is:

$$
I_{SR} = I_{LBIC} \oplus F_{REC}(F_{U-DECODER}(F_{CEN})),
$$

where $F_{U-DECODER}$ is the function of U-shaped network decoder, $F_{REC}$ is the function of reconstruction module.

In the training phase, the HR image $I_{HR}$ is also input into CIPRNet to obtain the HR camera internal parameters representation. $I_{SR}$ is fed into the CPN (including encoder, representation perception module and decoder) to finally obtain the generated LR image $I_{LR}'$, and the L1 norms of $I_{LR}'$ and $I_{LBIC}$ are calculated as the cycle loss and optimized to limit the search space of the network.

### 3.1. Camera Internal Parameters Representation (CIPR) Learning

The goal of CIPR learning is to learn to extract differentiated representations in LR-HR thermal images in a supervised manner, which is performed before the overall network training. As shown in Figure 3, the LR thermal images from different cameras are fed into a classification framework for binary supervised learning, and eventually the feature vectors (CIP representations) learned by encoder can represent the camera internal parameters. When training the overall network, the network parameters obtained from CIPR learning are used as the initial parameters of CIPRNet. CIPRNet also participates in the update and optimization of the overall network to make its CIPR more capable.
3.2. Contrastive Learning adapting misaligned data

Pixel misalignment data can push the training of the SR model to the wrong direction. To reduce the effect of misalignment data, we introduce a training scheme based on contrastive learning. As shown in Figure 1, for each batch of data, the $I_{LR}$ is passed through U-shaped encoder, CIPRNet and U-RPM to output an encoded feature vector $F_{LR}$. Meanwhile, $I_{LR}$ is randomly rotated and transformed into three simulated misaligned data, which are also passed through U-Encoder, CIPRNet and U-RPM to generate three negative sample feature vectors $F^1_{LR}$, $F^2_{LR}$, $F^3_{LR}$. Since $F_{LR}$, $F^1_{LR}$, $F^2_{LR}$, $F^3_{LR}$ come from the same LR thermal image, the Euclidean distance among them on the feature space should be smaller. In other words, we can consider $F^1_{LR}$, $F^2_{LR}$, $F^3_{LR}$ as the positive samples of $F_{LR}$, and by the same method we can get $F^1_{HR}$, $F^2_{HR}$, $F^3_{HR}$ as their negative samples.

Positive and negative samples are back-propagated by contrastive loss, which ultimately enables the network to adapt different misalignment data and improve the network performance. The expression for the contrastive loss is:

$$L_{con} = \sum_{i=1}^{B} -\log \frac{\exp(F_{LR_i} \cdot F^1_{LR_i} / \tau)}{\sum_{j=1}^{3} \exp(F_{LR_i} \cdot F^j_{LR_i} / \tau)},$$

where $F_{LR_i}$ represents the feature vector generated by the $i$th $I_{LR}$ in the current batch. $F^1_{LR_i}$ is one of the corresponding positive samples, $F^3_{LR_i}$ is one of the corresponding negative samples. $\tau$ is a hyper-parameter.

3.3. Loss functions

The weights of CIPPSRNet was updated by pixel-level $L_1$ loss, SSIM loss and cyclic loss between SR and HR image in each training iteration. The total loss function is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{ssim} + \lambda_3 \mathcal{L}_{cycle},$$

$$\mathcal{L}_{ssim} = \frac{1}{N} \sum (1 - SSIM(I_{SR}, I_{HR})),$$

$$\mathcal{L}_{cycle} = \frac{1}{N} \sum (||I_{LR} - I'_{LR}||_1),$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$ is the weight coefficient of different loss items. $N$ is the batch size and SSIM is the function of structural similarity.

After updating the overall parameters using $\mathcal{L}$, the U-encoder, CIPRNet and U-RPM parameters are then updated using $L_{con}$ shown in Equation 5 to reduce the negative impact of misaligned data on the network.
Table 1. The quantitative comparison of the proposed network with other TISR algorithms on validation dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>29.11</td>
<td>0.7428</td>
</tr>
<tr>
<td>EDSR [20]</td>
<td>29.24</td>
<td>0.7278</td>
</tr>
<tr>
<td>TherISuRNet [3]</td>
<td>29.53</td>
<td>0.7686</td>
</tr>
<tr>
<td>PBVS2021-Winner [28]</td>
<td>29.60</td>
<td>0.7702</td>
</tr>
<tr>
<td>ChaSNet [23]</td>
<td>29.63</td>
<td>0.7583</td>
</tr>
<tr>
<td>Ours</td>
<td>29.85</td>
<td>0.7706</td>
</tr>
</tbody>
</table>

Table 2. The quantitative comparison of the proposed network with other TISR algorithms on PBVS2022 test dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIPPSRNet</td>
<td>29.85</td>
<td>0.7706</td>
</tr>
<tr>
<td>w/o CIPPRNet</td>
<td>29.21</td>
<td>0.7504</td>
</tr>
<tr>
<td>w/o CPN</td>
<td>29.84</td>
<td>0.7628</td>
</tr>
<tr>
<td>w/o Contrastive Learning</td>
<td>29.73</td>
<td>0.7609</td>
</tr>
</tbody>
</table>

Table 3. Ablation study and further investigation experimental results

4. Experiments

4.1. Dataset

Our model is mainly trained based on a benchmark thermal image dataset [1]. This dataset provides 951 images in the training set and 50 in the validation set as well as 10 in the test set. Since the model is mainly dedicated to solving the cross-domain SR task, we aim to accomplish x2 upscaling from the LR dataset (i.e. Axis (MR) dataset) to the corresponding HR dataset (i.e. Flir (HR). The content captured by the corresponding images in dataset is consistent, but there is still a pixel-level misalignment due to the difference in perspective. Hence, we use the combination of classic and DNN methods to register the images. However, the registered images still do not achieve completely pixel-wise accuracy due to the inherent shortcomings of geometric correction. To eliminate the impact of misalignment, we use data augmentation and contrastive learning to make the network more robust in real scenarios.

4.2. Training Setup

The proposed networks are trained with Adam optimizer on the default $\beta$ values and with the learning rate of $5 \times 10^{-4}$ which is decayed by cosine decay. We have trained the proposed modules up to $2 \times 10^3$ number of epochs with a batch size of 32. In training, the 96 × 96 patch from HR images and its corresponding patch from LR images are extracted randomly and then augmented using random horizontal flipping and rotation with 0°, 90°, 180° and 270° operations and randomly adding gaussian noise. The weighting coefficients $\lambda_1$, $\lambda_2$ and $\lambda_3$ associated to Equation 6 are set to 10, 1 and 1.

4.3. Quantitative Evaluation

Our proposed method, CIPPSRNet, was quantitatively evaluated by comparing it with other existing state-of-the-art methods in terms of PSNR and SSIM. On the validation set of 50 images, we tested the performance of Bicubic, EDSR [20], TherISuRNet [3], PBVS2021-Winner [28], and ChasNet [23], as shown in Table 1. We choose the central 50% of the SR results for validation to eliminate the effect of black edges due to alignment, which would make the results more accurate. As can be seen, our proposed architecture outperforms both the previous algorithms and bicubic interpolation in terms of performance. Besides, Table 2 shows the results of our proposed method in the PBVS2022 TISR challenge with other teams in 10 images test set.

4.4. Qualitative Evaluation

In Figure 5, we also demonstrated the ability of various approaches to reconstruct perceptually convincing images. It can be noticed that our proposed network also achieves qualitative improvement over other state-of-the-art algorithms. As can be seen in the Figure 5, our proposed network brings rich structural detail information of SR image closer to real HR image than the previous methods. Its ability to recover some of the missing texture properties of the LR image validates the effectiveness of our method in the cross-sensor TISR task.

4.5. Ablation Study

In this section, we compare the impact of three modules on network performance: 1) CIPPRNet, 2) contrastive learning strategy, 3) Cycle Perceptual Network. While ensuring that the rest of the network modules were set up consistently and the training environment was consistent, we removed one module each and then evaluated the network by PSNR/SSIM on validation set. The experimental results in
Table 3 demonstrate that all our proposed methods can improve the performance of the outputs.

Further analysis, CIPRNet has a significant impact on network performance, which proves that the camera internal parameter representation information from CIPRNet is extremely useful. Even though we have almost eliminated the bad impact from misalignment data by multiple registration methods, contrastive learning still played a crucial role in our network.

5. Conclusion

In this paper, we proposed a novel thermal image super-resolution network named CIPPSRNet, which could also be applied to other cross-camera super-resolution tasks. Compared with existing state-of-the-arts thermal and natural SR methods, our method simultaneously improved the visual experience and quantitative result for LR thermal images in the cross-camera setting. Furthermore, contrastive learning makes our proposed network easier to adapt misaligned data in real applications. CIPRNet extracts the camera internal parameter representations of different camera that improving the robustness of our network.

Acknowledgment

The authors would like to thank anonymous reviewers for their insightful suggestions. Our work is supported by Xidian University - SenseTime Joint Lab of Smart Health Care and the National Natural Science Foundation of China (No. 61977052).

References


computer vision and pattern recognition, pages 4681–4690, 2017. 2


