특집논문 (Special Paper)
방송공학회논문지 제25권 제6호, 2020년 11월 (JBE Vol. 25, No. 6, November 2020)
https://doi.org/10.5909/JBE.2020.25.6.898
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

초고해상도 복원에서 성능 향상을 위한 다양한 Attention 연구

문 환 복[a], 윤 상 민[b)‡]

A Study on Various Attention for Improving Performance in Single Image Super Resolution

Hwanbok Mun[a] and Sang Min Yoon[b)‡]

요 약

컴퓨터 비전에서 단일 영상 기반의 초고해상도 영상 복원의 중요성과 확장성으로 관련 분야에서 많은 연구가 진행되어 왔으며, 최근 딥러닝에 대한 관심이 증가하면서 딥러닝을 활용한 단안 영상 기반 초고해상도 연구가 활발히 진행되고 있다. 대부분의 딥러닝을 기반으로 하는 단안 영상 기반 초고해상도 복원 연구는 복원 성능을 향상시키기 위해 네트워크의 구조, 손실 함수, 학습 방법에 초점이 맞추어 연구가 진행되었다. 한편, 딥러닝 네트워크를 깊게 쌓지 않고 초고해상도 영상 복원 성능을 향상시키기 위해 추출된 특징 맵을 강조하는 Attention Module에 대한 연구가 다양한 분야에 적용되어 왔다. Attention Module은 다양한 관점에서 네트워크의 목적에 맞는 특징 정보를 강조 및 스케일링 한다. 본 논문에서는 초고해상도 복원 네트워크를 기반으로 다양한 구조의 Channel Attention과 Spatial Attention을 설계하고, 다양한 관점에서 특징 맵을 강조하기 위해 다중 Attention Module 구조를 설계하여 성능을 분석 및 비교한다.

Abstract

Single image-based super-resolution has been studied for a long time in computer vision because of various applications. Various deep learning-based super-resolution algorithms are introduced recently to improve the performance by reducing side effects like blurring and staircase effects. Most deep learning-based approaches have focused on how to implement the network architecture, loss function, and training strategy to improve performance. Meanwhile, Several approaches using Attention Module, which emphasizes the extracted features, are introduced to enhance the performance of the network without any additional layer. Attention module emphasizes or scales the feature map for the purpose of the network from various perspectives. In this paper, we propose the various channel attention and spatial attention in single image-based super-resolution and analyze the results and performance according to the architecture of the attention module. Also, we explore that designing multi-attention module to emphasize features efficiently from various perspectives.

Keyword : Super-Resolution, Attention Module, Image Decomposition

Ⅰ. 서 론

컴퓨터 비전 및 컴퓨터 그래픽스의 한 분야인 초고해상도 복원은 카메라 센서 혹은 주변 환경으로 인해 열화되거나 정보가 손실된 저해상도의 영상을 고해상도의 영상으로 복원하는 것으로 정의된다. 초고해상도 복원은 단순히 입력 영상의 크기만을 키우는 것을 넘어서서 영상의 섬세한 부분이나 중요한 정보를 복원할 수 있기 때문에 그 중요성으로 인해 많은 연구가 진행되어 왔다. 하지만 하나의 저해상도 영상 혹은 영상 내 픽셀에서 복원될 수 있는 고해상도 영상 혹은 픽셀과 경우의 수는 매우 많고 고유할 수 없기 때문에 Ill-posed problem으로 여겨지며, 저해상도 영상으로부터 가장 원본 혹은 정답에 가까운 고해상도 영상을 얻기 위해 다양한 방법들이 연구되었다.

Dong[1]등은 기존의 Reconstruction기반의 초고해상도 복원 방법을 3개의 Convolution Layer로 구현한 SRCNN을 소개하였으며, SRCNN같은 초기 연구를 시작으로 딥러닝을 활용한 많은 초고해상도 복원 방법들이 제안되었다. 초기의 딥러닝 기반의 초고해상도 복원 연구들은 깊은 네트워크를 효율적으로 설계하거나 네트워크를 학습하는 방법에 대한 연구들이 주를 이뤄왔으며 다양한 관점에서 복원 성능을 평가하기 위해 RGB가 아닌 고차원의 특징 맵으로 손실 함수를 설계한 연구들도 제안되었다.

한편, 네트워크의 성능을 향상시키기 위해 네트워크 구조나 학습 방법이 아닌 네트워크로 부터 추출된 특징 맵에 관심을 갖는 연구들도 있었다. SE-Net[2]의 Channel Attention은 암묵적으로 고려되던 채널 간 상관관계를 명시적으로 고려하도록 하였다. Channel Attention은 중요한 정보를 포함하는 채널의 강조 및 스케일링하여 네트워크 깊이는 유지하면서 네트워크의 표현 능력과 성능을 향상시킬 수 있다. SCA-CNN[3]은 네트워크의 목적과 관련이 있는 특징 맵의 공간을 강조하는 Spatial Attention을 제안하였으며, 이를 통해 네트워크가 목적이나 성능에 관련이 있는 특징 맵의 중요한 부분에 집중하도록 하였다.

Attention Module이나 특징 맵 자체에 대한 연구들이 소개되면서 보다 정확한 복원을 위해 초고해상도 복원에서도 Attention Module을 적용하는 방법들이 제안되었다. RCAN[4]은 Channel Attention을 적용해서 영상 복원에 도움이 되는 유용한 채널을 강조하고 복원 성능을 향상시켰으며, SAN[5]은 채널 간의 Covariance를 활용한 Second order Channel Attention을 제안하여 채널들 간의 상관관계를 잘 파악하고 분별력을 키웠다.

본 논문에서는 정확한 복원을 위해 Structure와 Texture를 분리하여 복원하는 영상 분할 기반 초고해상도 복원 네트워크를 제안한다. 제안하는 네트워크의 손실 함수 가중치와 학습의 균형을 맞추기 위해 GradNorm[14]을 적용한다. 또한 네트워크의 복원 성능을 향상시키기 위해 다양한 구조의 Attention Module과 배치 방법을 설계하였으며 실험을 통해 구조와 배치에 따른 성능을 비교 및 분석한다.

Ⅱ. 관련 연구

1. 초고해상도 복원

컴퓨터 비전을 비롯한 다양한 연구 분야에서 딥러닝이 활용되기 이전 전통적인 초고해상도 복원 방법에는 주변에 인접한 픽셀들과 가중치를 이용하여 복원하는 Bilinear와 Bicubic같은 보간법이 있었으며 저해상도 영상 패치와 고해상도 영상 패치의 관계를 학습한 사전을 이용하는 Reconstruction기반 방법들이 있었다.

Dong등은 Reconstruction 기반 복원 방법의 특징 추출, 맵핑, 복원 3단계를 3개의 Convolution Layer로 구현해서

a) 국민대학교 컴퓨터공학과(Department of Computer Science, Kookmin University)
b) 국민대학교 소프트웨어융합대학(College of Computer Science, Kookmin University)
‡ Corresponding Author : 윤상민(Sang Min Yoon)
E-mail: smyoon@kookmin.ac.kr
Tel: +82-2-910-4645
ORCID: https://orcid.org/0000-0003-0001-1845
※ 이 논문의 연구 결과 중 일부는 “2020년 한국방송·미디어공학회 하계 학술대회”에서 발표한 바 있음.
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.2020-0-01826, AI 기반 선도적 실전문제 해결 연구인재 양성)
※ This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korean Government(MSIP)(No. Grant Number – 2015R1A5A7037615)
※ Following(or This research) was results of a study on the "HPC Support" Project, supported by the ‘Ministry of Science and ICT’ and NIPA.
· Manuscript received September 17, 2020; Revised October 21, 2020; Accepted October 21, 2020.

기존 방법보다 높은 복원 성능을 보였다. 이후 VDSR[6]은 복원 성능을 향상시키기 위해 보다 깊은 네트워크와 Residual Connection을 적용한 네트워크를 제안하였으며, DRCN[7]과 DRRN[8]은 가중치 공유 및 반복적인 적용을 통해 적은 파라미터로도 초고해상도 복원을 할 수 있음을 보였다. EDSR[9]은 깊은 네트워크를 효율적으로 학습시키기 위해 여러 개의 Residual Block과 Skip Connection을 사용하여 네트워크를 설계하였으며, L1-Norm 손실 함수를 사용해서 복원 영상이 흐릿한 문제를 완화시킴과 동시에 복원 성능을 향상시켰다. ESPCN[10]은 Upsampler의 높은 계산 복잡도를 개선하기 위해 특징 맵을 재배치하는 방식의 Pixel-Shuffler를 제안하였다. MemNet[11]과 SRDenseNet[12]은 네트워크의 모든 레이어에서 추출된 특징 맵을 연결하고 활용하는 Dense Connection을 도입해서 저차원부터 고차원까지 다양한 특징과 정보를 이용한 초고해상도 복원 네트워크를 제안하였다.

딥러닝의 발전으로 초고해상도 복원 연구는 많은 성과가 있었지만, 섬세한 선분이나 문양 같은 Texture를 복원하는 것은 여전히 어려운 문제이다. 본 논문에서는 정확한 Texture를 복원하기 위해서 Structure와 Texture를 분리 및 복원하는 네트워크를 제안한다.

2. Attention Module

Attention Module은 네트워크의 성능을 향상시키기 위해 특징 맵에서 중요한 정보를 포함하고 있는 채널을 강조 및 스케일링한다. 영상 분류 분야에서 제안된 SE-Net[2]은 Squeeze and Excitation 혹은 Channel Attention으로 불리는 Attention Module을 사용해서 기존에 암묵적으로 고려되던 채널의 상관관계를 명시적으로 고려하도록 만들었으며, 영상 분류에 도움이 되는 정보를 포함하는 특징 맵의 채널을 강조 및 스케일링해서 네트워크의 깊이를 유지함과 동시에 분류 정확도를 높였다. Image Captioning 분야에서 제안된 SCA-CNN[3]은 영상에서 관심 있는 객체에 집중하고 영상을 잘 표현하는 자막을 만들기 위해 Channel Attention과 Spatial Attention을 사용해서 특징 맵의 채널과 공간 정보를 강조하였다.

초고해상도 복원에서도 정확한 복원을 위해서 Attention Module을 적용 및 응용한 연구들이 소개되었다. RCAN[4]은 Residual In Residual 네트워크 구조를 기반으로 Channel Attention을 적용하여 중요한 정보를 포함하고 있는 특징 맵의 채널을 스케일링 및 강조하였으며, CSFM[13]은 Channel Attention과 Spatial Attention을 동시에 적용한 Residual Block과 DenseNet을 기반으로 Multi-Level에서 전역적 및 지역적 정보를 활용한 초고해상도 복원을 하였다. SAN[5]은 복잡한 채널 간의 관계를 보다 정확하게 고려하기 위해 Covariance를 기반으로 채널의 대표 값을 구하고 강조하는 방법을 제안하였다.

Ⅲ. 본 론

본 논문에서 제안하는 네트워크는 효율적으로 학습하기 위해 RCAN[4]의 Residual In Residual 구조를 기반으로 하며 Residual Block에 Attention Module을 적용하여 추출된 특징 맵을 강조한다. 또한 제안하는 네트워크는 하나의 입력과 두개의 출력을 갖기 때문에 멀티태스크 네트워크 구조로 볼 수 있으며, 네트워크의 학습의 균형과 손실함수의 가중치를 최적화하기 위해 GradNorm[14]을 적용했다.

1. Network Structure

본 논문에서 제안하는 네트워크는 다음 그림1과 같이 저해상도 영상I_{LR}을 입력으로 받아 복원된 고해상도의 StructureS_{SR}와 TextureT_{SR}영상을 출력하고 두 영상을 더해서 최종적으로 초고해상도 복원 영상I_{SR}을 얻는다. 제안하는 네트워크의 복원 과정을 수식으로 표현하면 다음 수식 1과 같다.

$$I_{SR} = S_{SR} + T_{SR} = F(I_{LR}) \tag{1}$$

제안하는 네트워크 F 는 Feature Extractor와 2개의 SR Network로 이루어져 있으며 각 네트워크들은 효율적인 학습과 고차원의 특징 추출을 위해 Residual Group과 Residual Block으로 구성된다. 하나의 Residual Group는 여러 개의 Residual Block과 Skip Connection으로 구성되어 있으며 각 Residual Block은 2개의 Convolution Layer, Skip Connection으로 구성된다. 제안하는 Attention Module

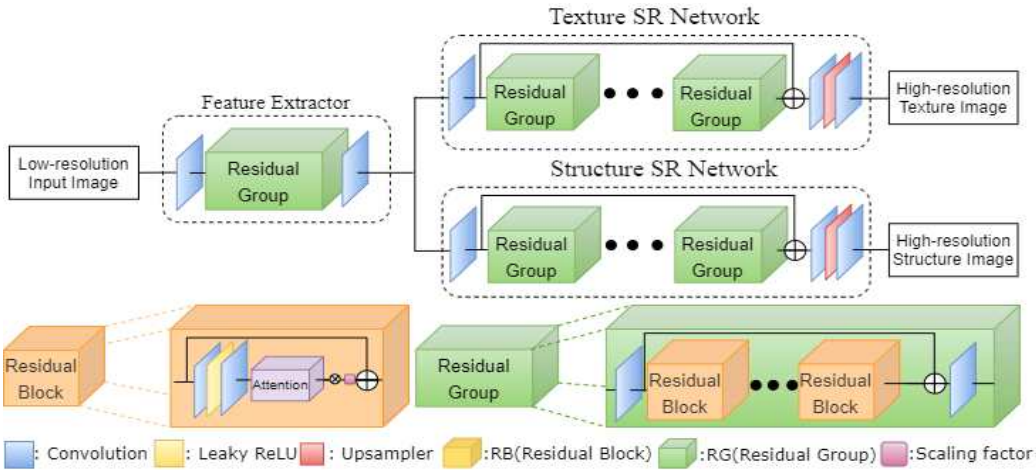


그림 1. 제안하는 네트워크 구조
Fig. 1. Proposed Network Architecture

은 Residual Block 내에 위치하며 추출된 특징을 강조하는 데에 사용된다.

네트워크의 가장 앞에 위치한 Feature Extractor는 저해상도 입력 영상을 RGB채널에서 고차원의 특징 맵으로 변환함과 동시에 유용한 특징 맵을 추출한다. 추출된 고차원의 특징 맵은 다양한 정보들을 포함하고 있으며 초고해상도 복원을 위한 두개의 SR Network의 입력으로 전달된다. Texture SR Network와 Structure SR Network는 Feature Extractor로 추출된 특징 맵을 입력받아 각 네트워크의 목적에 맞는 정보를 추출 및 복원하며 Upsampler로 ESPCN의 PixelShuffler를 사용하였다.

제안하는 Channel Attention과 Spatial Attention은 Residual Block에 Convolution Layer 다음에 위치하여 추출된 특징 맵의 채널 혹은 공간 정보를 강조한다. 다음 표 1은 제안하는 네트워크에 대한 세부 사항이다.

2. 손실 함수

제안하는 네트워크를 학습하기 위한 손실 함수는 다음 수식 2와 같다.

$$L = \lambda_T L_T + \lambda_S L_S \tag{2}$$

L은 전체 네트워크에 대한 손실 함수로, 두 SR Network의 손실 함수인 L_T, L_S와 각각에 대한 가중치인 λ_S와λ_T로 계산된다. L_S와 L_T의 정의는 다음과 같다.

$$L_T = \|T_{SR} - T_{HR}\|_1,\ L_S = \|S_{SR} - S_{HR}\|_1 \tag{3}$$

T_{SR}과 S_{SR}은 복원한 고해상도의 Texture와 Structure 영상이며, T_{HR}과 S_{HR}은 각 복원된 Texture와 Structure의 참조 영상이다. 영상에서 Structure와 Texture에 대한 명확한 정의나 Ground Truth은 존재하지 않기 때문에 기존 영상 분할 알고리즘으로 학습을 위한 참조 영상을 만들어서 학습에 사용했으며, L_T와 L_S는 복원 영상과 참조 영상의 차이의 L1-Norm으로 계산된다. 한편, 제안하는 네트워크는 하나의 입력과 두개의 출력을 갖기 때문에 멀티 태스크 네트워크로 볼 수 있으며, Feature Extractor와 SR Network는

표 1. 네트워크 세부사항
Table 1. Network Specification

Feature Extractor			SR Network			Local Skip Connection	Global Skip Connection
#Conv in RR	#RB in RG	#RG in Network	#Conv in RR	#RB in RG	#RG in Network		
3	10	1	2	10	3	O	O

각각 루트 네트워크와 파생된 브런치 네트워크라고 할 수 있다. 각 브런치 네트워크는 학습 시 다른 브런치 네트워크의 손실 함수에 의해 영향을 받지 않으며 독립적으로 학습된다. 하지만 루트 네트워크는 두 브런치 네트워크에 연결되어 있기 때문에 두 브런치 네트워크에 의존적이며 동시에 영향을 받는다. 대부분의 멀티태스크 네트워크는 학습의 방향성 혹은 목적에 따라 실험적으로 손실 함수의 가중치를 정하고 학습한다. 그러나 이러한 실험적 가중치 결정 방법은 시간과 비용이 발생시킬 수 있으며 결정된 가중치는 최적이 아닐 수 있다.

본 논문에서는 네트워크 학습 시 발생할 수 있는 불균형 문제를 완화하고 최적의 가중치를 찾기 위해 GradNorm을 적용하여 가중치 λ_S와 λ_T를 결정하고 학습한다.

3. GradNorm

본 논문에서는 네트워크의 학습 균형을 맞추고 최적의 가중치를 찾기 위해 GradNorm[14]을 적용한다. GradNorm은 학습 시 각 브런치 네트워크에서 루트 네트워크로 전달되는 기울기와 브런치 네트워크의 학습률을 고려하여 손실 함수의 가중치를 결정하고 학습의 균형을 맞춘다. GradNorm을 적용하기 위한 손실 함수는 다음과 같다.

$$\mathcal{L}_{grad} = \sum_{i \in \mathrm{S,T}} |G_W^i(t) - \overline{G}_W(t) \times [r_i(t)]^{\alpha}|_1 \quad (4)$$

브런치 네트워크에서 루트 네트워크로 전달되는 기울기를 측정하기 위해 Feature Extractor의 가장 마지막에 위치한 Convolution Layer를 타겟 레이어 W로 지정하였으며 GradNorm의 손실 함수를 구성하는 각 항의 정의는 다음과 같다.

- $G_W^S(t) = \|\nabla_W \lambda_S(t) L_S(t)\|_2$, $G_W^S(t) = \|\nabla_W \lambda_S(t) L_S(t)\|_2$: 네트워크 학습 시 각 브런치 네트워크에서 타겟 레이어 W로 전달되는 기울기의 L2-Norm이다. 루트 네트워크가 학습 시 각 브런치 네트워크로 부터 영향 받는 정도를 정량적으로 나타낸 값이며 GradNorm은 손실 함수의 가중치 λ_S와 λ_T를 조절하여 위 두 기울기가 같아지도록 학습한다.
- $\overline{G}_W(t) = (G_W^S(t) + G_W^T(t))/2$: 두 기울기의 L2-Norm의

표 2. 네트워크 학습 알고리즘
Table 2. Network training strategy

Initialize:
  Network weight parameters F
  Set parameters $\lambda_S(0) = \lambda_T = 1$
  Choice α and the target layer W
Iteration t=0 to train_step :
  Given batch image (I_{LR}, $S_{HR_{ref}}$, $T_{HR_{ref}}$)
  Estimate S_{HR} and T_{HR} using $F(I_{LR})$
  Compute network loss $\mathcal{L}(t)$ with $\lambda_S(t)$ and $\lambda_T(t)$
  Compute Grad loss $\mathcal{L}_{grad} = \sum_{i \in \mathrm{S,T}} |G_W^i(t) - \overline{G}_W(t) \times [r_i(t)]^{\alpha}|_1$
  Compute gradient of network loss $\nabla_F \mathcal{L}(t)$
  Compute gradient of Grad loss $\nabla_{\lambda_{ST}} \mathcal{L}_{grad}(t)$
  Update :
    Update network F using $\nabla_F \mathcal{L}(t)$
    Update weight $\lambda_{S(t)}$ and $\lambda_{T(t)}$ using $\nabla_{\lambda_{ST}} \mathcal{L}_{grad}(t)$
    Normalize $\lambda_S(t+1)$ and $\lambda_T(t+1)$ so that $\sum_{i \in \{\mathrm{S,T}\}} \lambda_i = 2$
end

평균값이다.

- $\tilde{L}_S(t) = L_S(t) / L_S(0)$, $\tilde{L}_T(t) = L_T(t) / L_T(0)$: 각 브런치 네트워크의 손실 함수의 학습 초기 대비 감소율이며 각 브런치 네트워크의 학습률을 나타내는 지표로 사용한다.
- $r_{ST}(t) = \tilde{L}_{ST}(t) / E_{task}[\tilde{L}_{ST}(t)]$: 각 브런치 네트워크의 상대적인 학습률이다.

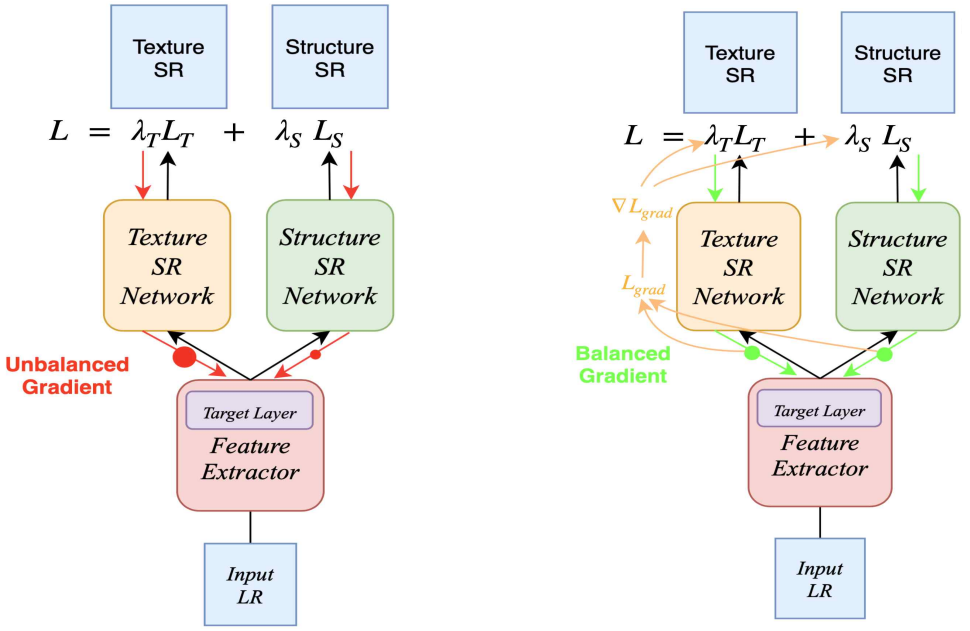
파리미터 α는 클수록 브런치 네트워크의 학습률을 크게 고려하고 0에 가까울수록 기울기의 크기만을 고려하여 손실 함수의 가중치를 업데이트한다. 제안하는 논문에서는 α를 0.1로 설정하고 실험을 진행하였다. 매 학습 시 업데이트된 손실 함수 가중치 λ_S와 λ_T는 합이 2가 되도록 재정규화 과정을 거친다. 다음 표 2는 GradNorm을 적용한 네트워크 학습 알고리즘이며 그림 2는 Grad- Norm을 적용한 네트워크의 학습 시 기울기 전달을 표현한 그림이다.

4. Attention

본 논문에서는 Attention Module을 사용하여 Residual Block에서 추출된 특징 맵을 강조 및 스케일링 한다. 또한 특징 맵의 채널과 공간 정보를 동시에 강조하기 위해 Channel Attention과 Spatial Attention을 배치하는 방법을 설계하고 성능을 비교 및 분석 한다.

- Channel Attention

Channel Attention은 특징 맵의 채널 간 상관관계를 이용하여 네트워크의 목적에 맞는 유용한 정보를 강조한다. Channel Attention은 입력 특징 맵 F_{in}을 입력으로 받아 Global Average Pooling을 통해 특징 맵의 각 채널을 대표하는 채널 크기의 벡터를 생성한다. 이후 벡터는 Fully Connected Layer와 LeakyReLU를 통해 비선형성을 가짐과 동시에 유용한 정보가 압축된 잠재 벡터로 변환된다. 잠재 벡터는 다음 Fully Connected Layer와 Sigmoid를 통해 0부터 1사이의 값을 갖는 채널 크기의 스케일링 벡터가 된다. 해당 벡터는 입력 특징 맵 F_{in}을 스케일링 해서 최종적으로 유용한 정보를 포함하는 채널이 강조된 특징 맵 F_{out}을 출력한다. Channel Attention을 수식으



(a) Network without GradNorm (b) Network with GradNorm(Proposed)

그림 2. GradNorm 적용 네트워크 학습 비교
Fig. 2. Comparison networks using GradNorm

로 표현하면 다음 수식 5와 같다.

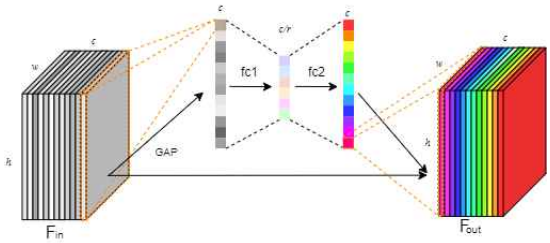
$$F_{out} = CA(F_{in}) \tag{5}$$

Channel Attention은 Global Average Pooling과 마지막 Fully Connected Layer 사이의 압축된 벡터의 구조와 Fully Connected Layer의 개수에 따라 다양한 형태를 가질 수 있으며 다음 그림 3은 설계한 다양한 형태의 Channel Attention 구조와 수식이다.

제안하는 Channel Attention은 3개로 Global Average Pooling과 마지막 Fully Connected Layer를 사이에 Fully Connected Layer의 개수에서 차이를 갖는다. One layer Channel Attention은 가장 기본적인 형태의 Channel Attention이며 한 개의 Fully Connected Layer를 적용한다. Two layer Channel Attention은 보다 유용한 정보를 갖는 벡터를 만들기 위해서 2개의 Fully Connected Layer를 사용했으며 잠재 벡터는 똑같은 길이의 벡터로 다시 한 번 추출된 뒤 입력 채널 크기의 벡터로 변환한다. 마지막 Three layer Channel Attention은 앞선 Two layer Channel Attention과 같이 복잡하고 유용한 정보를 찾기 위해 3개의 Fully Connected Layer를 사용하였다. 각 Channel Attention에서 마지막에 위치한 활성화 함수는 강조된 벡터가 0부터 1사이의 값을 가지도록 만들기 위한 Sigmoid이며 이를 제외한 나머지 활성화 함수는 LeakyReLU이다. c는 입력 특징 맵의 채널의 크기이며 r은 상수로 본 논문에서는 실험상 4의 값의 적용하였다.

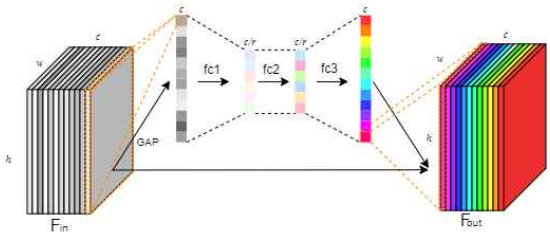
• Spatial Attention

Spatial Attention은 Convolution Layer를 이용해서 네트워크의 목적과 연관이 있는 특징 맵의 공간 정보를 강조 및 스케일링한다. Spatial Attention은 입력 특징 맵 F_{in}에서 Convolution Layer와 Sigmoid를 이용하여 0부터 1사이의 값을 갖는 1채널의 스케일링 특징 맵을 만든다. 스케일링 특징 맵은 입력 특징 맵F_{in}과 같은 높이와 너비를 가지며 공간 정보를 강조하기 위한 가중치를 갖고 있다. Spatial Attention은 1채널의 스케일링 특징 맵으로 입력 특징 맵 F_{in}을 스케일링하고 최종적으로 공간 정보가 강조된 특징 맵 F_{out}을 출력한다. 이를 수식으로 표현



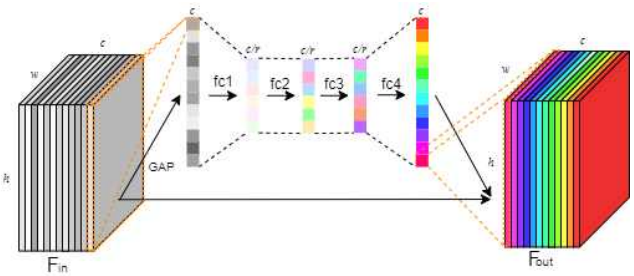
$$F_{out} = \sigma_2(fc_2(\sigma_1(fc_1(F_{in}))))F_{in}$$

(a) One layer Channel Attention



$$F_{out} = \sigma_3(fc_3(\sigma_2(fc_2(\sigma_1(fc_1(F_{in}))))))F_{in}$$

(b) Two layer Channel Attention



$$F_{out} = \sigma_4(fc_4(\sigma_3(fc_3(\sigma_2(fc_2(\sigma_1(fc_1(F_{in}))))))))F_{in}$$

(c) Three layer Channel Attention

그림 3. 제안하는 다양한 Channel Attention

Fig. 3. Proposed various Channel Attention

하면 다음과 같다.

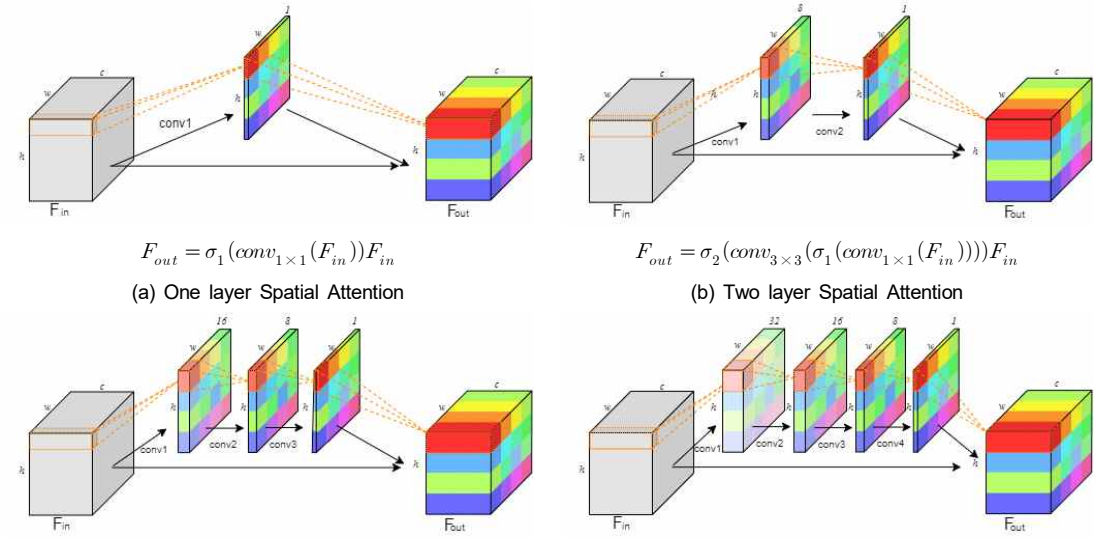
$$F_{out} = SA(F_{in}) \tag{6}$$

본 논문에서 제안하는 다양한 Spatial Attention의 구조는 그림 4와 같다.

제안하는 Spatial Attention은 스케일링 특징 맵을 만들기 위해 적용한 Convolution Layer의 개수와 커널의 크기에서 차이를 가진다. One layer Spatial Attention은 1x1 커널을 갖는 한 개의 Convolution Layer를 이용해서 채널의 정보를 취합하는 방식으로 공간 정보를 강조한다. Two layer Spatial Attention부터 Four layer Spatial Attention은 Convolution Layer를 추가로 사용해서 인접한 픽셀 정보 및 특징을 활용한다. Two layer Spatial Attention은 3x3의 커널을 갖는 Convolution Layer를 이용해서 인접한 9개의 픽셀을 추가로 고려하며, Three layer Spatial Attention과 Four layer Spatial Attention은 각각 49개, 169개의 주변 픽셀을 고려하여 강조한다. 또한 유용한 정보를 보존 및 압축하기 위해 특징 맵의 채널을 크기를 점진적으로 줄여서 강조된 스케일링 특징 맵을 만든다. 스케일링 특징 맵이 0부터 1사이의 값을 가지도록 만들기 위해 마지막 활성화 함수는 Sigmoid를 사용하였으며 이를 제외한 나머지 활성화 함수는 LeakyReLU이다.

• Parallel and Serial Attention

특징 맵의 채널과 공간 정보를 한 번에 강조하기 위해 Channel Attention과 Spatial Attention을 직렬 혹은 병렬로 배치하는 방법을 설계하였다. 첫 번째 방법은 Attention Module을 직렬로 배치하여 채널 혹은 공간 정보가 강조된 특징 맵을 다른 Attention Module로 다시 한 번 강조하는 것이며 배치하는 순서에 따라 2개의 배치 방법을 적용할 수 있다. 두 번째 방법은 Attention Module을 병렬로 배치하고 강조된 2개의 특징 맵을 합치는 것으로 합치는 방식에 따라 2개의 병렬 배치 방법이 있다. 직렬 배치는 Spatial Attention을 먼저 적용하고 뒤에 Channel Attention을 적용하는 방식과 반대로 적용하는 방식이 있으며, 병렬 방식은 강조된 특징 맵을 단순히 더하는 방식과 두 요소 중 큰 값을 사용하는 Max방식이 있다. 다음은 제안하는 직렬 혹은 병렬 배치 방법에 따른 Attention의 구조와 이를 수식으로 나타낸 것이다.



$F_{out} = \sigma_1(conv_{1\times1}(F_{in}))F_{in}$

(a) One layer Spatial Attention

$F_{out} = \sigma_2(conv_{3\times3}(\sigma_1(conv_{1\times1}(F_{in}))))F_{in}$

(b) Two layer Spatial Attention

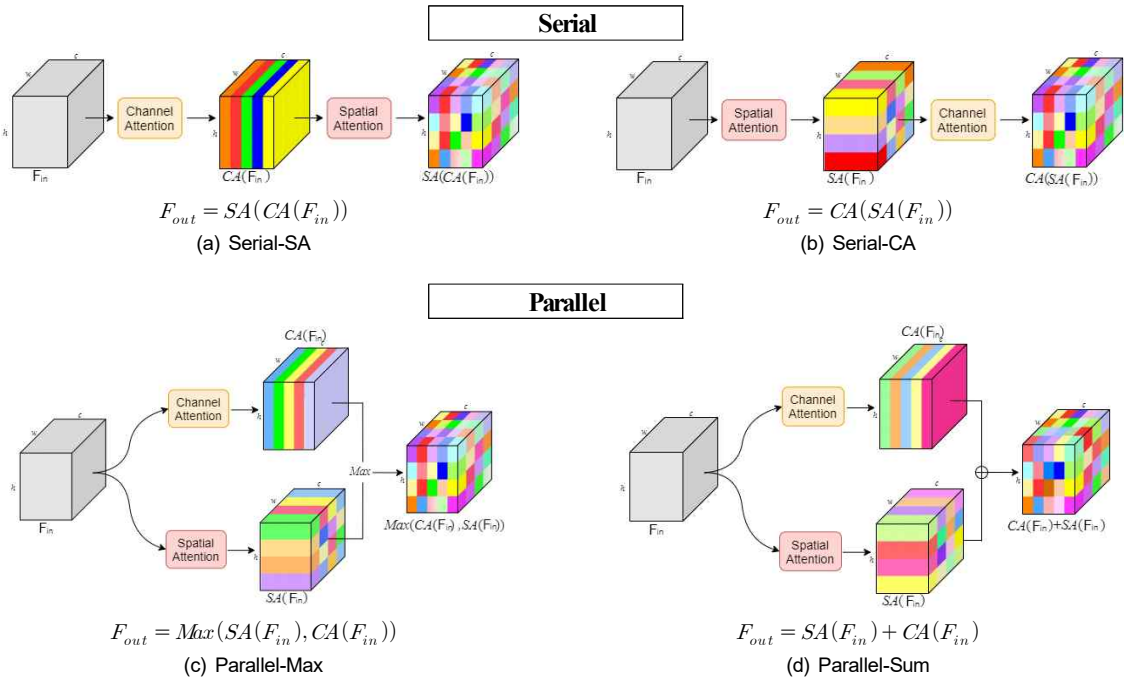
$F_{out} = \sigma_3(conv_{5\times5}(\sigma_2(conv_{3\times3}(\ldots(F_{in})\ldots))F_{in}$

(c) Three layer Spatial Attention

$F_{out} = \sigma_4(conv_{7\times7}(\sigma_3(conv_{5\times5}(\ldots(F_{in})\ldots))F_{in}$

(d) Four layer Spatial Attention

그림 4. 제안하는 다양한 Spatial Attention
Fig. 4. Proposed various Spatial Attention



$F_{out} = SA(CA(F_{in}))$
(a) Serial-SA

$F_{out} = CA(SA(F_{in}))$
(b) Serial-CA

$F_{out} = Max(SA(F_{in}), CA(F_{in}))$
(c) Parallel-Max

$F_{out} = SA(F_{in}) + CA(F_{in})$
(d) Parallel-Sum

그림 5. 제안하는 Attention 다양한 배치 방법
Fig. 5. Proposed various arrangement method for Attention module

Ⅳ. 실험 및 결과

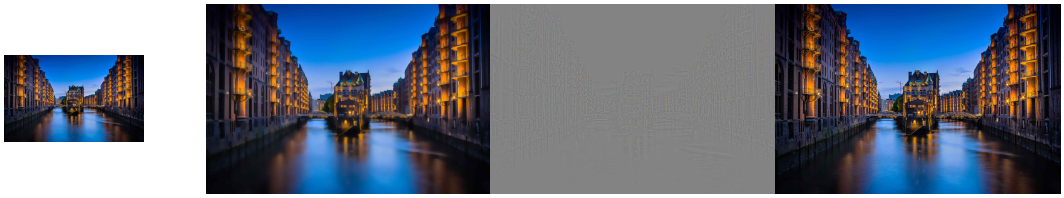
1. 실험 방법 및 세부 사항

본 논문에서는 Tensorflow를 기반으로 제안하는 초고해상도 복원 네트워크와 Attention Module를 설계하고 x2 복원 스케일에서 실험을 진행하였다. 학습 데이터 셋은 DIV2K[15]를 사용하였으며 SATV[16]를 이용해서 참조 영상 Structure S_{HR}와 Texture T_{HR}을 만들었다. 다음은 학습에 사용한 영상 예시이다.

또한 실험 결과를 검증하기 위해 Set5, Set14, BSD100, Urban100, Manga109를 검증 데이터셋으로 사용하였으며 Attention Module을 사용하지 않은 방법을 Base로 두고 비교하였다. 평가 방법은 PSNR과 SSIM을 사용하였으며 가장 높은 값은 굵게 표시하고 두 번째로 높은 값은 밑줄로 표시한다.

2. Channel Attention

본 논문에서는 Convolution Layer의 개수에 따른 3개의 Channel Attention을 설계하고 실험하였다. 다음 표 3은 제안하는 Channel Attention의 결과이다.



(a) Input LR Image (b) Reference Structure S_{HR} (c) Reference Texture T_{HR} (d) Ground Truth I_{HR}

그림 6. 학습 영상 예시
Fig. 6. Training image examples

제안하는 Channel Attention은 Set14, BSD100, Urban-100에서 Base보다 낮거나 비슷한 PSNR을 보였으며 Set5와 Manga109에서만 Base보다 높은 PSNR을 얻었으며 Set14와 BSD100에서 Base에 비해 높은 SSIM을 얻었다. Channel Attention이 각 데이터셋의 특징과 도메인을 고려하지 못한 채 채널을 강조해서 복원 성능이 일관적이지 않고 경우에 따라 오히려 성능을 저하시킨 것으로 보인다. 한편, One layer CA는 Manga109에서만 Base보다 높은 PNSR와 SSIM을 보였으나 Convolution Layer를 추가한 Two layer CA와 Three layer CA는 Set5, BSD100, Urban109에서도 Base보다 높거나 비슷한 PSNR을 얻었으며 SSIM의 경우 Set5, Set14, Manga109에서 원본보다 높거나 비슷한 결과를 보였다. 이는 Channel Attention의 Fully Connected Layer를 추가함으로써 데이터셋 도메인에 대한 표현 능력과 일반화 성능이 비교적 향상된 것으로 생각된다.

3. Spatial Attention

다음 표 4는 Spatial Attention의 실험 결과이다. Spatial Attention은 Convolution Layer 의 갯수와 채널의 크기에 따라 4개의 Spatial Attention을 설계하고 실험하였다.

표 4에서 One layer SA의 경우 모든 검증 데이터셋에 대해서 Base보다 복원 성능이 낮은 것을 알 수 있다. One layer SA는 1x1 커널을 갖는 하나의 Convolution Layer를 사용해서 인접한 픽셀들에 대한 고려 없이 채널별 정보만을 취합하기 때문에 적절한 공간 정보를 강조하지 못하였고 결과적으로 복원 성능을 감소시켰다. Convolution Layer를 추가한 Two layer SA와 Three layer SA의 경우 Convolution Layer를 추가할수록 점진적으로 복원 성능이 높아지는 것을 확인 할 수 있다. 이는 보다 많은 Convolution Layer와 3x3, 5x5 커널의 넓은 수용 영역으로 인접한 픽셀을 고려하고 점진적으로 채널을 줄여 정보를 압축하는 것이 보다 정확한 강조된 특징 맵을 만드는 데에 도움을 준다고 할 수 있다. 마지막 Four Layer SA는 더욱 넓은 수용 영역을 이용해서 모든 검증 데이터셋에 대해 Base와 같거나 높은 PSNR과 SSIM을 얻었다. 위 실험 결과를 토대로 넓은 수용 영역으로 주변의 인접한 픽셀들을 고려하는 것이 정확한 강조된 특징 맵을 만들고 초고해상도 복원 성능을 높이는 데에 도움이 된다는 것을 알 수 있다.

4. Parallel and Serial Attention

본 논문에서는 다양한 관점에서 특징 맵을 강조하기 위

표 3. 다양한 Channel Attention의 초고해상도 복원 결과
Table 3. Results of various Channel Attention on validation datasets

	Set5	Set14	BSD 100	Urban 100	Manga 109
Base	37.85/0.9597	**33.59**/0.9174	**32.16**/0.9027	**32.10/0.9292**	38.23/0.9739
One layer CA	37.84/0.9597	33.51/0.9177	32.11/0.9019	32.03/0.9275	**38.33/0.9741**
Two layer CA	37.83/0.9597	33.56/**0.9180**	**33.16/0.9028**	**32.10**/0.9290	38.22/0.9739
Three layer CA	**37.87/0.9598**	33.50/0.9177	32.12/0.9023	32.06/0.9284	**38.33**/0.**9741**

표 4. 다양한 Spatial Attention의 초고해상도 복원 결과
Table 4. Results of various Spatial Attention on validation datasets

	Set5	Set14	BSD 100	Urban 100	Manga 109
Base	37.85/0.9597	**33.59**/0.9174	**32.16**/0.9027	32.10/0.9292	38.23/0.9739
One layer SA	37.84/0.9595	33.52/0.9168	32.15/0.9027	32.05/0.9283	38.18/0.9736
Two layer SA	37.85/0.9596	33.53/0.9174	32.15/0.9027	32.05/0.9284	38.21/0.9738
Three layer SA	37.85/0.9597	33.57/0.9178	32.15/0.9027	32.08/0.9289	38.25/**0.9740**
Four layer SA	**37.88/0.9598**	33.57/**0.9180**	**32.16/0.9029**	**32.16/0.9297**	**38.27**/0.9739

표 5. 다양한 Attention Module 배치에 따른 결과
Table 5. Results of Parallel and Serial Attention module

	Set5	Set14	BSD 100	Urban 100	Manga 109
Base	37.85/0.9597	**33.59**/0.9174	**32.16**/0.9027	32.10/0.9292	38.23/0.9739
Three layer CA	37.87/0.9598	33.50/0.9177	32.12/0.9023	32.06/0.9284	38.33/0.9741
Four layer SA	37.88/0.9598	33.57/0.9180	**32.16**/**0.9029**	32.16/**0.9297**	38.27/0.9739
Serial-SA	37.86/0.9597	33.56/0.9179	32.12/0.9022	38.08/0.9283	32.29/0.9741
Serial-CA	37.90/**0.9599**	33.51/0.9179	32.13/0.9023	32.08/0.9281	38.35/0.9742
Parallel-Max	37.86/0.9598	33.54/0.9180	32.12/0.9022	32.06/0.9285	38.32/0.9741
Parallel-Sum	**37.91**/**0.9599**	33.57/**0.9184**	32.15/0.9025	**32.19**/0.9295	**38.39**/**0.9743**

해 Channel Attention과 Spatial Attention을 직렬 혹은 병렬로 배치하였다. 각 배치 방법의 Channel Attention과 Spatial Attention은 각각 Three layer CA와 Four layer SA이며 다음 표 5는 배치 방법에 따른 결과이다.

위 실험 결과에서 병렬-합(Parallel-Sum)이 Set5, Urban-100, Manga109에서 다른 배치 방법들에 비해 가장 높은 PSNR을 보였으며 BSD를 제외한 나머지 검증 데이터셋에서 가장 높은 SSIM을 얻었다. PSNR은 Set14와 BSD100에서는 Base보다 낮았지만 그 차이가 다른 방법들에 비해 적었다. Attention Module을 직렬로 배치한 방법의 경우 앞선 Attention Module에 의해 강조된 특징 맵의 정보가 뒤에 이은 Attention Module에 의해 약화될 수 있기 때문에 의도와 달리 온전치 않게 부분적으로 강조되었고 그 결과로 낮은 복원 성능을 가져온 것으로 생각된다. 병렬-최댓값(Parallel-Max)은 각각의 강조된 특징 맵은 보존하였으나, 두 강조된 특징 맵 간의 Max연산 과정에서 강조 정보가 약화된 것으로 생각된다.

5. 실험 결과 비교

본 논문에서 제안하는 방법의 성능을 검증하기 위해 기존의 방법들과 정량적 및 정성적으로 비교하였다. 다음은 저해상도 영상에서 제안하는 방법을 통해 저해상도 영상에서 복원한 구조 및 질감 요소 영상과 최종 복원 결과이다.

다음은 제안하는 다양한 Attention Module과 기존의 복원 방법들을 정량적으로 비교한 표이다.

표 6. 제안하는 방법의 정량적 평가 결과
Table 6. Quantitative comparisons of previous state-of-the-art methods

	Set5	Set14	BSD 100	Urban 100	Manga 109
Bicubic	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.51/0.8946	35.72/0.9680
VDSR	37.53/0.9587	33.05/0.9127	31.90/0.8960	30.77/0.9141	37.16/0.9740
LapSRN	37.52/0.9591	32.99/0.9124	31.80/0.8949	30.42/0.9101	37.53/0.9710
MemNet	37.78/0.9591	33.28/0.9142	32.08/0.8978	32.26/0.9195	37.72/0.9740
EDSR	38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351	39.10/0.9773
RCAN	38.27/0.9614	34.12/0.9216	32.41/0.9027	33.34/0.9384	39.44/0.9786
Base	37.85/0.9597	33.59/0.9174	32.16/0.9027	32.10/0.9292	38.23/0.9739
Three layer CA	37.87/0.9598	33.50/0.9177	32.12/0.9023	32.06/0.9284	38.33/0.9741
Four layer SA	37.88/0.9598	33.57/0.9180	32.16/0.9029	32.16/0.9297	38.27/0.9739
Serial-SA	37.86/0.9597	33.56/0.9179	32.12/0.9022	38.08/0.9283	32.29/0.9741
Serial-CA	37.90/0.9599	33.51/0.9179	32.13/0.9023	32.08/0.9281	38.35/0.9742
Parallel-Max	37.86/0.9598	33.54/0.9180	32.12/0.9022	32.06/0.9285	38.32/0.9741
Parallel-Sum	37.91/0.9599	33.57/0.9184	32.15/0.9025	32.19/0.9295	38.39/0.9743

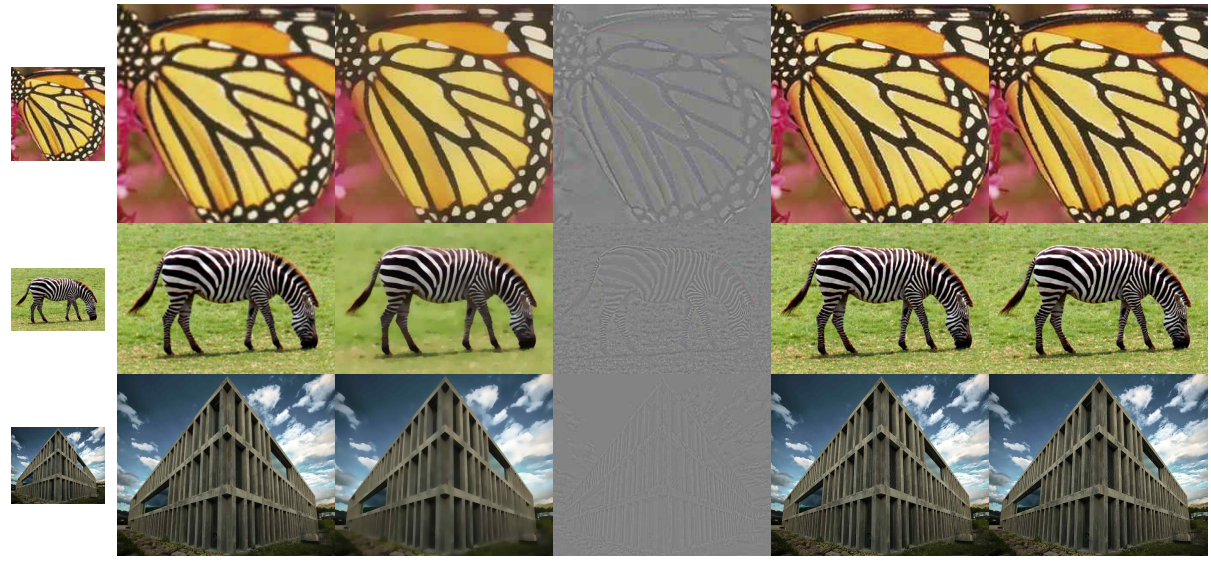


그림 7. 제안하는 방법을 통한 복원 결과
Fig. 7. Results of the proposed network

V. 결 론

본 논문에서는 초고해상도 복원을 기반으로 다양한 구조의 Attention Module과 효율적으로 배치 방법을 실험하고 성능을 평가 및 분석하였다. 그 결과 Spatial Attention과 Channel Attention에서 Layer를 추가해서 수용 영역을 넓히고 채널의 잠재된 정보를 추출하는 것이 강조된 공간 정보와 채널 정보를 얻는 데에 효과가 있고 초고해상도 복원 네트워크의 성능을 향상시키는 데에 도움이 된다는 것을 확인하였다. 또한 다중 Attention Module을 사용하여 다양한 관점에서 특징 맵을 강조하고자 할 때, 병렬로 배치하는 것이 직렬로 배치하는 것보다 효과적으로 특징을 강조할 수 있다는 것을 확인하였다.

참 고 문 헌 (References)

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

[2] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp.7132 - 7141, 2018.

[3] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp.6298-6306, 2017.

[4] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, "Image super-resolution using very deep residual channel attention networks," *Proceedings of the European Conference on Computer Vision*, Munich, Germany, pp.286-301, 2018.

[5] T. Dai, J. Cai, Y. Zhang, S. Xia, L. Zhang, "Second-order attention network for single image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 11065-11074, 2019.

[6] J. Kim, J. Kwon, K. Lee, "Accurate image super-resolution using very deep convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.1646 - 1654, 2016.

[7] J. Kim, J. Kwon Lee, K. Lee "Deeply-recursive convolutional network for image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.1637 - 1645, 2016.

[8] Y. Tai, J. Yang, X. Liu, "Image super-resolution via deep recursive residual network," P*roceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 3147-3155, 2017.

[9] B. Lim, S. Son, H. Kim, S. Nah, K. Lee, "Enhanced deep residual networks for single image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, Hawaii, USA, pp.1132 - 1140, 2017.

[10] W. Shi, J. Caballero, F. Huszar, J. Totz, A. Aitken, R. Bishop, D. Rueckert, Z. Wang, "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*, Las Vegas, NV, USA, pp.1874-1883, 2016.

[11] Y. Tai, J. Yang, X. Liu, C. Xu, "Memnet: A persistent memory network for image restoration," *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 4539-4547, 2017.

[12] T. Tong, G. Li, X. Liu, Q. Gao, "Image super-resolution using dense skip connections," *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 4799-4807, 2017.

[13] Y. Hu, J. Li, Y. Huang, X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.30, No.11, pp.3911-3927, Nov 2020.

[14] Z. Chen, V. Badrinarayanan, C. Lee, A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," *Proceedings of the 35th International Conference on Machine Learning*, PMLR, Vol.80, pp.794-803, 2018.

[15] E. Agustsson, R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, Hawaii, USA, pp.1100 - 1121, 2017.

[16] J. Song, H. Cho, J. Yoon, S. Yoon, "Structure adaptive total variation minimization-based image decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.28, No.9, pp.2164-2176, Sep 2018.

저 자 소 개

문 환 복

- 2019년 2월 : 인하공업전문대학 컴퓨터시스템공학과 공학사
- 2019년 3월 ~ 현재 : 국민대학교 컴퓨터공학과 석사과정
- ORCID : https://orcid.org/0000-0001-7360-1016
- 주관심분야 : 인공지능, 컴퓨터 비전

윤 상 민

- 2000년 : 고려대학교 전자공학과 공학사
- 2002년 : 고려대학교 전자공학과 공학석사
- 2010년 : 독일 다름슈타트공대 컴퓨터공학 공학박사
- 2002년 ~ 2005년 : 삼성종합기술원 연구원
- 2010년 ~ 2011년 : 일본 AIST 박사후 연구원
- 2011년 ~ 2012년 : 연세대학교 조교수
- 2018년 ~ 2019년 : 미국 MIT 방문연구원
- 2012년 ~ 현재 : 국민대학교 소프트웨어융합대학 부교수
- ORCID : https://orcid.org/0000-0003-0001-1845
- 주관심분야 : 컴퓨터비전, 인공지능, HCI