electronics

MDPI

Article

An Effective Multi-Task Two-Stage Network with the Cross-Scale Training Strategy for Multi-Scale Image Super Resolution

Jucheng Yang [1,*,†,‡], Feng Wei [1,2,‡], Yaxin Bai [1], Meiran Zuo [1], Xiao Sun [1] and Yarui Chen [1]

[1] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China; wf_xiaobo@163.com (F.W.); m15731106512@163.com (Y.B.); zuomeiran@mail.tust.edu.cn (M.Z.); sunx@mail.tust.edu.cn (X.S.); yrchen@tust.edu.cn (Y.C.)
[2] College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300457, China
* Correspondence: jcyang@tust.edu.cn; Tel.: +86-022-6060-0978
† Current address: Tianjin Economic and Technological Development Zone, No.9, 13th Street, Tianjin 300457, China.
‡ These authors contributed equally to this work.

Abstract: Convolutional neural networks and the per-pixel loss function have shown their potential to be the best combination for super-resolving severely degraded images. However, there are still challenges, such as the massive number of parameters requiring prohibitive memory and vast computing and storage resources as well as time-consuming training and testing. What is more, the per-pixel loss measured by L_2 and the Peak Signal-to-Noise Ratio do not correlate well with human perception of image quality, since L_2 simply does not capture the intricate characteristics of human visual systems. To address these issues, we propose an effective two-stage hourglass network with multi-task co-optimization, which enables the entire network to focus on training and testing time and inherent image patterns such as local luminance, contrast, structure and data distribution. Moreover, to avoid overwhelming memory overheads, our model is capable of performing real-time single image multi-scale super-resolution, so it is memory-friendly, meaning that memory space is utilized efficiently. In addition, in order to best use the underlying structure and perception of image quality and the intermediate estimates during the inference process, we introduce a cross-scale training strategy with 2×, 3× and 4× image super-resolution. This effective multi-task two-stage network with the cross-scale strategy for multi-scale image super-resolution is named EMTCM. Quantitative and qualitative experiment results show that the proposed EMTCM network outperforms state-of-the-art methods in recovering high-quality images.

Keywords: CNN; per-pixel loss; HVS; EMTCM; multi-task co-optimization; cross-scale training

Citation: Yang, J.; Wei, F.; Bai, Y.; Zuo, M.; Sun, X.; Chen, Y. An Effective Multi-Task Two-Stage Network with the Cross-Scale Training Strategy for Multi-Scale Image Super Resolution. *Electronics* **2021**, *10*, 2434. https://doi.org/10.3390/electronics10192434

Academic Editors: Otoniel Mario López Granado and Stefanos Kollias

Received: 19 August 2021
Accepted: 30 September 2021
Published: 7 October 2021

1. Introduction

Image super-resolution (SR) is the process of recovering a high-resolution (HR) image from a low resolution (LR) image. This important computer vision task has found many real-world applications, including medical imaging [1–3], surveillance and security [4–6]. Despite the considerable success in CNN-based SR methods [7–14], they face the following issues:

(1) How HR images are degraded is unknown and the degradation process can be affected by various factors (e.g., defocusing, compression artefacts, anisotropic degradation, sensor and speckle noise). This makes learning to map from LR to HR images an ill-posed problem, since there exist infinitely many HR images that can be downscaled to the same LR image [15]. In other words, the number of possible functions mapping LR to HR images can be extremely large, thus severely limiting learning performance. Moreover, the solution space increases exponentially with the increase of scale factors. High magnifications aggravate the problem of networks' time and resource consumption. Although

generative adversarial networks (GANs) can improve the performance of single image super-resolution (SISR) [16], they suffer from training instability. Efforts have been devoted to stabilizing GAN training, such as introducing various regularization terms [17–19], handling complex losses [16,20,21] and adding training recipes [22]. However, these models usually have an enormous number of parameters, require vast computing and storage resources and have long training and testing time.

(2) Despite the successful collaboration between the CNN and per-pixel loss (L_2), there exist some limitations. The L_2 measure correlates poorly with image quality perceived by a human observer [23]. The use of L_2 assumes that the impact of noise is independent of the local characteristics of the image. However, the sensitivity of human visual systems (HVS) [24] to noise depends on local luminance, contrast and structure [17]. Moreover, the L_2 loss should be considered under the assumption of white Gaussian noise, but the validity of this assumption is generally unknown.

To resolve the above issues, we develop an effective multi-task two-stage network with a cross-scale training strategy for multi-scale image SR. The proposed network is called EMTCM, which is memory-friendly since memory space is utilized efficiently. Our EMTCM network is able to gradually extract features, thereby decreasing channel dimensions and reducing the time of training and testing, while achieving real-time image SR. The contributions of this work are three-fold, summarized as follows.

First, We propose a simple yet effective SISR model, which is scalable and has the best combination in terms of functionality for image SR. Our EMTCM network goes beyond the widely used bicubic degradation assumption. It can be readily implemented in real time and works for multi-scale SR in a two-stage hourglass model, thus making a substantial step forward in developing a CNN-based super-resolver for real-world applications.

Second, we propose a multi-task co-optimization method without calculating pixel-wise L_2 loss for training and optimizing the entire network. Specifically, this method focuses on local huminace, contrast, structure and fitting data distribution. It not only improves the performance of the quantitative metric, Peak Signal-to-Noise Ratio (PSNR), but also produces visually desirable results on LR images.

Third, with a memory-friendly encoder, a recovery decoder and the cross-scale training strategy, the proposed EMTCM network can avoid large amounts of memory overheads and attain strong performance. As the memory-friendly encoder can gradually learn high-frequency feature maps, channel dimensions are decreased. In order to solve problems of the local receptive field, we impose a sufficient number of convolution layers with a long skip connection operator to capture an input image's high-level information.

2. Related Works

CNN-based methods: SRCNN [25] was the first work to use CNN to solve SISR, which has a three-layer. C.Dong et al. [26] explored the impact of depth on SISR and empirically showed that the difficulty of training deeper model hinders the performance improvement of CNN super-resolvers. To address those problems, VDSR [7] with residual learning strategy was proposed. At the same time, VDSR can handle multi-scales super-resolution in a single model. Furthermore, Zhang et al. [27] showed that CNN-based methods mainly model the prior information and they empirically demonstrated that they can handle multi-scales super-resolution in a single model. Nevertheless, in spite of the achieved superior performance compared to the state-of-the-art methods, those methods not only suffer from overwhelming memory overheads but also hinder the effective expansion of receptive field.

To avoid overwhelming memory overheads, some researchers directly applied an end-to-end network that took original LR images as input and adopted an upscaling operation at the end of the network. A deconvolution layer at the end of the network to perform upsampling was adopted by Dong et al. [28]. An efficient sub-pixel convolution was proposed by Shi et al. [29] to upscale the LR feature maps into HR images. A Laplacian pyramid SISR network (LapSRN) [13] that took an original LR image as input and

progressively predicted the sub-band residuals with transposed convolutions in a coarse-to-fine manner. To improve the HVS at a significant scale factor, a generative adversarial network [30] based super-resolution (SRGAN) method was proposed by Ledig et al. [9]. However, CNN-based methods or GANs and SISR suffer from notoriously difficult super-resolution on the problem of overwhelming memory overheads, training instability.Various CNN-based methods for SISR are always used to solve the widely-used settings of bicubic degradation, thus neglecting their limited applicability for practical scenarios. However, model-based optimization framework [31–33] can go beyond bicubic degradation. For example, Zhang et al. proposed [33], which can address the widely-used Gaussian degradation as in [34]. However, the above methods cannot solve different scales that result in difficult training and testing. Thus, it is desirable to learn a single and real-time SISR model which can address multi-scales. This paper attempts to give a positive answer.

Pixel-wise loss: The loss layer, despite being the effective driver of the network's learning, has received little attention from the image processing research community. The choice of the cost function generally defaults to the squared L_2 norm of the error [25,35,36]. This is understandable, considering many desirable properties that this norm possesses. There is also a less well-founded but just as relevant reason for the continued popularity of L_2, that is, standard neural network packages, such as Caffe [37], only offer implementations for this metric. However, with the development of other frameworks (e.g., Pytorch and Tensorflow), we can impose a multi-task co-optimizing strategy to further improve the performance of SISR. Wang et al. [38] apply SSIM, observing that the scale at which local structure should be analyzed is a function of factors such as the image-to-observer distance. To account for these factors, they propose MS-SSIM, a multi-scale version of SSIM that weighs SSIM computed at different scales according to the sensitivity of the HVS. Experimental results show the superiority of SSIM-based indexes over L_2. However, the $SSIM$ does not provide quantitative improvement. We thus design a multi-task training strategy that can meet the requirement of quantitative and qualitative improvement.

3. Methods

Our single image super-resolution is composed of a two-stage hourglass network, including a memory-friendly encoder and recovery decoder. Here we impose a deconvolution layer that achieves $2\times$, $3\times$, $4\times$ super-resolution, yet our methodology can be applied for higher upsampling goals. As aforementioned, we bootstrap the super-resolution process by a multi-task co-optimizing method to focus on different inherent patterns of an image such as local luminance, contrast, structure and fitting the ground-truth data distribution. Meanwhile, we impose cross-scale training strategy to improve performance further. Our memory-friendly encoder approach progressively extracts high-level feature maps and decreases the channel dimension, thus avoiding the overwhelming memory overheads.

3.1. EMTCM

Inspired by [28], we choose the CNN-based method as our basic approach. However, in the supervised model based on traditional image SR such as SRCNN, the image needs to be processed and interpolated to the desired size, and SRCNN model learns mapping in high dimensional space, leading to time-consuming and computing-consuming problem. Therefore, we create an end-to-end EMTCM to solve aforementioned weakness. As shown Figure 1, our overall network is a two-stage hourglass architecture, including a memory-friendly encoder and a recovery decoder. What is more, we impose a number of residual blocks with long skip connection in order to solve the problem of CNN's local receptive field and enhance the capability of model feature representation.

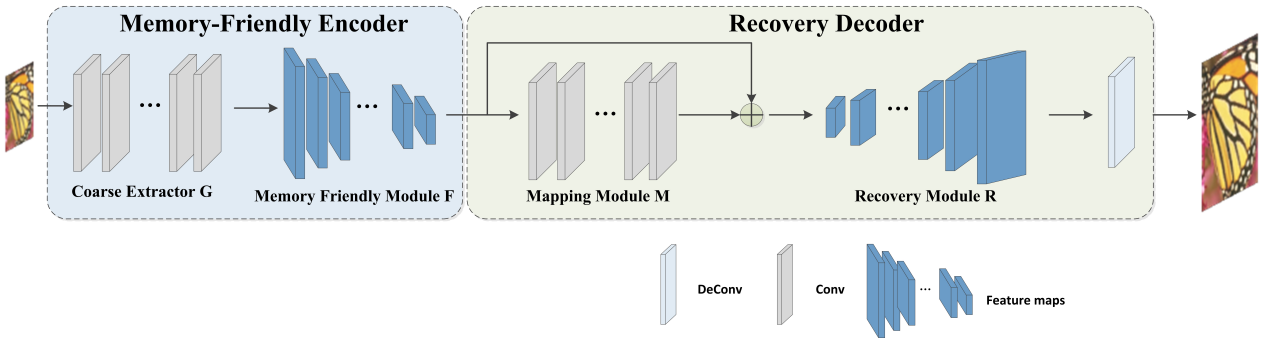

Figure 1. Overall framework of proposed two-stage EMTCM methods. The architecture of EMTCM is composed of two networks, Memory-Friendly Encoder and Recovery Decoder.

A Memory-Friendly Encoder: as shown in Figure 1, our memory-friendly encoder consists of two stages—Coarse Extractor G and Memory Friendly Module F. Each stage stacks several convolution layers. Coarse Extractor G is a coarse network, which is a network with a simple structure and can restore a coarse HR image. Specifically, the module G takes the original LR image I^{LR} as input without interpolation, performs a series of convolutions and extracts coarse feature maps. The coarse feature maps are treated as high-dimensional feature vectors, leading overwhelming memory overheads. Therefore, the high-dimensional feature vector can be formulated by:

$$F_{HV} = G(I^{LR}) \tag{1}$$

where $G(\cdot)$ denotes the coarse extractor operation, which consists of a series of convolution layers, I^{LR} is the degraded low-resolution image and F_{HV} denotes the high-dimensional feature vectors extracted by G, serving as the inputs to the memory-friendly module.

As aforementioned, our EMTCM model takes the original LR image as the input through G with a sufficient number of convolution layers, thus becoming high dimensional vectors. The convolution operators result in a prohibitive memory of EMTCM. Therefore, we apply a memory-friendly module F that gradually extracts higher feature maps while decreasing the channel dimension. The low-dimensional feature vectors can be formulated by

$$F_{LV} = F(F_{HV}) = F(G(I^{LR})) \tag{2}$$

where F_{LV} is given by the function $F(\cdot)$ which takes F_{HV} in (1) as the input and gradually extracts higher information while decreasing the channel dimension.

As aforementioned, even though we impose an end-to-end architecture to infer the overall network, the LR image is extracted by high-dimensional feature vector, leading prohibitive memory. Therefore, we apply two coherent efforts and innovations to avoid the problem of notoriously prohibitive memory. On the one hand, a naive option directly takes raw pixel without any interpolation as input, but it still cannot sufficiently solve the problem of overwhelming memory and computation. On the other hand, we build a memory-friendly module to fix the weakness of the first effort. EMTCM with mutual collaboration between en-to-end network and memory-friendly can solve prohibitive memory and computation.

Recovery Decoder: as shown in Figure 1 right, we propose a recovery decoder, following the memory-recovery encoder, including mapping module M, recovery module R and a deconv layer De. In mapping module, we impose a sufficient number of CNN to capture more context information of LR in order to avoid the inherent weakness of local receptive field of CNN operation. However, that could cause the loss of feature resolution, fine details and gradient vanishing or gradient exploding. Another parallel way to effectively address the above issue is to apply a residual block with long skip connection. By doing so, we can capture more high-level information to guide the network to generate

super-resolver results. Meanwhile, we can relieve the inherent weakness of CNN. Here we define M function, as below:

$$F_{HL} = M(F_{LV}) = M(F(G(I^{LR}))) \tag{3}$$

where F_{HL} is given by the function $M(\cdot)$ that maps F_{LV} in (2) to F_{HL}. F_{HL} is the high-level information and the foundation of our multi-task co-training strategy.

As the following, we apply a recovery module R after the M module. In R module, we aim to increase the channel dimension of low-dimensional feature map vectors. Although, M module reduces the channel dimension of high-dimensional feature vectors for the sake of the computational efficiency, if we generate the HR image directly from low-dimensional feature vectors, the final performance quality will be poor. Therefore, we apply recovery module to boost performance further. It is defined as:

$$F_R = R(F_{HL}) = R(M(F(G(I^{LR})))) \tag{4}$$

where R denotes the recovery module function. F_R is obtained by function $R(\cdot)$ that recovers the resolution of the feature maps. F_R is significant for attaining final visual super resolution images.

The last module is Upsampling Operation. The operation is the learning based on upsampling Transposed Convolution Layer De, also known as deconvolution layer, which tries to perform a transformation opposite a normal convolution, i.e., to predict the possible input based on feature maps of the output size of convolutional layers. Specifically, it improves the image resolution by expanding the image by inserting zero values and performing convolution. Since the transposed convolution layer can enlarge the image size in an end-to-end manner while maintaining a connectivity pattern compatible with vanilla convolution, we impose deconv as learning-based upsampling method. Then the output is directly the reconstructed HR image. The deconvolution layer learns a set of upsampling kernels for the input feature maps. These kernels are diverse and meaningful. If we force these kernels to be identical, the parameters will be used inefficiently (equal to summing up the input feature maps as one). The final result is expressed as:

$$I^{SR} = De(F_R) = De(R(M(F(G(I^{LR}))))) \tag{5}$$

where De denotes the Transposed Convolution Operation and I^{SR} is the final output of the EMTCM model.

3.2. Multi-Task Co-Optimization Strategy

Image super-resolution tasks in the SISR domain benefit from the currently best collaboration between CNN-based and pixel wise loss. However, the collaboration seems like not the best partner in SISR. This is because there are some disadvantages to both of them. Meanwhile, in the span of just a couple of years, neural networks have been employed for virtually every computer vision and image processing task known to the research community. Much research has focused on the definition of new architectures that are better suited to a specific problem. A large effort was also made to understand the inner mechanisms of neural networks and what their intrinsic limitations are. However, loss function as an effective driver of network's learning has attracted little attention within the SISR research community: most CNN-based methods impose pixel wise L_2 loss. Note that L_2 loss and the Peak Signal-to-Noise Ratio, PSNR, do not correlate well with human's perception of image quality: L_2 is a single-task to just optimize PSNR. Here, we introduce a multi-task co-optimizing strategy to fix the aforementioned weakness. Interestingly, adding the multi-task co-optimizing strategy can improve performance. Therefore, that makes it a natural idea to incorporate multi-take co-optimizing strategy into EMTCM, which may help it capture more useful and meaningful information. Specifically, we construct a multi-task of super-resolution, instead of L_2 loss. That makes EMTCM focus on

inherent pattern of images, including local luminance, contrast, structure and fitting the ground-truth data distribution.

HVS task: Human visual system correlates well with inherent pattern of images, including local luminance, contrast and structure. Moreover, $SSIM$ loss can force the overall network to focus on inherent pattern of images and is proven effective in recovering high-field images. Hence we introduce the $SSIM$ loss to generate realistic images. We define the $SSIM$ loss as:

$$L^{HVS}(\theta_{EMTCM}, X) = \sum_{n=1}^{N} 1 - SSIM(X, \hat{X}) \tag{6}$$

where the network EMTCM is parameterized by θ_{EMTCM}, $SSIM(X, \hat{X})$ is a spatial similarity map between X and $\hat{X}$. X is ground-truth HR image and $\hat{X}$ is the last step to finally output I^{LR}.

Fitting ground-truth data distribution task: Another task co-optimizing is to fit ground-truth data distribution. In order to fix the weakness of $SSIM$ loss, we introduce Cross Entropy Loss as fine-tuning strategy to fit the distribution of HR. We define Cross Entropy Loss, as follows:

$$L^{FD}(\theta_{EMTCM}, X) = -\sum_{n} p_X(n) \log \hat{p}_X(n) \tag{7}$$

where p_X denotes the ground-truth probability distribution of image , $\hat{p}_X$ stands for the output probability distribution produced by EMTCM based on X. n stands for batch size of training data.

Multi-task co-optimizing strategy: the combination of two losses can achieve the goal of multi-task co-optimizing and improve performance further. Based on the two tasks and loss introduced above, we define the overall loss of our EMTCM super-resolution model as follows:

$$L^{EMTCM}(\theta_{EMTCM}, X) = L^{HVS}(\theta_{EMTCM}, X) + \beta L^{FD}(\theta_{EMTCM}, X) \tag{8}$$

where $L^{EMTCM}(\theta\_EMTCM, X)$ is the multi-task co-optimizing overall loss of EMTCM, β is a trade-off parameter that balances overall objective loss in order to fine-tune model. We set the $\beta = 0.0001$.

4. Results

4.1. Datasets

We conduct experiments on two widely used datasets: 91-images and General-100. Specifically, 91-image has 91 images and General-100 dataset contains 100 bmp-format images (with no compression). The size of the newly introduced 100 images ranges from 710 × 704 (large) to 131 × 112 (small). They are all of good quality with clear edges but fewer smooth regions (e.g., sky and ocean), thus they are very suitable for the training. However, as deep learning generally benefits from big data, 91-image and General-100 are not enough for training phase. In order to address data-hungry, we carry out data augmentation as in [39]. We augment data 19 times for both datasets from scaling and rotation two ways. (i) Scaling: each image is downscaled with the factor 0.6, 0,7, 0.8 and 0.9. (ii) Rotation: each image is rotated with the degree of 270, 90 and 180.

To prepare the training data, we first downsample the original training images by the desired scaling factors to form the LR images. Then we crop the LR training images into a set of f × f-pixel sub-images with a stride n. The corresponding HR sub-images (with size $((sf^2))$ are also cropped from the ground truth images. These LR/HR sub-image pairs are the primary training data. Test and validation datasets: we select Set5, Set14, BSD200 as test datasets. Another 20 images from the validation set of the BSD500 dataset are selected for validation.

Metrics: PSNR and SSIM are of the most popular metrics in super resolution. Therefore, we adopt them as performance measures and compute them on the Y channel of YCbCr.

4.2. Implement Details

Training Datasets Strategy: We adopt the 91-image dataset for training. In addition, we also explore a two-step training strategy. First, we train a network from scratch with the 91-image dataset. Then, when the training is saturated, we add the General-100 dataset for fine-tuning. With this strategy, the training converges much earlier than training with the two datasets from the beginning. When training with the 91-image dataset, the learning rate of the convolution layers is set to be 10^{-3} and that of the deconvolution layer is 10^{-4}. Then during fine-tuning, the learning rate of all layers is reduced by half. For initialization, the weights of the convolution filters are initialized with the method designed for PReLU. Meanwhile, we set $\beta = 0.0001$ for trade-off parameter to fine-tuning further. Our experiments are implemented on Pytorch 1.0.0 with NVIDIA TITAN RTX (24G).

Cross-Scale Training Strategy: Our EMTCM can be further trained in a cross-scale way with a cross-scale feature promotion promoting method. Specifically, we firstly have obtained a well-trained EMTCM under the upscaling factor 3, we then train the network for $\times 2$ on the basis of that for $\times 3$. To be specific, the parameters of G, F, M, R convolution filters in the well-trained EMTCM are shared to the four modules aforementioned EMTCM of $\times 2$. During training, we only fine-tune the deconvolution layer on the 91-image and General-100 datasets of $\times 2$. We conduct the same way for $\times 4$ experiment. By dong so, EMTCM can learn a better representation across different scales and receive the exchanged features from other scales by up/down-shared parameters, In such a design, the information transferred from the basic information is exchanged across each scale, which achieves a more powerful feature representation. This cross-scale training strategy further improves the performance of our approach.

Network Architecture: here we describe more details of our EMTCM networks. Given input LR images, LR features are extracted by G. In G module, the size and number of filter are 5×5 and 56. It denotes Conv(5,56,1). Then through F module, which consists of 12 convolution layers, it decreases the channel dimension of features. F denotes Conv(1,12,56).F is followed by M, which consists of nine residual blocks with long skip connection to capture more high-level information. M denotes 9Conv(1,12,12). After M module is R module, which consists of 56 convolution layers. R denotes Conv(1,56,12). Finally, SR images are recovered by deconvolutional layer with kernel size of 9. *De* denotes Deconv(9,1,56).

4.3. Results and Analysis

We compare the proposed EMTCM method with state-of-the-art SISR methods. Table 1 lists the quantitative evaluation results on 91 images using SSIM. Tables 2 and 3 show the PSNR and testing time in different scales, respectively. It can be observed that our EMTCM method achieves the best PSNR. Moreover, the speed of EMTCM is the fastest of all, satisfying the real time requirement. Table 4 shows the result of PSNR on datasets 91-image and General-100. Thus, EMTCM achieves comparable performance to SISR methods which perform PSNR-oriented tasks. This indicates that our EMTCM method is able to preserve pixel-wise accuracy while increasing the perceptual quality of super-resolved images. We also give a parameter comparison between EMTCM and state-of-the-art SR methods to demonstrate that our EMTCM is an effective memory-friendly network. Details are reported in Table 5, where we can see the high efficiency of parameter use in our EMTCM method compared with other models of SR.

We visualize some SR results of different methods [13,14,26,28,33,40–43] as shown in Figures 2–5. We see that EMTCM recovers correct details while other methods fail in giving pleasant results. This indicates that our method is able to produce more stable SR results than other methods. Note that our method has a significant advantage in handling large pose and rotation variations. The reason is that EMTCM is not only a PSNR-oriented task; it is also a HVS-oriented task. Hence, EMTCM can predict progressively more accurate feature maps to guide the reconstruction in each step. Therefore, our method performs better in preserving facial structures and generating better details even though images have large pose and rotation. Furthermore, EMTCM produces more realistic textures of images.

Therefore, the qualitative comparison with state-of-the-art face SR methods demonstrates the powerful generative ability of our methods.

4.4. Ablation Study

We further implement an ablation study to measure the effectiveness of the co-optimizing strategy. On the one hand, in order to validate the effectiveness of the HVS task, we remove the HVS loss. This model is called EMTCM-HVS, which is equivalent to a single task without $SSIM$ loss. On the other hand, we remove the fitting ground-truth data distribution to evaluate the effects of the proposed CE loss. This model is named EMTCM-FD. The PSNR on the 91-image dataset is presented in Table 6, where it is clear that when the EMTCM network loses a task, SR quality deteriorates since its ability to capture meaningful configuration weakens.

Table 1. The results of SSIM on three test datasets.

Test Dataset	Scale	KK	SRF	SRCNN	FSRCNN	Ours
Set5	2	0.9511	0.9556	0.9521	0.9558	0.9632
	3	0.9033	0.9098	0.9033	0.9140	0.9258
	4	0.8541	0.8600	0.8530	0.8657	0.8823
Set14	2	0.9026	0.9074	0.9039	0.9088	0.9152
	3	0.8132	0.8206	0.8145	0.8242	0.8420
	4	0.7419	0.7497	0.7413	0.7535	0.8065
BSD200	2	0.9000	0.9053	0.9287	0.9074	0.9201
	3	0.8016	0.8095	0.8038	0.8137	0.8374
	4	0.7282	0.7368	0.7291	0.7398	0.7596

Table 2. The results of PSNR (dB) on three test datasets. All models are trained on the 91-image dataset.

Test Dataset	Scale	SRF	SRCNN	SRCNN-EX	SCN	FSRCNN	Ours
Set5	2	36.84	36.33	36.67	36.67	36.94	37.15
	3	32.73	32.45	32.83	33.04	33.06	33.73
	4	30.35	30.15	30.45	30.82	30.55	30.85
Set14	2	32.46	32.15	32.35	32.48	32.54	32.33
	3	29.12	29.01	29.26	29.37	29.37	29.90
	4	27.14	27.21	27.44	27.62	27.50	27.67
BSD200	2	31.57	31.34	31.53	31.63	31.73	34.05
	3	28.40	28.27	28.47	28.54	28.55	29.54
	4	36.55	26.72	26.88	27.02	26.92	28.14

Table 3. The results of the testing time (in second) on three test datasets. All models are trained on the 91-image dataset.

Test Dataset	Scale	SRF	SRCNN	SRCNN-EX	SCN	FSRCNN	Ours
Set5	2	2.1	0.18	1.3	0.94	0.068	0.054
	3	1.7	0.18	1.3	1.8	0.027	0.023
	4	1.5	0.18	1.3	1.2	0.015	0.012
Set14	2	3.9	0.39	2.8	1.7	0.16	0.098
	3	2.5	0.39	2.8	3.6	0.061	0.056
	4	2.1	0.39	2.8	2.3	0.029	0.018
BSD200	2	3.1	0.23	1.7	1.1	0.098	0.088
	3	2.0	0.23	1.7	2.4	0.035	0.030
	4	1.7	0.23	1.7	1.4	0.019	0.016

Table 4. The results of PSNR on three datasets in comparison with state-of-the-art methods. EMTCM is trained on the 91-image and General-100 datasets.

Test Dataset	Scale	KK	A+	SRF	SRCNN	SCN	FSRCNN	Ours
Set5	2	36.20	36.55	36.89	36.43	36.93	36.94	37.13
	3	32.28	32.59	32.72	32.39	33.10	33.16	33.30
	4	30.03	30.28	30.35	30.09	30.86	30.71	30.88
Set14	2	32.11	32.28	32.52	32.18	32.56	32.63	33.08
	3	28.94	29.13	29.23	29.00	29.41	29.43	29.73
	4	27.14	27.32	27.41	27.20	27.64	27.59	27.73
BSD200	2	31.30	31.44	31.66	31.38	31.63	31.80	33.95
	3	28.19	28.36	28.45	28.28	28.54	28.60	29.39
	4	26.68	26.83	26.89	26.73	27.02	26.98	27.18

Table 5. Parameter comparison between our EMTCM and state-of-the-art methods to substantiate the efficacy of our method.

Model	Year	Parameters
SRCNN	2014	5.73 M
EDSR	2017	40.7 M
RCAN	2018	15.6 M
SAN	2019	15.7 M
IRN	2020	4.35 M
Ours	2021	1.89 M

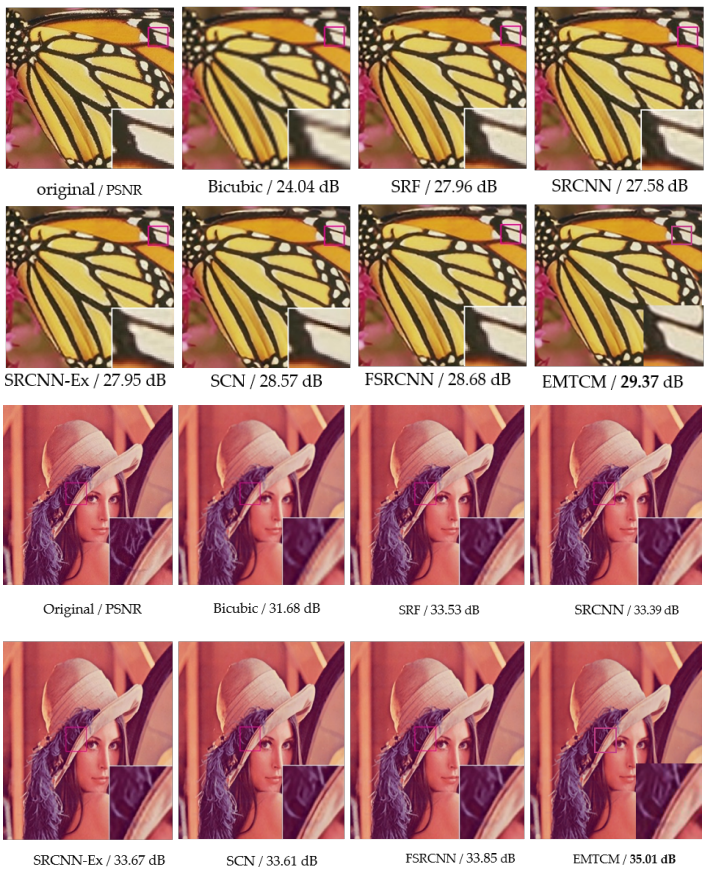

Figure 2. Qualitative and quantitative comparisons between our EMTCM and state-of-the-art SR models with the scale factor 3. Best viewed zoomed in.

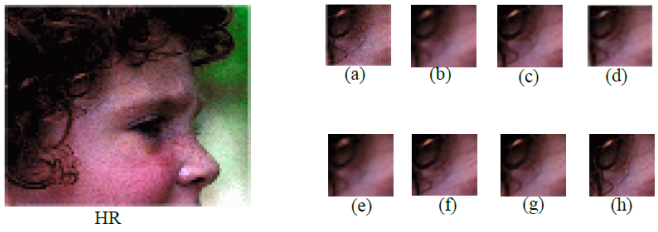

Figure 3. Qualitative comparison between our EMTCM and state-of-the-art SR models with the scale factor 3. (**a**) HR (**b**) Bicubic (**c**) FSRCNN (**d**) LapSRN (**e**) VDSR (**f**) RDN (**g**) SRFRN (**h**) Ours. Best viewed zoomed in.

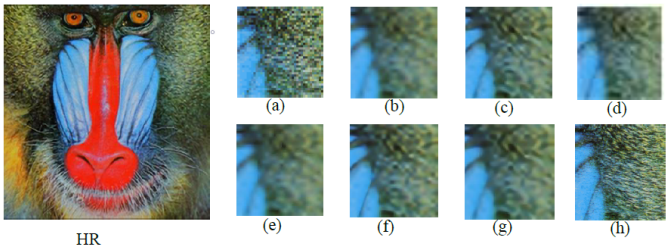

Figure 4. Qualitative comparison between our EMTCM and state-of-the-art SR models with the scale factor of 3. (**a**) HR (**b**) Bicubic (**c**) FSRCNN (**d**) LapSRN (**e**) VDSR (**f**) RDN (**g**) SRFRN (**h**) Ours. Best viewed zoomed in.

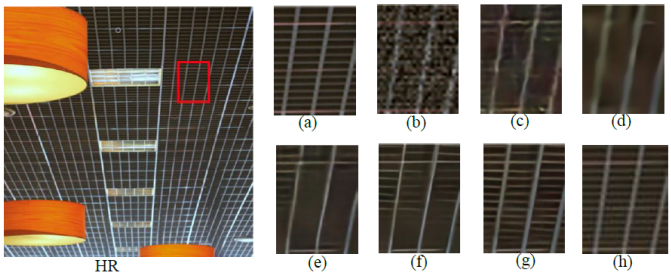

Figure 5. Qualitative comparison between our EMTCM and state-of-the-art SR models with the scale factor 3. (**a**) HR (**b**) Bicubic (**c**) SRCNN (**d**) IRCNN (**e**) SRMD (**f**) RDN (**g**) SRFRN (**h**) Ours. Best viewed zoomed in.

Table 6. PSNR under different settings of the loss function on the 91-image dataset with the scale factor 3.

Settings	Set5	et14	BSD200
EMTCM-FD	33.43	29.75	29.20
EMTCM-HVS	33.63	29.82	29.25
EMTCM	33.73	29.90	29.54

5. Conclusions

In this paper, we proposed an effective multi-task two-stage network for multi-scale image SR. The proposed EMTCM network exploits the the best collaboration between the CNN and multi-task co-optimization. The memory-friendly encoder in our model can avoid excessive memory overheads because it gradually extracts high-level features, which significantly decreases channel dimensions. Moreover, the proposed multi-task co-optimization method is applied to optimize the entire network. In addition, we utilize the residual blocks with the long skip connection to capture high-level information. Furthermore, with the cross-scale training strategy, our EMTCM network can further improve performance by learning different exchange features. Quantitative and qualitative experi-

ments on benchmark datasets demonstrate the competitive performance of the proposed EMTCM network, as compared with the state-of-the-art SISR methods.

Author Contributions: Conceptualization, J.Y. and F.W.; methodology, J.Y. and F.W.; software, Y.B.; validation, M.Z. and Y.C.; formal analysis, X.S. and Y.B.; investigation, F.W. and Y.B.; resources, J.Y. and Y.B.; data curation, X.S.; writing—original draft preparation, F.W.; writing—review and editing, J.Y. and Y.C.; visualization, X.S. and M.Z.; supervision, J.Y.; project administration, F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Tianjin Research Innovation Project for Postgraduate Students (2020YJSZXB11).

Data Availability Statement: All datasets at https://pan.baidu.com/s/1ZOqTF71GK-R3Lk9cGsydKA (access on 1 October 2021). Extraction code: sota.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, G.; Lv, J.; Tong, X.; Wang, C.; Yang, G. High-Resolution Pelvic MRI Reconstruction Using a Generative Adversarial Network with Attention and Cyclic Loss. *IEEE Access* **2021**, *9* , 105951–105964. [CrossRef]
2. Isaac, J.S.; Kulkarni, R. Super resolution techniques for medical image processing. In Proceedings of the 2015 International Conference on Technologies for Sustainable Development (ICTSD), Mumbai, India, 4–6 February 2015.
3. Huang, Y.; Shao, L.; Frangi, A.F. Simultaneous Super-Resolution and Cross-Modality Synthesis of 3D Medical Images using Weakly-Supervised Joint Convolutional Sparse Coding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
4. Farooq, M.A.; Khan, A.A.; Ahmad, A.; Raza, R.H. Effectiveness of State-of-the-Art Super Resolution Algorithms in Surveillance Environment. In Proceedings of the Conference on Multimedia, Interaction, Design and Innovation, online, 9–10 December 2020; pp. 79–88.
5. Zhang, L.; Zhang, H.; Shen, H.; Li, P. A super-resolution reconstruction algorithm for surveillance images. *Signal Process.* **2010**, *90*, 848–859. [CrossRef]
6. Rasti, P.; Uiboupin, T.; Escalera, S.; Anbarjafari, G. Convolutional neural network super resolution for face recognition in surveillance monitoring. In Proceedings of the International Conference on Articulated Motion and Deformable Objects, Palma de Mallorca, Spain, 13–15 July 2016; pp. 175–184.
7. Menon, S.; Damian, A.; Hu, S.; Ravi, N.; Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2437–2445.
8. Johnson, J.; Alahi, A.; Li, F. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
9. Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; Tan, M. Closed-loop matters: Dual regression networks for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5407–5416.
10. Cao, Q.; Lin, L.; Shi, Y.; Liang, X.; Li, G. Attention-aware face hallucination via deep reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 690–698.
11. Xu, X.; Sun, D.; Pan, J.; Zhang, Y.; Pfister, H.; Yang, M.H. Learning to super-resolve blurry face and text images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 251–260.
12. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
13. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
14. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
15. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep image prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9446–9454.
16. Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. *arXiv* **2018**, arXiv:1807.00734.
17. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
18. Kurach, K.; Lučić, M.; Zhai, X.; Michalski, M.; Gelly, S. A large-scale study on regularization and normalization in GANs. In Proceedings of the International Conference on Machine Learning—PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 3581–3590.

19. Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? In Proceedings of the International Conference on Machine Learning—PMLR, Stockholm Sweden, 10–15 July 2018; pp. 3481–3490.
20. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
21. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
22. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
23. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. A comprehensive evaluation of full reference image quality assessment algorithms. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012; pp. 1477–1480.
24. Haris, M.; Shakhnarovich, G.; Ukita, N. Task-driven super resolution: Object detection in low-resolution images. *arXiv* **2018**, arXiv:1803.1131.
25. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
26. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef] [PubMed]
27. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
28. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 391–407.
29. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
30. Ian, G.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
31. Bigdeli, S.A.; Jin, M.; Favaro, P.; Zwicker, M. Deep mean-shift priors for image restoration. *arXiv* **2017**, arXiv:1709.03749.
32. Meinhardt, T.; Moller, M.; Hazirbas, C.; Cremers, D. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1781–1790.
33. Zhang, K.; Zuo, W.; Gu, S.; Zhang, L. Learning deep CNN denoiser prior for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3929–3938.
34. Dong, W.; Zhang, L.; Shi, G.; Li, X. Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.* **2012**, *22*, 1620–1630. [CrossRef] [PubMed]
35. Jain, V.; Seung, S. Natural image denoising with convolutional networks. *Adv. Neural Inf. Process. Syst.* **2008**, *21*, 769–776.
36. Wang, Y.Q. A multilayer neural network for image demosaicking. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 1852–1856.
37. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
38. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
39. Wang, Z.; Liu, D.; Yang, J.; Han, W.; Huang, T. Deeply improved sparse coding for image super-resolution. *arXiv* **2015**, arXiv:1507.08905.
40. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
41. Ajith, M.; Kurup, A.R.; Martínez-Ramón, M. Time accelerated image super-resolution using shallow residual feature representative network. *arXiv* **2020**, arXiv:2004.04093.
42. Zhang, K.; Zuo, W.; Zhang, L. Learning a single convolutional super-resolution network for multiple degradations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3262–3271.
43. Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; Wu, W. Feedback network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3867–3876.