



Model Compression and Acceleration for Single Image Super-Resolution

Xuanzhen Li

School of Software, North University of China, Taiyuan, Shanxi, 030000, China
2313040348@st.nuc.edu.cn

Abstract. While deep learning-based methods for Single Image Super-Resolution (SISR) have consistently set new performance benchmarks, their substantial computational and memory footprints pose a significant barrier to deployment on resource-constrained devices. This challenge is particularly acute for SISR, a low-level vision task highly sensitive to the preservation of fine-grained texture details. To bridge this critical gap, two primary strategies are investigated: the design of lightweight network architectures and the application of model compression, focusing on network pruning and parameter quantization. This paper provides a critical investigation into why general-purpose compression algorithms, often developed for high-level vision tasks, frequently yield suboptimal results for SISR. This paper analyzes the inherent challenges, such as the structural constraints imposed by residual connections on pruning and the unique statistical distributions of feature maps that complicate quantization. Ultimately, the objective of this paper is to elucidate the intricate trade-offs among reconstruction fidelity, model size, and actual on-device latency, providing a principled foundation for designing genuinely efficient SISR models for real-world applications.

Keywords: Single Image Super-Resolution, Model Compression, Lightweight Network Architecture, Deep Learning.

1 Introduction

The demand for high-definition visual content is rapidly increasing, spurred by the adoption of 4K/8K displays, virtual reality (VR), and critical AI systems used in autonomous navigation and medical imaging. Single Image Super-Resolution (SISR) directly addresses this need. It is a computer vision task focused on the complex challenge of reconstructing a high-resolution (HR) image from a single low-resolution (LR) input.

Convolutional Neural Networks (CNNs) are now the standard approach in SISR, a trend started by the SRCNN model which proved the value of end-to-end training [1]. After SRCNN, the field split into two primary research paths. The first path focused on improving objective metrics like PSNR and SSIM, using deeper residual networks such as VDSR, EDSR, and RCAN to set new performance records [2] [3]. The second path used Generative Adversarial Networks (GANs) in models like SRGAN to create

images that looked more realistic and detailed. Despite their success, these leading state-of-the-art (SOTA) models all require significant computational power and memory [4]. This high resource cost makes them impractical for everyday consumer devices, highlighting a critical need for efficient, lightweight alternatives that maintain high quality.

Two main strategies have been developed to solve this efficiency problem. The first involves creating lightweight network architectures from scratch. This has produced innovative and compact models like FSRCNN and ESPCN (using post-upsampling), CARN (with cascaded blocks), LapSRN (progressive upsampling), and IMDN (information distillation), all designed to balance good performance with low computational cost. The second strategy uses model compression. Instead of designing new models, this approach takes large, existing models and applies techniques like network pruning and parameter quantization to make them smaller and faster without losing their original performance.

This paper provides a technical guide for creating efficient SISR models, aiming to encourage new work in the field. This paper surveys and analyzes the main strategies for model acceleration and compression. This paper first examines two core approaches: the design of lightweight architectures and the application of compression methods like network pruning and quantization, highlighting key models for each. This paper then compares these methods, focusing on the trade-offs between image quality, compression size, and actual inference speed. Finally, this paper summarizes the current state of the field, points out existing research gaps, and suggests potential avenues for future work.

2 Fundamentals of SISR and Model Compression

2.1 Single Image Super-Resolution (SISR)

Single Image Super-Resolution (SISR) addresses a classic and fundamentally ill-posed challenge in low-level computer vision. The difficulty arises because any single low-resolution (LR) input 'y' can correspond to a multitude of plausible high-resolution (HR) images, each differing in fine-grained details. This ambiguity necessitates the use of strong prior knowledge during the reconstruction phase to effectively narrow the range of potential solutions. The primary goal of SISR is therefore to infer the most faithful HR image 'x' from its LR counterpart.

The methodology for tackling SISR has evolved significantly over time. Early efforts were characterized by conventional techniques, including various inter-polation and reconstruction-based algorithms. A dramatic shift occurred with the introduction of SRCNN, which pioneered a new and now-dominant paradigm based on end-to-end learning with Convolutional Neural Networks (CNNs). This deep learning approach fundamentally re-conceptualizes the task, framing it as the process of training a network to learn a direct, non-linear transformation from the low-resolution space to the high-resolution space. Fig.1 illustrates a classic three-stage architecture that exemplifies this process.

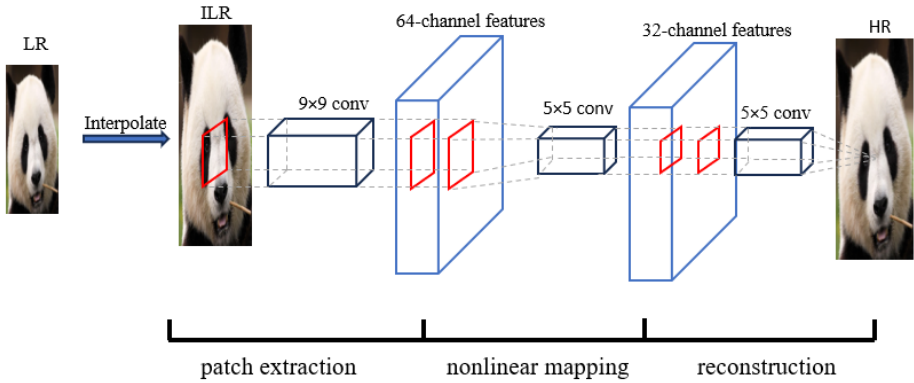


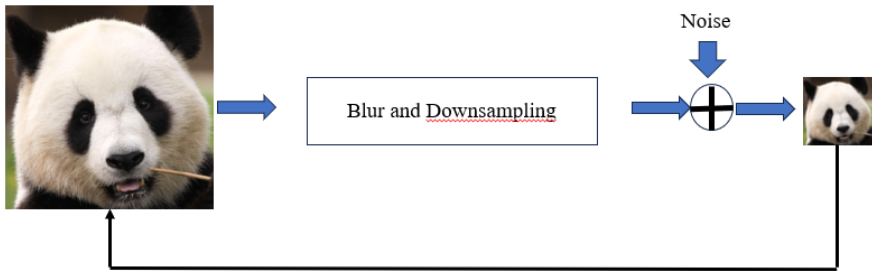
Fig. 1. The three-stage network architecture of SRCNN [1].

To generate the extensive datasets needed for robustly training deep models, a canonical image degradation model is frequently employed. The underlying principle of this model is that a given low-resolution (LR) image y is treated as the degraded output of a pristine high-resolution (HR) source image x . As illustrated in the pipeline shown in Fig. 2, this degradation process is mathematically formalized in Equation (1)

$$y = (x \otimes k) \downarrow_s + n \tag{1}$$

Where k denotes the blur kernel and n represents the additive noise, with \otimes indicating the convolution operation and \downarrow_s denoting downsampling with a scaling factor of s .

By applying this degradation model to datasets of high-quality HR images, vast quantities of paired (LR, HR) training samples can be generated.



SISR: Try to recover HR from its LR counterpart

Fig. 2. Process flow of the classic image degradation model [1].

The quality of SISR models is generally assessed using objective image fidelity metrics. Evaluation conventionally relies on metrics such as the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). The former is used to assess pixel-wise accuracy, a value derived from the mean squared error between the model's output and the original ground-truth image. SSIM, on the other hand, provides an evaluation that more closely mirrors human visual perception by assessing similarities in luminance, contrast, and structural information.

2.2 Model Compression

To enable the deployment of neural networks on resource-constrained hardware like mobile devices, a collection of techniques known as model compression is employed. The central objective of these methods is to reduce the storage foot-print and computational overhead of a network. This is ideally achieved with minimal impact on the model's predictive accuracy. Within the scope of this paper, the focus is placed on two such principal techniques: network pruning and parameter quantization.

Network pruning is based on the idea that deep neural networks are often inefficient, with many redundant or unnecessary parameters. By finding and removing these less important parameters, the technique results in smaller, faster models. Pruning methods fall into two main types: unstructured and structured. Unstructured pruning sets individual weights to zero, creating sparse weight tensors. Structured pruning, however, removes whole components like convolutional filters or channels, resulting in a smaller but still dense network. The structured approach is often a better fit for standard hardware accelerators, leading to greater speed improvements during inference [5].

Parameter quantization is another key compression method that lowers a model's numerical precision. This usually involves changing the standard 32-bit floating-point (FP32) weights to a smaller data type, such as 8-bit integers (INT8). This change offers two main advantages: it reduces the model's memory usage and bandwidth needs, and it can make inference much faster on hardware built for integer math. There are two common approaches. The first, Post-Training Quantization (PTQ), is a straightforward method where the conversion is done on a model that has already been trained. The second, Quantization-Aware Training (QAT), incorporates the effects of quantization into the training process itself, which usually results in better accuracy after compression. Fig. 3 shows a comparison of these two workflows [6].

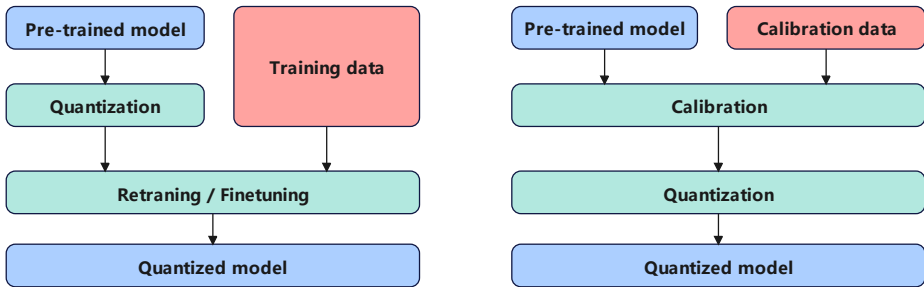


Fig. 3. A Comparison of PTQ and QAT Workflows [6].

3 Lightweight Architectures and Model Compression

Engineers and researchers have taken two main paths to make Single Image Super-Resolution (SISR) work on devices with limited power. The first path is to design lightweight network architectures, making efficiency a core part of the model. The second path is to take large, effective SISR models and shrink them using model compression. This chapter examines both methods. Based on recent studies, this paper

will analyze the unique features and limitations of each strategy in the context of SISR and conclude with potential areas for future research.

3.1 Lightweight Network Architecture Design

The design of lightweight network architectures integrates computational efficiency into a model's structure from the very beginning. The main objective is to achieve the best balance between performance and computational cost, working within a tight budget for parameters and FLOPs. This is accomplished by engineering novel, computationally efficient components, such as specialized modules or operators, to achieve maximum impact. Recent studies in this domain have highlighted two particularly promising avenues of investigation.

A significant focus of current research is on improving information flow within lightweight models by considering both dimensionality and scale. The Single Image Super-Resolution (SISR) task, which demands the reconstruction of fine texture details, necessitates the modeling of features across various scales. To address this, researchers have developed modules such as the Omni-Scale Aggregation Group (OSAG), which integrates operations from different receptive fields [7]. This approach tackles a key limitation of conventional self-attention in more compact networks, where the flow of information is often constrained by a narrow focus on either spatial or channel dimensions. To resolve this issue, the Omni Self-Attention (OSA) module was introduced, enabling the concurrent modeling of dependencies in both the spatial and channel domains within a single block.

Another line of research explores more effective feature spaces for attention mechanisms. The work on EMASRN, for instance, suggests that the target high-resolution (HR) space is the optimal domain for applying attention, as it offers better guidance for pixel reconstruction [8]. A significant challenge, however, is the high computational cost of calculating attention directly on HR feature maps. The High-Resolution Expectation-Maximization Attention Block (HREMAB) was developed to address this limitation. It constructs an efficient attention model by leveraging the iterative principles of the Expectation-Maximization (EM) algorithm. This methodology facilitates the direct capture of long-range spatial dependencies on the HR feature maps, bypassing the quadratic complexity inherent to such operations.

Despite its advantages, this design philosophy faces significant hurdles. The creation of innovative and efficient modules demands deep domain-specific knowledge, making the field less accessible to non-specialists. Moreover, the resulting architectures tend to be highly tailored to the SISR task, which restricts their transferability to other computer vision problems. Compounding these issues, the process of validating a new architecture is empirically driven and requires extensive experimentation, which in turn necessitates a substantial investment in research and development. The advantages and disadvantages of lightweight network architectures are summarized in Table 1.

Table 1. Pros and Cons of Lightweight Network Architecture Design.
(Data from: this study)

Advantages	Disadvantages & Limitations
Inherent High Performance	High Barrier to Entry
Streamlined Deployment	High Task Specificity
Promotes Field Innovation	High Development and Validation Costs

In response to the limitations discussed, this paper proposes three promising research directions for the design of lightweight architectures:

1) **Deeper Integration with Neural Architecture Search (NAS):** The application of NAS can automate the architectural design process. By setting up a multi-objective optimization that concurrently considers SISR quality metrics like PSNR and hardware performance indicators such as latency, NAS can autonomously identify superior network structures. This approach could significantly reduce the reliance on expert intuition and potentially uncover novel, highly efficient structural motifs.

2) **Decoupling Training and Inference Architectures via Structural Re-parameterization:** Future work should explore more architectures that exhibit a training-time and inference-time asymmetry, a concept popularized by methods like RepVGG. This involves utilizing complex, multi-branch structures during the training phase to improve representational capacity and stabilize optimization. For deployment, these complex structures can be equivalently fused into a simple, monolithic linear architecture (e.g., a single 3x3 convolution), thereby boosting the performance ceiling of lightweight SISR models without compromising their final inference speed.

3) **Synergistic Co-design with Compression Techniques:** The design of lightweight architectures and the application of model compression should not be viewed as mutually exclusive. A promising direction is the co-design of architectures that are inherently robust to compression. For instance, an architecture could be specifically designed to be pruning-friendly or quantization-friendly, minimizing the performance drop after compression is applied. This synergy could yield results where the combined effect is greater than the sum of the individual parts (a $1+1 > 2$ outcome).

3.2 Model Compression Techniques

A different strategy for creating efficient models involves compressing large, high-performance, pre-trained SISR networks. However, directly applying general-purpose compression algorithms to SISR often leads to less than ideal outcomes, as the distinct properties of the task must be carefully considered. The process typically involves using techniques like network pruning or parameter quantization to diminish a model's memory requirements and computational load, aiming for the smallest possible decline in performance.

1) **Network Pruning and its Challenges in SISR:** A primary obstacle to applying standard pruning methods to SISR networks arises from their heavy reliance on residual blocks. The summation operation within these blocks imposes a rigid structural

requirement: the channel dimensions of the input and output feature maps must be identical. This constraint, prevalent in deep SISR architectures, is incompatible with most conventional pruning algorithms. To address this issue, Structure-Regularized Pruning (SRP) was introduced [9]. SRP employs a progressively increasing L2 regularization penalty during the training phase, which guides the weights of targeted filters toward zero. This technique allows for effective pruning while adhering to the network's architectural demands and mitigating performance degradation.

2) **Parameter Quantization and its Challenges in SISR** : Similarly, applying conventional quantization methods directly to SISR models is often unsuccessful. The core challenges are threefold: the high sensitivity of the SISR task to pixel-level precision; the heterogeneous and wide-ranging distributions of intermediate feature maps; and the fact that many SISR models lack Batch Normalization (BN) layers, which are components that generally help to stabilize the quantization procedure. DAQ employs a dynamic, channel-wise quantization scheme that is sensitive to the data distribution [10]. This allows it to adapt effectively to the unique statistical properties of SISR feature maps, enabling high-quality image reconstruction even at ultra-low bit-widths, such as 2-bit.

The primary advantage of this approach is its broad applicability; it can be leveraged on virtually any pre-trained model. For large, over-parameterized networks, the potential for compression is well-defined and often substantial. Parameter quantization is particularly notable for being highly hardware-friendly, enabling significant inference acceleration on compatible hardware.

Nevertheless, this approach presents considerable challenges. Firstly, SISR is a task that is highly sensitive to numerical precision, creating a significant risk of performance degradation during the compression process. Secondly, the overall workflow—which involves pre-training, compression, and subsequent fine-tuning—is often complex and time-intensive. Lastly, designing effective compression algorithms for this domain requires deep expertise in both SISR and model compression, presenting a significant technical challenge for researchers. The strengths and weaknesses of model compression techniques are detailed in Table 2.

Table 2. Pros and Cons of Model Compression Techniques.
(Data from: this study)

Advantages	Disadvantages & Limitations
Architecture-Agnostic & Broadly Applicable	Risk of Significant Performance Degradation
Well-defined Optimization Goal	Complex and Time-Intensive Workflow
Amenable to Hardware Acceleration	Requires Interdisciplinary Expertise

To mitigate the issues of performance degradation and visual artifacts that arise when applying general-purpose compression algorithms to SISR, future research must move beyond one-size-fits-all approaches and embrace a paradigm of deep co-design between the task and the algorithm. Specifically, research efforts should be directed toward developing SISR-specific compression methods. This could include, for

example, creating pruning criteria that are aware of and can preserve critical texture details, or designing quantization-aware training schemes that are inherently robust against reconstruction artifacts. At the same time, future work should explore hybrid compression approaches that create synergy by combining methods like pruning, quantization, and knowledge distillation. Furthermore, establishing a standardized evaluation protocol focused on real-world deployment is essential. Such a framework must emphasize key performance indicators like true inference latency and energy usage, thereby reconciling the difference between theoretical computational estimates and actual on-device performance.

4 Conclusion

This paper provided a comprehensive examination of the two principal strategies for creating efficient Single Image Super-Resolution (SISR) models: the design of lightweight architectures and the application of model compression. Rather than merely cataloging existing techniques, this investigation has centered on an in-depth analysis of the specific challenges and inherent compromises that arise when applying general acceleration and compression methods to the unique demands of this low-level vision task.

Based on a detailed exposition and comparative analysis of these two technical avenues, this paper arrives at the following conclusions:

1) **Lightweight Network Architecture Design:** exemplified by recent works such as Omni-SR and EMASRN, shows immense potential for addressing the efficiency problem at a fundamental level. Its strengths lie in its native efficiency and higher performance ceiling. However, this approach is constrained by significant design complexity, a heavy reliance on expert intuition, and prohibitive development costs.

2) **Model Compression Techniques:** particularly network pruning and parameter quantization, offer a potent pathway for optimizing large-scale SISR models. Nevertheless, the direct application of generic compression algorithms developed for high-level vision tasks typically yields suboptimal results. The success of SISR-specific methods like SRP and DAQ underscores this point. Both were developed from a deep understanding of SISR's properties: SRP was designed to resolve the pruning index constraint inherent to residual architectures, while DAQ was engineered to overcome the quantization distortion challenge, which arises from the task's high demand for pixel-level precision and the unique statistical distributions of its feature maps.

In conclusion, regardless of the approach—whether designing efficient models from the ground up or compressing pre-existing ones—Task-Algorithm Co-design has emerged as the critical paradigm for driving future breakthroughs in this domain. Future research must therefore move beyond generic, "one-size-fits-all" solutions to focus on the intrinsic requirements of SISR. Future research should focus on systematic advancement in the following directions: leveraging Neural Architecture Search (NAS) to mitigate the reliance on manual design; decoupling training and inference architectures via techniques like structural re-parameterization; and exploring hybrid compression strategies. Crucially, a standardized evaluation framework oriented

toward real-world hardware deployment must be established to holistically assess the practical value of these different technologies.

References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) European Conference on Computer Vision (ECCV) 2014, LNCS, vol. 8695, pp. 184–199. Springer, Heidelberg (2014).
2. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 136–144. IEEE, Piscataway, USA (2017).
3. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) European Conference on Computer Vision (ECCV) 2018, LNCS, vol. 11207, pp. 286–301. Springer, Cham, USA (2018).
4. Ledig, C., Theis, L., Huszár, F., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690. IEEE, Piscataway, USA (2017).
5. Blalock, D., Gonzalez Ortiz, J.J., Frankle, J., Gutttag, J.: What is the state of neural network pruning?. In: Proceedings of Machine Learning and Systems, pp. 129–146. MLSys, Austin, USA (2020).
6. Nagel, M., Fournarakis, M., Amjad, R.A., Bondarenko, Y., van Baalen, M., Blankevoort, T.: A white paper on neural network quantization. arXiv preprint, arXiv:2106.08295 (2021).
7. Wang, H., Chen, X., Ni, B., Liu, Y., Liu, J.: Omni aggregation networks for lightweight image Super-Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22378–22387. IEEE, Piscataway, USA (2023).
8. Zhu, X., Guo, K., Ren, S., Hu, B., Hu, M., Fang, H.: Lightweight image Super-Resolution with Expectation-Maximization attention mechanism. IEEE Transactions on Circuits and Systems for Video Technology 32(3), 1273–1284 (2022).
9. Zhang, Y., Wang, H., Qin, C., Fu, Y.: Learning efficient image super-resolution networks via structure-regularized pruning. In: Proceedings of the International Conference on Learning Representations, ICLR (2022).
10. Hong, C., Kim, H., Baik, S., Oh, J., Lee, K.M.: Daq: Channel-wise distribution-aware quantization for deep image super-resolution networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2675–2684. IEEE, Piscataway, USA (2022).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

