JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 12, pp. 37-49, December 2025
https://doi.org/10.9708/jksci.2025.30.12.037

Minimization of Performance Degrading in Lightweight and Quantized Super-Resolution Models Through Feature-based Knowledge Distillation

Ho-min Jung*, Tae-Young Lee*, Byung-In Choi*

*Engineer, Software Team(Intelligent), Hanwha Systems. Co., Ltd., Seongnam, Korea

[Abstract]

This study proposes knowledge distillation (KD) method that minimizes the performance degradation caused by lightweighting and quantization in super-resolution (SR) tasks. The method has been redesigned to leverage simultaneously local and global feature information to maintain the detail restoration performance, and has been optimized the network into the edge device for validations. The spatial L1 loss function is used, in local level, to preserve the feature information such as boundaries, textures, and fine patterns. Meanwhile, in global level, the 2D FFT-based frequency transformation is employed to reflect the spatial characteristics and emphasize high-frequency components. This considerations of semantic context and spatial structure in images ensures to preserve fine details and structural consistency during the SR process. For verification, the network was optimized based on the performance comparison across different active functions for real-time operation on edge devices, and the local/global feature-based KD strategy was applied during initial training and quantization-aware training (QAT) to minimize performance loss. As results in optimized network, the inference speed has been improved by more than 7% on edge devices compared to the baseline. In our proposed method, it showed less performance degradation up to 0.12%, whereas the conventional QAT-based quantized models exhibited approximately 1.15% performance degradation in terms of PSNR. Thus, with our proposal, high-quality SR can be achieved even with lightweight models.

▸ **Key words**: Light-weight Super-Resolution, Knowledge Distillation, Quantization, Edge Device, NPU

• First Author: Ho-min Jung, Corresponding Author: Byung-In Choi
*Ho-min Jung (homin.jung@hanwha.com), Software Team(Intelligent), Hanwha Systems. Co., Ltd.
*Tae-Young Lee (ty.lee@hanwha.com), Software Team(Intelligent), Hanwha Systems. Co., Ltd.
*Byung-In Choi (byungin.choi@hanwha.com), Software Team(Intelligent), Hanwha Systems. Co., Ltd.
• Received: 2025. 11. 11, Revised: 2025. 11. 26, Accepted: 2025. 12. 10.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 연구는 초해상화 모델에서 경량화/양자화로 인한 성능 저하를 최소화하기 위해, 모델의 지역 및 전역 특징 정보를 함께 활용하는 지식 증류 방법을 제안하고 이를 검증하기 위한 엣지 디바이스에 최적화된 네트워크를 설계하였다. 제안하는 성능 저하 최소화 기법은 지역적 단위에서 공간 L1 손실 함수를 적용하여 경계, 질감, 세부 패턴 등 구조적 정보를 보존하며, 동시에 전역적 단위에서 2D FFT 기반 주파수 변환을 통해 공간 정보의 전역 특성을 반영하면서 고주파 성분을 강조하여 디테일 복원 성능을 보존하는 방식이다. 이미지의 의미적 맥락과 공간적 구조를 동시에 고려함으로써, 초해상화 과정에서 세부 정보의 보존과 구조적 일관성을 함께 만족할 수 있도록 설계했다. 이를 검증하기 위하여, 엣지 디바이스 환경에서 실시간 동작이 가능하도록 활성화 함수별 성능 비교 실험을 바탕으로 네트워크를 최적화하였고, 모델의 초기 학습과 양자화 인식 훈련(QAT) 과정에 지역/전역 특징 기반 지식 증류 방식을 적용하여 성능 저하를 최소화하였다. 실험 결과, 네트워크 최적화를 통해 엣지 디바이스에서 기존 대비 약 7% 이상 추론 속도를 개선하였고, 제안한 지식증류 기법의 경우, 기존 QAT 기반 양자화 모델은 원본 대비 PSNR 기준 약 1.15%의 성능 저하가 발생한 반면, 제안한 방법을 적용한 경우 성능 저하는 0.12% 수준에 그쳐, 경량화된 모델에서도 고품질 초해상화 복원이 가능함을 입증하였다.

▸ **주제어**: 경량 초해상화, 지식 증류, 양자화, 엣지 디바이스, NPU

I. Introduction

초해상화(Super-Resolution, SR) 기술은 저해상도 영상을 고해상도로 복원하는 기법으로, 의료 영상, 위성 영상, 감시·정찰 영상 등 다양한 분야에서 활용되고 있다. 최근에는 합성곱 신경망(Convolutional Neural Network, CNN) 기반 모델 [1],[2],[3]에서 높은 연산량이 요구되는 트랜스포머(Transformer) 모델 [4],[5],[6]을 활용한 연구로 확장됨에 따라 초해상화 복원 성능도 크게 향상되었다. 하지만 많은 연산량, 높은 메모리, 어텐션(Attention mechanism) 연산 미지원 등의 이유로 인해 엣지 디바이스나 NPU 환경에서 실시간 동작 및 모델 운용의 한계가 있다. 이러한 제약은 모바일 기기, 임베디드 IoT 제품 뿐 아니라 국방 감시 정찰 장비에 적용할 때 확연히 나타난다. 특히 국방 감시 정찰 장비는 단독 구동 형태로 운용되며, 저사양 감시 정찰 센서가 많이 사용됨에 따라 제한된 하드웨어 자원을 통해 실시간 영상 복원이 가능해야되기 때문에 경량화된 초해상화 모델이 필수적이다. 또한 초해상화 복원 결과는 객체 탐지 및 추적 등의 상황 인식 알고리즘의 사전 처리 단계로 활용되므로, 처리 속도와 성능 저하 최소화를 동시에 고려한 경량화 기법 설계가 중요하다.

기존 연구들은 네트워크 구조 단순화, 채널 수 축소, 양자화 등을 통해 연산량을 줄였으나, 모델 축소 및 양자화에 따른 질감 손실 등 이미지 복원에 대한 성능 저하가 발생 하였다. 특히 임베디드, NPU 등 엣지 디바이스에 적용하기 위한 양자화 과정은 부동 소수점 연산을 정수 연산으로 변환하는 과정에서 반올림 오차의 발생으로 발생하는 정보 소실만큼 복원 성능 저하를 야기한다. 이런 성능 저하는 픽셀 변화량이 큰 고주파 성분(경계, 모서리 등)에서 더욱 두드러지며, 결과적으로 영상의 정보 손실로 인한 품질 저하를 유발한다 [7],[8].

따라서 본 연구는 경량화 및 양자화 과정에서 발생하는 초해상화 성능 저하를 최소화하고자, 지역/전역 특징 기반 지식 증류(Knowledge Distillation, KD) 방법을 제안한다. 원본(Teacher)-모사(Student) 구조에서 지역 특징은 공간 L1 손실 함수를 적용하여 경계 및 세부 구조를 보존하고, 전역 특징은 2D FFT 기반의 주파수 변환을 통해 공간적 정보를 전역 특성을 반영하면서 고주파 성분을 강조하여 세부 복원에 대한 성능을 강화하였다. 전역 특징의 경우, 세부 복원에 기여도가 높은 고주파 성분을 중심으로 복원 손실 값을 계산하므로써 성능 저하를 완화하였다. 아울러, NPU 환경에서 실시간 동작이 가능하도록 활성화 함수 별 성능 비교 결과를 기반으로 네트워크를 효율적으로 최적화하고, 지역/전역 잔차 연산을 도입하여 성능과 연산 효율성을 동시에 확보하였다. 더불어, 양자화 인식 훈련(Quantization - Aware Training, QAT) 과정에서도 동

일한 지역/전역 특징 기반 지식 증류 전략을 적용하여, 양자화로 인한 성능 저하를 최소화하였다.

본 연구의 주요 기여점은 다음과 같다.

1. 엣지 디바이스 최적화 네트워크 설계: 활성화 함수별 연산 시간 분석과 지역/전역 잔차 연산을 통한 효율적 네트워크 설계를 통해 추론 속도를 최대화하였다.

2. 지역/전역 특징 분리 기반 지식 증류 기법: 지역 특징에는 공간 L1 손실 함수를, 전역 특징은 주파수 공간에서의 L1 손실 함수를 적용하여 지역/전역 특징이 분리되도록 증류 기법을 설계하였다. 모델의 복원 성능을 향상시키고, 동일한 전략을 양자화 인식 훈련에도 적용함으로써 모델 양자화로 인한 성능 저하를 억제하였다.

II. Preliminaries

2.1 Advancements in Super-Resolution Models and Lightweight Design Research

딥러닝 기반 초해상화 연구는 초기에는 주로 CNN을 중심으로 발전하였다. SRCNN[9], VDSR[10], EDSR[2] 등은 심층 CNN 구조를 통해 저해상도 영상의 세부 정보를 복원하였고, 이후 Residual Block, Dense Connection 등을 도입하여 복원 성능을 더욱 향상시켰다. 최근에는 트랜스포머 기반 초해상화 기법 연구가 활발해지면서, SwinIR[4], HAT[6] 등의 모델이 공개되었고, 이를 통해 전역적 구조나 패턴 등을 고려한 고품질 복원이 가능해졌다. 그러나 고성능 모델은 많은 연산량과 높은 메모리 사용이 요구되어, 임베디드, NPU 등 엣지 디바이스에서 실시간 동작이 어렵다. 이를 해결하기 위해 채널 수 축소, 경량 블록 설계 등 다양한 최적화 기법이 제안되었다.

2.2 Knowledge Distillation-Based Super-Resolution

지식 증류는 고성능 원본 모델이 학습한 지식을 상대적으로 파라미터 수가 적은 모사 모델로 전달하는 방법으로, 모델의 성능을 유지하면서도 추론 속도를 개선하는 경량화 방법 중 하나로 연구되어 왔다 [11],[12]. 최근 초해상화 영상 복원 분야에서도 실시간 추론이 요구도가 증가함에 따라, 지식 증류 기법이 적극적으로 활용되고 있다. 원본 모델과 모사 모델 간 L1 또는 L2 손실 함수를 사용해 복원 차이를 최소화하거나, 특징 정렬 방식을 주로 사용하였다 [13],[14]. 그러나 대부분의 모든 특징에 동일 손실 함수를 적용하므로 복원 과정에서 구조적 정보와 세부 질감처럼 서로 다른 표현 수준을 효과적으로 학습하기 어렵다. 이러한 한계를 극복하기 위해 본 연구에서는 지역/전역 특징을 분리하고 각각의 특징에 서로 다른 손실 함수를 적용하여 이미지의 의미적 맥락과 공간적 구조를 동시에 반영하는 지식 증류 기법을 제안한다.

2.3 Utilization of Frequency Domain

주파수 영역 분석은 영상의 저주파 성분(구조)과 고주파 성분(세부 질감)을 분리하여 처리할 수 있는 장점이 있다. 초해상화 연구에서는 푸리에 변환(Fourier Transform, FT), 이산 코사인 변환(Discrete Cosine Transform, DCT) 등을 활용하여 고주파 성분을 강화하거나, 주파수 대역별 손실 함수를 설계하는 방법이 제안되었다 [15]. 손실 함수를 통해 주파수 정보를 활용하므로 세부 정보 복원력은 향상하였으나, 출력 이미지 수준에서만 적용되거나 단일 손실 함수에 통합되는 경우가 많다.

2.4 Techniques for Post-Training Quantization and Quantization-Aware Training

양자화는 인공지능 모델을 NPU, 임베디드 등 엣지 디바이스에 운용 하기 위한 필수 과정으로, 연산 효율을 높이고 메모리 사용량을 줄이는 과정이다 [16]. 그러나 실수값을 정수 값으로 근사하는 과정에서 필연적으로 정보 손실이 발생하며, 이는 모델의 성능 저하로 이어진다. 특히 초해상화 복원에서는 고주파 성분이 경계, 질감, 세부 패턴 등 디테일을 구성하는 핵심 요소로 작용하는데, 이들 성분은 상대적으로 진폭이 작고 변화가 급격하여 양자화 오차에 더 민감하게 반응한다. 저주파 성분은 상대적으로 진폭이 크고 변화가 완만하여 양자화 후에도 안정적으로 유지되지만, 고주파 성분은 작은 오차에도 구조적 왜곡이나 세부 정보 손실로 이어질 수 있다. 따라서 초해상화 등과 같은 고주파 성분 활용이 중요한 분야에서는 양자화로 인한 성능 저하가 더욱 두드러지게 나타난다.

사후 양자화(Post Training Quantization, PTQ) 기법은 학습이 완료된 모델에 양자화를 적용하는 방식으로, 구현이 간단하고 추가 학습이 필요 없다는 장점이 있다. 그러나 학습 과정과 무관하게 사후적으로 양자화를 적용하기 때문에 비교적 성능 손실이 크게 발생한다. 특히 초해상화 분야에서는 양자화 적용 간 고주파 손실로 인한 경계와 세부 구조가 손상되어 전반적인 복원 성능 저하가 더욱 두드러지게 나타난다.

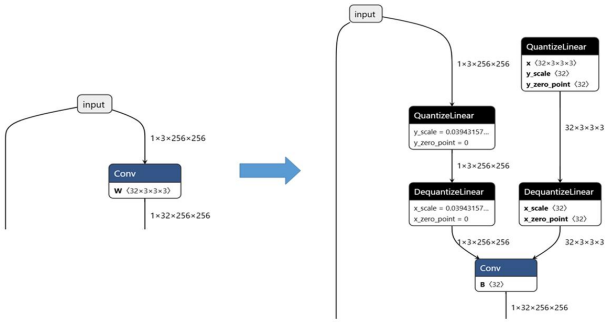

Fig. 1. Example of Applying Quantization Operations for QAT

QAT 기법의 경우, Fig. 1.과 같이 학습 과정에서 네트워크 연산 그래프에 QDQ (Quantize-DeQuantize) 연산을 추가하여 실제 양자화 과정을 모사 한다. 이때 추가 학습 간 발생하는 근사 오차가 손실 함수에 직접 반영되어, 역전파를 통해 파라미터가 양자화 환경에 적응하도록 학습된다. 결과적으로, QAT 기법은 모델 배포 시 실제 양자화된 연산에서도 정밀도 손실을 최소화할 수 있다. 해당 기법은 PTQ 기법 대비 성능 보존 측면에서 유리하지만, 추가 학습 비용과 구현 복잡성으로 인해 실제 적용은 비교적 제한적이다. 이에 따라 구현이 간단한 PTQ가 상대적으로 널리 활용되고 있으며, 양자화로 인한 성능 저하를 보완하기 위해 지식 증류를 결합한 연구는 아직 드문 편이다.

III. The Proposed Scheme

3.1 Network Structural Design: Limitations and Alternatives of Nonlinear Activation Functions

최근 초해상화 모델은 세밀한 질감 복원과 전역 문맥 반영을 위해 Pixel-wise attention 연산[17]과 더불어 SiLU[18] 등과 같은 비선형 활성화 함수가 널리 활용되고 있다. ReLU 계열의 활성화 함수 대비 더 부드러운 비선형성을 제공해 음수 입력에서의 정보 손실을 줄이고, 표현력을 향상 시켜 모델 학습 시 더 안정적인 수렴을 유도하여 복원 성능을 개선하고 있다.

그러나 Pixel-wise Attention 연산은 성능적 이점에도 불구하고, 모든 픽셀에 대해 비선형 연산을 수행하여 Attention map을 계산해야 하므로 연산량과 메모리 접근량의 증가로 인해 NPU에서의 실시간 추론이 어렵다. 또한 Sigmoid 기반 연산은 엣지 디바이스에서의 연산 지원이 제한적이어서, 부동소수점 변환을 거치는 Quantization - FP32 변환 - DeQuantization 과정에서 연산 병목 현상 및 정확도 손실 야기한다.

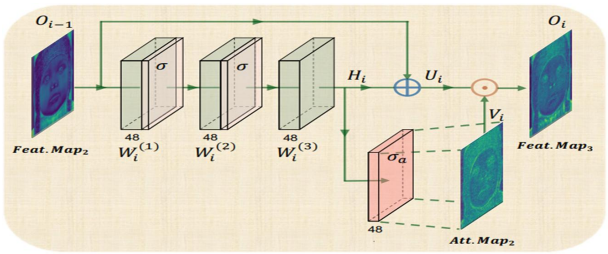

Fig. 2. Swift Parameter-free Attention Block (SPAB) in SPAN, with computations within the attention map contributing to faster processing

특히 SiLU 연산은 지수(exponential) 연산을 포함하기 때문에 출력 포화와 기울기 소실 문제가 뒤따르며, 무엇보다도 ReLU 계열에 비해 연산 복잡도가 높아 하드웨어 가속 효율이 낮다. 이러한 문제를 정량적으로 확인하기 위해 양자화 후 초해상화 분야에서 가장많이 사용되는 활성화 함수에 따른 성능 비교 결과는 아래 Table 1.과 같다.

Table 1. Performance Comparison Before and After Quantization by Activation Function under the Same Model

-		Leaky ReLU	PReLU	SiLU
Inference Speed	Base line	20.04ms	20.05ms	20.17ms
	quantized	6.95ms (65.33% ↑)	6.98ms (65.18% ↑)	7.48ms (62.91% ↑)
PSNR Drop (Urban 100)	Base line	28.21dB	28.33dB	28.72dB
	quantized	28.14dB (0.07 ↓)	28.09dB (0.14 ↓)	28.21dB (0.51 ↓)

Table 1.에서 확인할 수 있듯이, Parametric ReLU (PReLU)와 SiLU는 Leaky ReLU에 비해 양자화 후 추론 속도가 느린 경향을 보인다. PReLU[19]는 활성화 함수에 학습 가능한 파라미터가 포함되어 있어 복원 성능 향상에는 유리하지만, 해당 매개변수에 접근하는 과정에서 Leaky ReLU 대비 소폭의 지연이 발생할 수 있다. 또한, 양자화 과정에서 값이 근사화되면서 특히 작은 소수 값은 정수 표현에서 0으로 소실될 위험이 존재하며, 이로 인해 양자화 이후에도 Leaky ReLU에 비해 성능 소실이 더 크게 발생하는 것으로 나타났다.

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ ax, & \text{if } x \le 0 \end{cases} \quad (1)$$

또한 SiLU 함수는 복원 성능 측면에서는 음수 입력을 완전히 배제하지 않고, 연속적으로 처리한다는 점에서 복

원 성능이 가장 좋았다. 하지만 지수 및 분수 연산을 포함하기 때문에 엣지 디바이스 환경에서 동일 모델에서의 Leaky ReLU 대비 약 7% 이상의 속도 저하가 발생 하였다. 반면, Leaky ReLU는 단순한 연산 구조로 가장 빠른 처리 속도를 보였다. 특히 음수 기울기가 고정된 상수이므로 양자화 과정에서도 값이 변하지 않아, 실시간 처리나 저전력 임베디드 환경에서 가장 적합한 것으로 확인되었다. 특히 SiLU 대비 7% 이상 속도 개선을 보였고 양자화 적용간 성능 소실 역시 가장 적은 결과를 보여주었다.

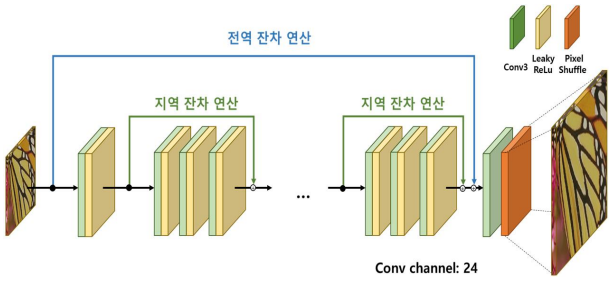

Fig. 3. Architecture of the Proposed Model

이러한 비교 결과를 바탕으로, 본 연구는 속도와 복원 성능 두가지 측면에서 엣지 디바이스 환경의 동작을 고려하여 PReLU, SiLU 대신 Leaky ReLU 활성화 함수를 적용하여 네트워크를 구성하였다. 더 나아가, 네트워크 최적화 및 양자화 과정에서 레이어 융합을 함께 고려함으로써 Leaky ReLU 활성화 함수 채택이 해당 네트워크의 연산 효율성을 한층 더 극대화하는 데 기여하였다.

또한, 본 연구에서는 네트워크의 구조적 관점에서 블록 단위의 세밀한 정보 전달을 담당하는 지역 잔차 연산과 전체적인 구조 및 고주파 성분 복원을 담당하는 전역 잔차 연산을 동시에 수행할 수 있는 구조를 설계하였다. 이러한 구조는 지역적 세부 정보와 전역적 맥락을 함께 반영함으로써 보다 정밀한 복원이 가능하다는 장점을 지닌다. 구체적으로, 지역 잔차 연산은 인접 영역의 질감과 경계 정보를 보존하는 데 효과적이며, 전역 잔차 연산은 전역 문맥과 구조적 일관성을 유지하는 데 기여한다. 이를 종합하여 Fig. 3.과 같이 Leaky ReLU 기반의 지역/전역 잔차 연산 기반의 경량 초해상화 모델인 lightweight Local-Global Residual super resolution Network (LGR-Net) 을 제안한다.

Table 2. Comparison of Performance Across Model Architectures (PSNR and Time), evaluated on NVIDIA Orin AGX 32GB under a 15W power setting.

Dataset	w/o residual	w/ local resual	w/ global residual	w/ local & global
Set5	33.79dB	33.92dB	34.11dB	**34.33dB** (0.54↑)
Set14	29.99dB	30.13dB	30.22dB	**30.52dB** (0.53↑)
BSD100	29.46dB	29.75dB	29.78dB	**29.91dB** (0.45↑)
Urban100	26.93dB	27.06dB	27.12dB	**27.31dB** (0.38↑)
Manga109	31.55dB	31.66dB	31.83dB	**31.98dB** (0.43↑)
Processing time	5.68ms	5.91ms	6.54ms	**6.95ms** (1.27↑)

제안된 모델은 경량 초해상화 모델로서 엣지 디바이스 환경에서 실시간 동작이 가능하도록 구조적 경량화를 수행하고, 양자화 과정에서의 레이어 융합(Conv + Leaky ReLU)을 고려하여 설계되었다. 또한 메모리 대역폭 병목 현상을 최소화하기 위해 구조를 최적화하였다. 특히 채널 수 변화에 따른 추가 커널 호출로 증가되는 연산량을 줄이기 위해 채널 수를 유지하는 방향으로네트워크를 구성하였다. 이를 통해 레이어 간 채널 변환으로 인한 불필요한 연산과 메모리 접근을 감소시키고, 하드웨어의 병렬 처리 효율을 극대화함으로써 메모리 효율 개선을 통하여 모델 추론 속도를 개선하였다.

Table 2.에서 확인할 수 있듯이, 지역/전역 잔차 연산을 모두 채택 할 경우 추론 시간이 소폭 증가하였으나, 동일 디바이스 내 동일 전력(15W) 조건에서 PSNR이 유의미하게 향상되어 결과적으로 동일 전력 소모 대비 복원 품질 측면에서 효율적인 구조임을 증명하였다.

다만, 해당 구조는 NPU 구동을 위한 경량화를 지향하는 설계 특성상 파라미터 수가 제한적이기 때문에, 고품질 복원이 요구되는 상황에서는 성능 저하가 발생할 수 있는 잠재적 한계를 내포하고 있다. 이를 보완하기 위해, 본 연구에서는 지역/전역 특징 기반 지식 증류 기법을 도입하여 모델 경량화 과정에서 손실된 세부 정보와 구조적 일관성을 효과적으로 보존하고 복원하였다. 따라서 제안된 네트워크는 단순한 경량성과 속도 향상뿐 아니라, 지역/전역 잔차 연산 내 지식 전이를 고려하여 설계됨으로써 구조적 일관성과 디테일 복원을 동시에 충족하도록 하였다.

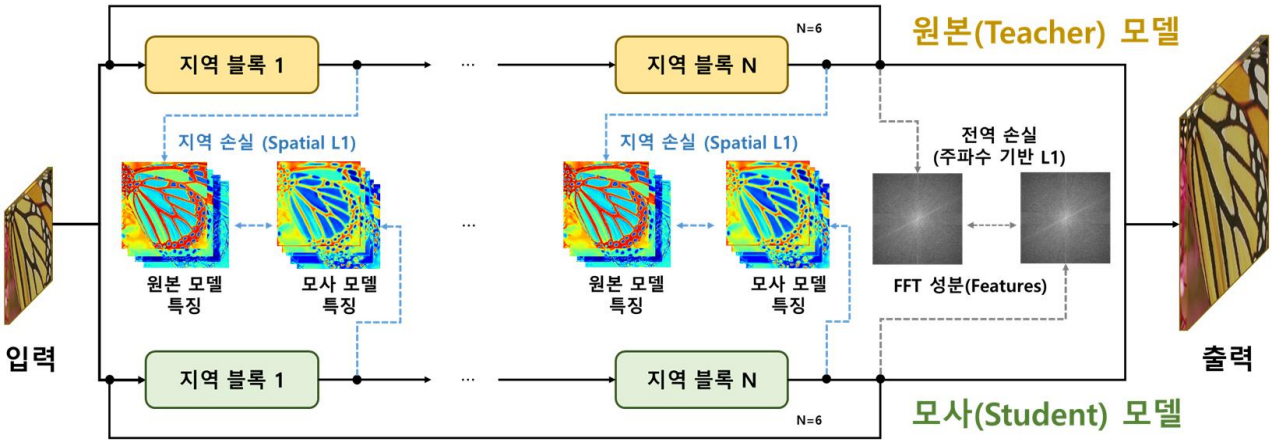

Fig. 4. Knowledge Distillation Flow between Local and Global Features of the Teacher Model and the Student Model

3.2 Knowledge Distillation Strategies for Compensating Performance Degradation

본 장에서는 초해상화 네트워크의 구조적 경량화로 인한 성능 저하를 보완하기 위해, 지역 및 전역 특징을 분리하여 손실 함수를 계산하는 새로운 지식 증류 기법을 제안한다. 제안 방법은 사전 학습된 고성능 초해상화 모델 SPAN[15]을 원본 모델로, 본 논문에서 설계한 LGR-Net을 모사 모델로 설정하였다. LGR-Net은 SPAN의 핵심 연산인 Pixel-Wise Attention을 지역 블록에 반영하여, 모사 네트워크가 원본 모델의 특징을 효과적으로 학습할 수 있도록 설계되었다. 본 연구에서 제안하는 지식 증류 전략은 기존의 단일 특징 기반 접근과 달리, 지역적 특징들과 전역적 특징을 분리하여 각각 최적화한다. 지역 경로에서는 경계, 질감, 세부 패턴 등 고주파 디테일 복원을 위해 공간 영역 L1 손실을 적용하였다. 전역 경로에서는 2D FFT 기반 주파수 변환을 통해 공간적 위치에 무관한 전역 분포를 반영하고, 고주파 성분을 강조하여 세부 질감과 경계 복원 능력을 강화하였다. 이를 통해 모사 모델은 전역적 표현 안정성과 세부 디테일 복원력을 동시에 확보할 수 있다. 최종적으로 지역 손실 함수와 전역 손실 함수를 결합하여 경량 모델이 구조적 단순화에도 불구하고 고품질 복원을 달성하도록 하였다. 제안하는 지식 증류 기법은 단순한 성능 향상을 넘어, 양자화 및 엣지 디바이스 환경에서도 고주파 손실을 최소화하며 실시간 동작이 가능한 경량 초해상화 모델 구현에 핵심적인 역할을 한다.

먼저 지역 블록에서는, 지역적 세부 정보 복원 능력을 올리기 위하여 Fig. 4.와 같이 원본 모델과 모사 모델의 각 지역 블록에서 특징들을 추출하고, 공간 영역에서의 L1 손실 함수를 기반으로 블록 간의 차이를 계산한다. 그러나 원본 모델은 구조적으로 더 크기 때문에 추출된 특징의 크기가 모사 모델과 달라 직접적인 손실 계산에 적합하지 않다. 이에 따라 아래 식 2와 같이, 모사 모델의 작은 특징 크기에 맞추기 위해 원본 모델의 특징에 적응형 평균 풀링(Adaptive Average Pooling) 연산을 적용하고, 채널 수를 축소를 통해 서로 다른 특징의 크기를 통일하여 손실을 계산하였다.

$$L_{Local} = \frac{1}{N-1}\sum_{i=1}^{N-1} \alpha \cdot \| (f_s^{(i)}) - R(Pool(f_t^{(i)})) \|_1 \quad (2)$$

여기서 N은 전체 특징의 개수이고, 해당 식에서는 지역 특징에 한해서 $N-1$까지 계산하며, $f_s^{(i)}, f_t^{(i)}$는 모사, 원본 모델의 i번째 지역 특징 맵이며, R은 채널 축소 연산자를 의미하며, $Pool(\cdot)$은 적응형 평균 풀링을 의미하고, α는 지역 손실의 가중치를 의미한다.

전역 잔차 연산에서는 2D FFT 연산을 적용하여 입력 이미지를 공간 영역에서 주파수 영역으로 변환한 뒤, 강도 스펙트럼을 추출한다. 이후 고주파 마스크를 적용해 세부 질감과 경계 정보를 강조한 주파수 성분을 분리하고, 원본 모델와 모사 모델 간의 스펙트럼 차이에 대해 L1 손실 함수로 전역 손실을 계산한다. 이 전역 손실은 공간적 위치에 의존하지 않고 영상 전체의 주파수 분포를 기반으로 계산되므로 전역적 특성을 반영하며, 동시에 고주파 성분을 통해 영상 복원 성능을 강화한다.

$$L_{Global} = \| H(|F(f_s^{(N)})|) - H(|F(R(Pool(f_t^{(N)})))|) \|_1 \quad (3)$$

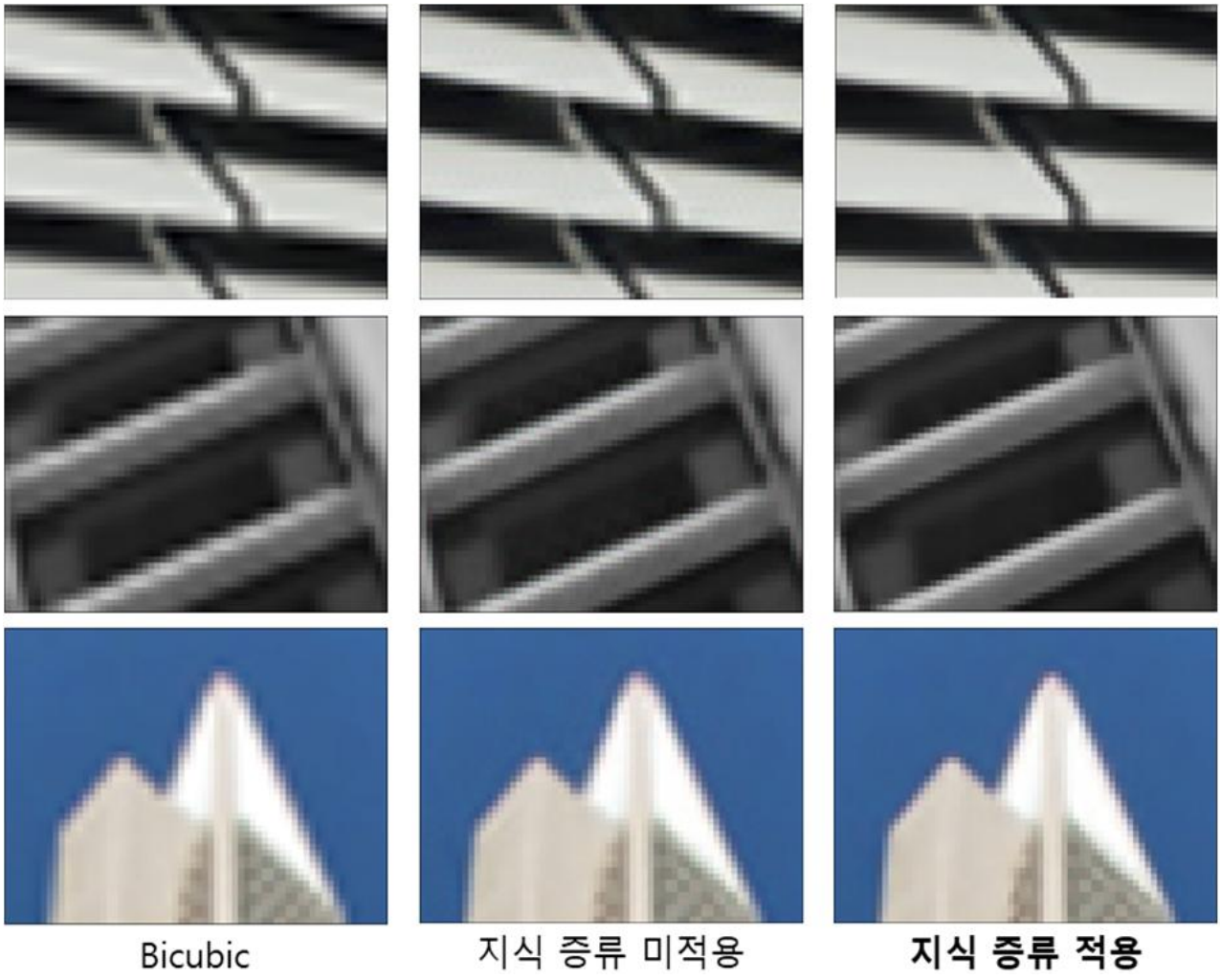

Fig. 5. Qualitative Comparison of the Model with and without Knowledge Distillation (at 2× scale-up). After applying the technique, unnecessary artifacts in edge regions are reduced.

여기서 $F(\cdot)$는 2D FFT 연산을 의미하며, $|\cdot|$는 복소수 magnitude이며, $H(\cdot)$는 고주파 마스크 적용 함수이다.

지역 잔차 연산에서의 지식 증류는 경계, 질감과 같은 지역적 세부정보에 특화된 손실 함수를 적용하여 선명도와 구조적 표현력을 향상시켰고, 전역 잔차 연산 모듈에서는 2D FFT 연산을 통해 주파수 영역에서의 전역적 특징 및 안정성을 보완하며, 동시에 고주파 성분을 강조함으로써 세부 질감과 경계 복원 능력을 강화하였다. 두 가지 손실 값을 결합함으로써 고품질 복원이 가능하도록 설계하였다.

다음으로는, 지역/전역 특징 분리와 손실 함수를 통해 지역적 세부 정보와 전역적 안정성을 학습하였다면, 이어서 부드러운 정답(Soft Target) 기반 증류를 적용하여 단순히 특징 수준의 모사에 그치지 않고 원본 모델이 생성한 최종 복원 이미지를 참조하도록 하였다. 원본 모델의 부드러운 정답 값은 정답 값(Ground Truth)와 달리 질감 조정, 색감 보정, 세부 패턴 반영 등 복원 과정에서의 전략이 내재된 부드러운(Soft) 분포를 가지며, 이를 활용함으로써 학생 모델은 완성된 이미지 수준에서 원본 모델의 복원 전략을 효과적으로 전이받을 수 있다. 결과적으로, 특징 모사와 이미지 모사가 상호보완적으로 작용하여 경량 모델임에도 고품질 복원을 달성할 수 있도록 한다.

$$L_{Teacher} = \| Y^S - Y^T \|_1 \quad (4)$$

지식 증류의 마지막 단계에서는 정답 라벨 기반 지도 학습 (Label Supervision)을 결합하여 정답 값과의 직접적인 차이를 최소화함으로써, 절대적 기준인 정답 값에 기반한 학습을 보장하였다.

$$L_{Label} = \| Y^S - Y^{GT} \|_1 \quad (5)$$

Table 3. Quantitative Results by Dataset with and without the Proposed Knowledge Distillation Method (at 2× Scale-up)

Proposed Method	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
O	**34.94**	**0.9462**	**30.94**	**0.8917**	**30.26**	**0.8665**	**28.21**	**0.8832**	**32.95**	**0.9536**
X	34.33	0.9407	30.52	0.8854	29.91	0.8512	27.31	0.8691	31.98	0.9454

이는 부드러운 정답이 제공하는 상대적 분포 정보나 블록 단위 증류가 전이하는 부분적 특징 모사만으로는 확보하기 어려운 객관적이고 확정적인 지도 신호를 제공한다는 점에서 의미가 있다. 즉, 부드러운 정답은 원본 모델의 복원 전략을 반영하고, 블록 단위 증류는 지역·전역 수준의 표현 학습을 가능하게 하지만, 두 방법 모두 궁극적으로는 원본 모델에 종속된 상대적 지식 전이에 해당한다. 반면 라벨 기반 지도 학습은 정답 값이라는 절대적 참조를 통해 학습 과정에서 발생할 수 있는 누적 오차를 교정하고, 최종 출력이 실제 고해상도 목표에 도달할 수 있도록 유도하는 과정이다. 결과적으로, 제안한 세 가지 지식 증류 과정은 블록 단위 특징 모사와 이미지 수준 모사에 기반한 상대적 지식 전이와, 정답 기반 학습을 통한 절대적 지도가 상호보완적으로 작용하는 단계적 학습 과정을 형성한다. 이를 통해 경량 모델임에도 세부 디테일 복원력과 전역적 표현 안정성을 동시에 확보할 수 있다. 총 세 단계를 통해 계산된 $L_{Loc}, L_{Glo}, L_{Teacher}, L_{Label}$을 바탕으로 최종 L_{total} 은 각 L의 가중 합으로 설계하였다.

$$L_{Total} = \alpha(L_{Local} + L_{Global}) + \beta L_{Teacher} + \gamma L_{Label} \quad (6)$$

단, $\alpha + \beta + \gamma = 1$이며, 본 연구에서는 손실 함수 간의 균형과 학습 목표의 우선순위를 고려하여 최종적으로 $\alpha = 0.25, \beta = 0.25, \gamma = 0.5$로 설정하였다. 모델의 기본 성능 확보를 위해서 L_{Label}에 가장 높은 가중치를 할당하였고, α, β의 경우 L_{Label}의 절반 수준인 0.25로 설정하여 label loss 대비 보조적인 역할을 수행하도록 유도하였다.

본 논문에서 제안한 지식 증류 기법이 모델의 복원 성능에 미치는 영향성을 검증하기 위해 비교 실험을 수행하였다. 검증 모델로는 본 논문에서 제안한 LGR-Net을 사용하였다. 학습 간에 사용한 데이터의 경우 DIV2K [23]와 LSDIR [24] 데이터셋을 패치 단위(256×256)로 정제하여 활용하였다. 학습 관련 하이퍼파라미터 설정은 배치 크기 64, 총 500 Epoch, 초기 학습률 1×10−4에서 시작하여 Cosine Annealing 스케줄러를 적용하여 학습을 진행하였다.

평가의 경우 Set5 [25], Set14 [26], BSD100 [27], Urban100 [28], Manga109 [29] 데이터셋을 대상으로 최대 신호 대 잡음비 (Peak Signal-to-Noise Ratio, PSNR)과 구조적 유사도 지수 (Structural Similarity Index Measure, SSIM) 지표를 측정하였다. 그 결과 제안한 지역/전역 특징 기반 지식 증류를 적용하였을 때 Table 3. 과 같이, 정량적 평가에서 PSNR에서 평균 2.1%, SSIM에서 평균 1.1% 성능 향상을 보였다. 또한, 본 논문에서 제안한 지식 증류 기법을 적용 후 정성적 평가에서도 Fig. 5. 와 같이 경계 영역에서 잡음 정보가 줄어들고 경계 성분이 선명하게 복원되는 것을 확인하였다.

3.3 Application of Knowledge Distillation Techniques in Quantization-Aware Training

모델 학습 단계에서는 모델의 안정적인 수렴을 위해 32비트 부동소수점 (Float32) 정밀도가 사용된다. 그러나 NPU, 임베디드 등 엣지 디바이스와 같은 자원 제약 환경에 모델을 배포할 경우, 실시간 처리 성능을 확보하기 위해 정수형 기반의 양자화가 필수적으로 요구된다. 양자화 기법은 크게 PTQ 기법과 QAT 기법으로 구분된다. PTQ 기법은 학습이 완료된 모델을 대상으로 보정(calibration) 과정을 수행하여 각 층의 Scale 및 Zero-point를 추정함으로써 정밀도를 유지하는 방식이다. 그러나 이 과정에서 발생하는 주요 오차는 Scale 및 Zero-point의 부정확한 추정과 정수화 과정에서의 반올림 오차에 따른 정보 손실로부터 발생한다. 이러한 오차는 객체 탐지와 같은 분야와는 달리, 초해상화 분야에서는 저해상도 이미지를 고해상도로 복원하는 과정에서 경계, 질감, 세부 패턴과 같은 고주파 정보를 재생성하는 것이 핵심이다.

이런 고주파 성분은 진폭이 작고 변화가 급격하여 양자화 오차에 취약하다. 즉 고주파 성분은 픽셀 간 변화가 빠르기 때문에, 저비트 양자화 환경에서는 표현 정밀도가 낮아져 작은 진폭의 변화를 충분히 표현하지 못하므로, 고주파 성분이 쉽게 소실된다. 그 결과 경계가 흐려지고 질감이 손실되며, 동일한 절대 오차라도 고주파 영역에서는 상

Table 4. Quantitative Performance Comparison of Models with and without Knowledge Distillation Based on Local/Global Features (at 2× Scale-up). For each model, the original version applies the proposed method, while networks other than LSRNet lack local blocks, so the loss is computed only from global components

model	QAT	propsed method	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LSRNet	baseline		34.94	0.9462	30.94	0.8917	30.26	0.8665	28.21	0.8832	32.95	0.9536
	O	O	**34.79**	**0.9443**	**30.89**	**0.8889**	**30.21**	**0.8646**	**28.14**	**0.8827**	**32.91**	**0.9515**
		X	34.49	0.9408	30.74	0.8856	30.09	0.8599	27.95	0.8797	32.57	0.9481
FSRCNN [1]	baseline		33.65	0.9352	30.03	0.8791	29.56	0.8532	27.14	0.8468	31.57	0.9361
	O	O	**33.35**	**0.9305**	**29.88**	**0.8749**	**29.42**	**0.8461**	**26.97**	**0.8375**	**31.24**	**0.9314**
		X	33.16	0.9278	29.67	0.8721	29.34	0.8432	26.88	0.8349	31.11	0.9287
ESPCN [14]	baseline		34.54	0.9407	30.61	0.8836	29.96	0.8569	27.65	0.8676	32.13	0.9450
	O	O	**34.16**	**0.9384**	**30.44**	**0.8825**	**29.56**	**0.8532**	**27.45**	**0.8615**	**32.01**	**0.9415**
		X	34.02	0.9355	30.21	0.8787	29.33	0.8501	27.33	0.8597	31.89	0.9387
XLSR [15]	baseline		34.28	0.9419	30.47	0.8779	29.95	0.8593	27.33	0.8676	32.12	0.9464
	O	O	**34.22**	**0.9377**	**30.41**	**0.8765**	**29.83**	**0.8559**	**27.26**	**0.8641**	**31.93**	**0.9414**
		X	34.11	0.9356	30.29	0.8706	29.69	0.8499	27.21	0.8613	31.79	0.9368
Bicubic++ [16]	baseline		33.49	0.9336	30.17	0.8728	29.49	0.8536	26.68	0.8553	31.66	0.9403
	O	O	**33.37**	**0.9315**	**30.14**	**0.8706**	**29.42**	**0.8515**	**26.61**	**0.8540**	**31.61**	**0.9387**
		X	33.32	0.9301	30.06	0.8691	29.31	0.8501	26.41	0.8531	31.56	0.9365

대 오차가 크게 나타나 정보 손실이 더욱 심화 된다. Fig. 6.과 같이 초해상화에서는 출력 픽셀 단위의 미세한 오차도 PSNR, SSIM 등 품질 지표에 직접적으로 반영된다. 사람의 시각은 배경보다 경계와 질감의 손실에 훨씬 민감 하기에, 고주파 성분이 조금만 손실되어도 비교적 큰 차이를 느낄 수 있다.

PTQ 적용 모델 복원 결과 QAT 적용 모델 복원 결과

Fig. 6. Super-Resolution Results by Quantization Method

이러한 성능 저하의 문제점들을 해결하고자, 엣지 디바이스 탑재 시에는 PTQ 방식보다 QAT 방식이 주로 사용된다. QAT 방식은 학습 과정에서 양자화 연산을 모사함으로써 성능 저하를 줄일 수 있으나, 여전히 저비트 환경에서는 원본 모델 대비 품질 손실이 두드러지는 한계가 존재한다. 본 연구에서는 이러한 한계를 극복하기 위해, 앞서 식 (2)에서 정의한 지식 증류 기법을 QAT 과정에 통합 적용하였다. 구체적으로, 식 (2)의 각 손실 항목 중 모사 모델 관련 항은 모두 양자화 연산 $Q(\cdot)$을 거친 형태로 지역 손실 함수는 다음과 같이 정의된다. 양자화 전후의 특징맵은 동일한 크기와 채널을 유지하므로, 추가적인 변환 과정 없이 위와 같이 손실 함수를 정의할 수 있다.

$$L_{Local}^{QAT} = \frac{1}{N-1}\sum_{i=1}^{N-1} \alpha \cdot \|Q(f_s^{(i)}) - f_t^{(i)}\|_1 \quad (7)$$

지역/전역 기반의 손실 함수, 부드러운 정답 기반 손실 함수, 라벨 기반 지도학습 손실 함수 역시 동일한 방식으로 모사 모델 항에 $Q(\cdot)$을 적용하여 계산되며, 이는 앞서 제안한 단계적 증류 전략에 따라 순차적으로 수행하여 양자화 간 성능 손실을 최소화 한다.

이를 통해 양자화 추가 학습 과정에서 양자화로 인한 오차를 직접 반영할 수 있으며, 특히 고주파 성분의 손실을 최소화하도록 설계되었다. 이는 단순히 고주파 성분만을 강조하는 것이 아니라, 원본 모델의 전역적 구조와 색상 일관성을 함께 증류하기 때문에, 추가 학습 과정에서 가짜 물결무늬와 같은 울림 현상(Ringing)이나 경계선 주변에서의 명암이 튀는 현상(Overshoot) 과 같은 현상들을 방지할 수 있다.

제안한 기법의 성능 검증을 위해 경량화 전/후의 성능 비교 실험을 수행하였다. QAT 기법에서 추가 학습에 사용된 데이터셋과 하이퍼파라미터 설정은 3.2.2장에서 설정한 기존 학습 방식과 동일하게 유지하였으며, Epoch 수의 경

Fig. 7. Qualitative Comparison after Quantization with and without Local/Global Feature-based Knowledge Distillation (at 2× Scale-up)

우 기존 대비 20%로 축소하였고, 학습률의 경우 안정적인 수렴을 위하여 1×10−5로 설정하여 재학습을 진행하였다. 지식 증류에 활용한 원본 모델의 경우 사전 학습이 완료된 경량화 전 모델을 사용하여 학습을 진행하였다.

Table 4에서 확인할 수 있듯이, QAT 환경에서 제안한 지식 증류 기법을 적용한 결과, 일반적인 QAT 방식에 비해 정량적 성능 평가에서 일반적인 QAT 방식 대비 PSNR 감소폭을 1.15%에서 0.12%로 억제하여 성능 보존 효과를 입증하였다. 특히 구조적 유사성을 나타내는 SSIM 지표에서 성능 보존이 효과적으로 이루어졌으며, 이는 제안한 기법이 구조적 정보의 손실을 최소화하는 데 강점을 가진다는 것을 시사한다. 대부분의 데이터셋에서 PSNR 또한 원본 대비 성능 저하가 미미하며, 일반적인 QAT보다 높은 수치를 기록함으로써 제안 기법의 효과를 정량적으로 증명하였다. Fig. 7.의 정성적 평가 결과를 통해, 제안한 기법이 기존 QAT 방식에 비해 불필요한 노이즈를 효과적으로 억제함으로써 시각적 품질을 향상시킨다는 점을 확인할 수 있다. 양자화로 인한 성능 저하를 제안한 지식 증류 기법을 통해 보완함으로써, 복원 이미지의 노이즈를 효과적으로 억제하였다.

또한, 본 논문에서 제안한 지식 증류 기법의 성능 저하 보존 효과를 검증하기 위해 다양한 경량 초해상화 모델에서도 실험을 수행하였다. 일반적인 QAT 방식과 비교했을 때, 제안하는 지식 증류 기법을 적용하면 Table 4.에서 확인 할수 있듯이, 해당 기법 적용 전후로 PSNR과 SSIM에서 더 우수한 성능을 확인할 수 있었다. 이는 제안 기법이 단순히 양자화 손실을 방어하는 것을 넘어, 경량화된 구조에서도 고주파 디테일을 효과적으로 복원하여 정량적 성능을 유의미하게 개선 및 보존함을 입증하였다.

결과적으로, 지역/전역 특징 기반 지식 증류 기법은 저비트 양자화 환경에서도 제안한 기법은 전역 특징을 통해서 전체적인 색감과 구조적 일관성을 유지하고, 지역 특징을 통해 경계 질감 등 고주파 정보를 효과적으로 복원함으로써, 저비트 양자화 환경에서도 성능 저하 없이 원본에 가까운 품질을 달성하였다. 제안한 LGR-Net 모델 뿐만 아니라, Table 4.에서 확인할 수 있듯이 다른 경량 모델들에서도 기존 QAT 방식보다 해당 지식 증류 기법을 적용 시 더 높은 성능을 보였다. 복원 영상의 비교 결과, 기존 방식 대비 경계·질감·세부 패턴과 같은 고주파 정보가 더욱 충실히 재현되었으며, 원본 모델 대비 시각적 차이를 최소화하여, 유사한 색감과 품질을 확보하였다. 특히 Table 4.에서 확인할 수 있듯이, Urban100 [22]과 Manga109 [23]처럼 복잡한 텍스처와 세밀한 디테일이 많은 데이터셋에서 정량적으로 성능 저하 폭이 유의미하게 줄어들어, 제안한 접근 방식의 효과성을 정량적, 정성적 실험을 통해 타당성을 입증하였다.

IV. Conclusions

본 연구에서는 엣지 디바이스 환경에서 실시간으로 동작 가능한 초해상화 모델을 설계하고, 구조적 경량화 및 양자화 과정에서의 성능 저하를 최소화하기 위해 지역/전역 특징 기반의 지식 증류 기법을 제안하였다. 제안한 기법은 지역과 전역 특징을 분리하여 학습에 반영함으로써, 실시간 처리가 요구되는 환경에서도 우수한 복원 성능을 유지할 수 있음을 확인하였다.

세부적으로 해당 기법에서는, 지역 특징의 경우 블록단위에서의 공간 영역에 대한 L1 손실을 적용하여 경계와 세부 구조를 직접적으로 학습하도록 하였으며, 전역 특징의 경우에는 2D 푸리에 변환을 통해 주파수 스펙트럼을 추출한 뒤 고주파 마스크를 적용한 손실 함수를 설계하여 공간적 위치에 독립적인 전역 특성을 반영하면서 고주파 성분을 강조하여 디테일 복원 능력을 강화하도록 설계하였다. 제안한 지역/전역 특징 기반 지식 증류 전략은 경량화 및 양자화된 모델에서도 이미지 내 의미적 맥락과 공간적 구조를 동시에 고려한 초해상화 복원을 가능하게 하며, 단순한 픽셀 정렬을 넘어 구조적 일관성과 질감 복원력을 함께 확보할 수 있는 효과적인 초해상화 복원 전략을 제공한다.

또한, 실시간 처리를 보장하기 위해 활성화 함수 별 분석 결과를 바탕으로 네트워크 구조를 최적화하고, 지역/전역 잔차 연산을 도입하여 연산 효율성을 향상시켰다. 이를 통해 경량 초해상화 모델의 성능 저하를 최소화하면서도 우수한 복원 성능을 유지할 수 있었다. 뿐만 아니라, 기존 양자화 방식인 PTQ 기법과 QAT기법은 양자화 과정에서 고주파 성분 손실과 구조적 왜곡으로 인해 초해상화 복원 성능 저하가 발생하는 한계가 있었다. 본 연구에서는 이러한 문제를 극복하기 위해 QAT기반의 양자화 과정에 지역/전역 특징 기반 지식 증류 전략을 병행 적용하였다. 그 결과, 양자화 환경에서 두드러지게 나타나는 고주파 정보 손실을 효과적으로 보완하여, 경량 모델에서도 디테일 보존과 구조적 정확성을 동시에 달성할 수 있음을 확인하였다.

이는 경량 초해상화 모델이 가지는 한계를 극복하고, 저전력/저사양 환경에서도 실시간으로 고품질 영상을 복원할 수 있는 가능성을 제시하였다. 향후 연구에서는 제안한 지식 증류 전략을 영상 복원, 압축 영상 개선 등 다양한 영상 처리 과제에 확장하고, 다양한 감시 정찰 장비 운용 환경을 통해 제안하는 방법의 활용 가능성을 검증할 예정이다.

REFERENCES

[1] Dong, C., Loy, C. C., He, K., Tang, X., "Image Super-Resolution Using Deep Convolutional Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.38, No.2, pp.295-307, Feb. 2016. DOI: 10.1109/TPAMI.2015.2439281

[2] Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K., "Enhanced Deep Residual Networks for Single Image Super-Resolution," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp.136-144, Honolulu, USA, July 2017. DOI: 10.1109/CVPRW.2017.151

[3] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," European Conference on Computer Vision (ECCV), pp.286-301, Munich, Germany, Sept. 2018. DOI: 10.1007/978-3-030-01234-2_18

[4] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., "SwinIR: Image Restoration Using Swin Transformer," IEEE International Conference on Computer Vision (ICCV) Workshops, pp.1833-1844, Montreal, Canada, Oct. 2021. DOI: 10.1109/ICCVW54120.2021.00210

[5] Lee, J. W., Yoon, S. W., Lee, K. C., "Transformer-Based Deep Learning Models for ERS SAR Image Super-Resolution," Korean Journal of Remote Sensing, Vol.41, No.1, pp.143-152, Feb. 2025. DOI: 10.7780/kjrs.2025.41.1.12

[6] Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., Dong, C., "Activating More Pixels in Image Super-Resolution Transformer," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.22367-22377, Vancouver, Canada, June 2023. DOI: 10.1109/CVPR52729.2023.02148

[7] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D., "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2704-2713, Salt Lake City, USA, June 2018. DOI: 10.1109/CVPR.2018.00286

[8] AMD Research, "Best Practices for Post-Training Quantization (PTQ)," AMD Quark PyTorch Documentation, https://quark.docs.amd.com/latest/pytorch/quark_torch_best_practices.html

[9] Kim, J., Lee, J. K., Lee, K. M., "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1646-1654, Las Vegas, USA, June 2016. DOI: 10.1109/CVPR.2016.182

[10] Ren, B., Guo, H., Sun, L., Wu, Z., Timofte, R., Li, Y., Zhang, Y., Chai, X., Cheng, Z., Qin, Y., Yang, Y., Song, L., Yu, H., Xu, P., Wan, C., Huang, Z., Guo, P., Cui, S., Li, C., Hu, X., Pan, P., Zhang, X.,ss Zhang, H., Luo, Q., Jiang, L., Lei, H., Gao,

Q., Li, Y., Luo, W., Li, T., Wang, Q., Liu, Y., Wang, Y., An, H., Zhang, L., Zhao, S., Song, L., Sun, L., Pan, J., Dong, J., Tang, J., Wei, J., Wang, M., Guo, R., Wang, Q., Liu, Q., Cheng, Y., et al., "The Tenth NTIRE 2025 Efficient Super-Resolution Challenge Report," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.1-15, Vancouver, Canada, June 2025. DOI: 10.48550/arXiv.2504.10686

[11] Hinton, G., Vinyals, O., Dean, J., "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531, 2015.

[12] Mansourian, A. M., Ahmadi, R., Ghafouri, M., Babaei, A. M., Badali Golezani, E., Yasamani Ghamchi, Z., Ramezanian, V., Taherian, A., Dinashi, K., Miri, A., Kasaei, S., "A Comprehensive Survey on Knowledge Distillation," arXiv preprint arXiv:2503.12067, 2025. Available at: https://arxiv.org/abs/2503.12067

[13] Zhang, Y., Chen, H., Chen, X., Deng, Y., Xu, C., Wang, Y., "Data-Free Knowledge Distillation for Image Super-Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.7852-7861, Nashville, USA, June 2021. DOI: 10.1109/CVPR46437.2021.00776

[14] Hoo, V., "FAKD: Feature-Affinity Based Knowledge Distillation for Efficient Image Super-Resolution," IEEE International Conference on Image Processing (ICIP), pp.1726-1730, Abu Dhabi, UAE, Oct. 2020. DOI: 10.1109/ICIP40778.2020.9191212

[15] Fuoli, D., Van Gool, L., Timofte, R., "Fourier Space Losses for Efficient Perceptual Image Super-Resolution," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.2340-2349, Oct. 2021. DOI: 10.1109/ICCV48922.2021.00236

[16] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., Keutzer, K., "A Survey of Quantization Methods for Efficient Neural Network Inference," arXiv preprint arXiv:2103.13630, 2021. Available at: https://arxiv.org/abs/2103.13630

[17] Wan, C., Yu, H., Li, Z., Chen, Y., Zou, Y., Liu, Y., Yin, X., Zuo, K., "Swift Parameter-free Attention Network for Efficient Super-Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.1-10, Seattle, USA, June 2024. DOI: 10.48550/arXiv.2311.12770

[18] Elfwing, S., Uchibe, E., Doya, K. "Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning," Neural Networks, Vol.107, pp.3-11, 2018. DOI: 10.1016/j.neunet.2017.12.012

[19] He, K., Zhang, X., Ren, S., Sun, J. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.1026-1034, 2015. DOI: 10.1109/ICCV.2015.123

[20] 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," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1874-1883, Las Vegas, USA, June 2016. DOI: 10.1109/CVPR.2016.207

[21] Babu, A., Touvron, H., Vedaldi, A., Lanchantin, J., "XLSR: Cross-Scale Cross-Reference for Efficient Super-Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.4186-4196, New Orleans, USA, June 2022. DOI: 10.1109/CVPR52688.2022.00416

[22] Bilecen, B. B., Ayazoglu, M., "Bicubic++: Slim, Slimmer, Slimmest - Designing an Industry-Grade Super-Resolution Network," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1-10, Vancouver, Canada, June 2023. DOI: 10.48550/arXiv.2305.02126

[23] Agustsson, E., Timofte, R., "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.126-135, Honolulu, USA, July 2017. DOI: 10.1109/CVPRW.2017.150

[24] Liang, J., Zhang, K., Van Gool, L., Timofte, R., "LSDIR: A Large-Scale Dataset for Image Restoration," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1-10, Vancouver, Canada, June 2023. DOI: 10.48550/arXiv.2305.03045

[25] Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M. L., "Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding," Proceedings of the British Machine Vision Conference (BMVC), pp.1-10, Surrey, UK, Sept. 2012. DOI: 10.5244/C.26.16

[26] Zeyde, R., Elad, M., Protter, M., "On Single Image Scale-Up Using Sparse-Representations," Curated Technical Report, pp.1-11, Israel, 2010. Website. http://www.cs.technion.ac.il/~elad/publications

[27] Martin, D., Fowlkes, C., Tal, D., Malik, J., "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.416-423, Vancouver, Canada, July 2001. DOI: 10.1109/ICCV.2001.937655

[28] Huang, J. B., Singh, A., Ahuja, N., "Single Image Super-Resolution from Transformed Self-Exemplars," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5197-5206, Boston, USA, June 2015. DOI: 10.1109/CVPR.2015.7299156

[29] Fujimoto, D., Ogawa, T., Yamasaki, T., Aizawa, K., "Manga109 Dataset and Creation of Metadata," Proceedings of the IEEE MultiMedia Information Processing and Retrieval (MIPR), pp.484-487, Miami, USA, April 2019. DOI: 10.1109/MIPR.2019.00098

Authors

Ho-min Jung received the B.S. degree in Information and Communication Engineering from Sunmoon University, Asan, South Korea, in 2017, and the M.S. degree in Electrical and Electronic Engineering from Kyungpook National University (KNU), Daegu, South Korea, in 2021. He is currently a Researcher with the S/W Team (Intelligent), Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include super resolution, quantization and optimization in AI.

Tae-Young Lee received the B.S. degree in information and control engineering from robotics school, Kwangwoon University, Seoul, South Korea, in 2009, and the M.S. degree in control and instrumentation engineering from robotics school in Kwangwoon University, Seoul, South Korea, in 2011. He is currently a Senior Researcher with the S/W Team (Intelligent), Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include object detection, object tracking and segmentation with deep Learning also generative AI.

Byung-In Choi received the B.S., M.S., and Ph.D. degrees in electronic engineering from Hanyang University in Seoul, South Korea, in 2001, 2003, and 2008, respectively. He is currently a Leader of the S/W team(Intelligent) Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include object detection, object tracking, and super resolution.