Dual Super Resolution Learning(DSRL) 프레임워크 개선을 통한 Semantic Segmentation 성능 향상

연구참여학생: 가상은 (인 가상은), 지도교수: 이규중 (인)
성신여자대학교 AI융합학부

1. Abstract

Semantic Segmentation의 성능을 높이는 방법에는 여러 가지가 존재한다. 예를 들어, Semantic Segmentation에 high-resolution 입력을 적용하는 방법 또는 Natural Language Processing (NLP)로부터 비롯된 Vision Transformer (ViT)[1]를 CNN 대신 적용하는 방법 등 다양하다. Dual Super-Resolution Learning (DSRL)[2]는 Semantic Segmentation 성능을 높이기 위해 Super-Resolution을 보조적으로 사용한 대표적인 프레임워크 중 하나이다. DSRL은 기존의 Semantic Segmentation 네트워크로 이뤄진 Semantic Segmentation Super-Resolution (SSSR) 브랜치에 Single Image Super-Resolution (SISR) 브랜치와 Feature Affinity (FA) 모듈을 추가했다. SISR 브랜치로부터 생성된 Super-Resolution 이미지 정보는 FA 모듈을 통해 SSSR 브랜치에 반영되어 더 향상된 Semantic Segmentation task를 수행한다. 본 논문에서는 DSRL 프레임워크의 한계점을 찾고 이를 보완 및 수정하여 Semantic Segmentation 성능을 향상시킨 개선된 DSRL 프레임워크를 제안한다. 개선된 DSRL 프레임워크는 CityScapes[3] 데이터셋에 대해 61.58% mIoU를 이룰 수 있었다. 이는 기존 DSRL 프레임워크를 사용한 결과보다 약 14.12% 높은 mIoU이다.

2. Introduction & Related Work

2-1. Dual Super-Resolution Learning

'Dual Super-Resolution Learning (DSRL)' [2]은 Semantic Segmentation task를 수행하는 간단하고 유연한 모델 중 하나로써 추가적인 computation cost와 memory overload 없이도 low-resolution 입력 이미지에서 높은 성능을 보인다. DSRL 프레임워크는 **Figure 1**에서 보는 바와 같이 세 부분인 Semantic Segmentation Super-Resolution (SSSR) 브랜치, Single Image Super-Resolution (SISR) 브랜치, 그리고 Feature Affinity (FA) 모듈로 구성된다.

Semantic Segmentation Super-Resolution (SSSR) 브랜치는 DSRL의 핵심적인 목표가 되는 Semantic Segmentation task를 수행한다. Encoder-decoder 구조를 갖는 Semantic Segmentation 모델로 구성되어 있으며, decoder가 Post-upsampling SR[4, 5, 6] 과정을 따르기 때문에 Pre-upsampling SR[6]로 인한 label 정보 손실을 피할 수 있다는 장점을 갖는다. DSRL 논문에서는 Semantic Segmentation을 수행하는 모델로 ESPNetv2[7], DeepLabv3+[8], PSPNet[9] 등을 이용했다. 본 논문에서는 DeepLabv3+[8] 모델로만 실험을 진행하였다.

Single Image Super-Resolution (SISR) 브랜치는 low-resolution 입력 이미지를 high-resolution 이미지로 만드는 Super Resolution task를 수행한다. 이렇게 만들어진 high-resol

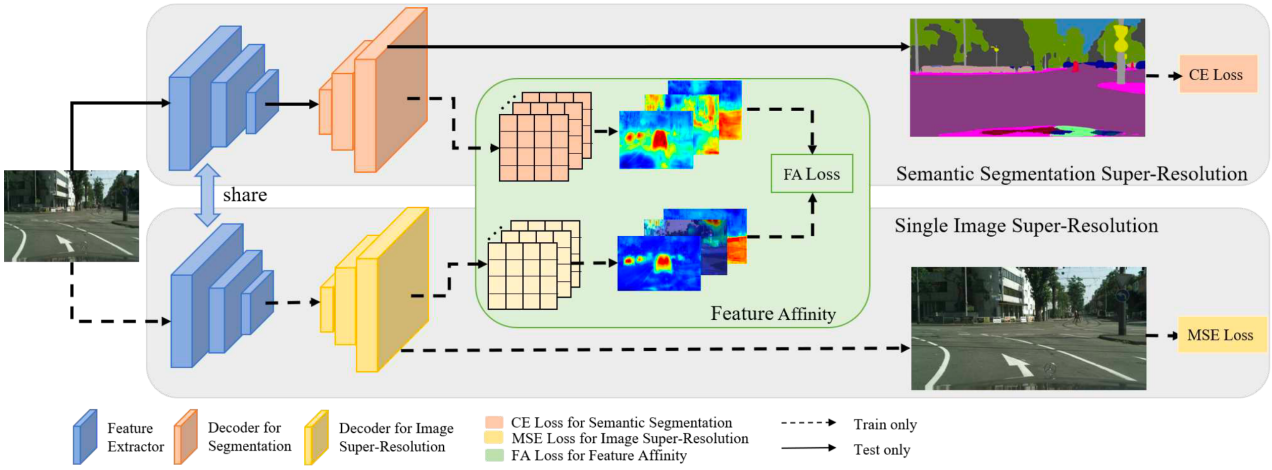

Figure 3. DSRL 프레임워크의 전체적인 구조를 나타낸 그림이다. 해당 그림은 'Dual Super-Resolution Learning for Semantic Segmentation' 논문[2]에서 발췌하였다.

ution 이미지는 Semantic Segmentation 수행에 도움을 준다. SSSR 브랜치에서는 보기 힘들고 놓칠 수 있는 이미지의 세밀하고 자세한 정보를 SISR로 생성된 high-resolution feature를 통해 얻을 수 있기 때문이다. 이처럼 SISR과 SSSR 브랜치는 계속해서 상호작용을 해가며 SSSR 브랜치가 Semantic Segmentation을 더 잘 수행할 수 있도록 보조한다. SISR 브랜치 역시 encoder-decoder 구조를 갖는데, SISR의 encoder는 SSSR의 encoder와 동일하여 함께 encoder를 공유하기 때문에 이 부분에서 연산량을 많이 감소된다.

Feature Affinity (FA) 모듈은 SISR로부터 얻은 high-resolution feature의 세부적인 표현 정보를 SSSR로 잘 전달할 수 있도록 반영해 주는 역할을 수행한다. SSSR과 SISR decoder의 출력 feature map을 대상으로 simiarity matrix을 만들어 픽셀과 픽셀 간의 거리를 학습한다. 그러나, 이 모듈에는 몇 가지 이해되지 않는 점들이 있다. 첫 번째는 FA 모듈의 대상이 되는 feature 선정에 대한 적절한 이유나 근거를 찾아볼 수 없는 것이다. 두 번째는 전체 픽셀 중 1/8 개만을 subsampling 하여 sampling 된 픽셀 간의 관계만 살펴보는 것이다. 이는 feature 전체의 특징을 반영하지 않는다는 한계가 명백히 존재한다. 이러한 이유들 때문에 본 논문에서는 FA 모듈을 Attention mechanism을 이용한 다른 모듈로 바꿈으로써 새롭게 개선된 DSRL 프레임워크를 제안한다. Attention mechanism에 대한 설명은 바로 아래 **Sec. 2-2**에 이어진다.

2-2. Self-Attention & Cross-Attention

2017년 NLP 분야에서 제안된 Self-Attention 기반의 Transformer[10]를 시작으로 Computer Vision에서도 이를 적용한 ViT[1] 모델이 등장하여 image classification, image segmentation 등의 여러 task에 쓰이고 있다. Self-Attention은 어떠한 대상의 각 부분에 대한 연관성을 파악하기 위한 Transformer의 mechanism 중 하나이다. 예를 들어, 문장에서 각 단어들이 문맥상으로 어느 정도 연관이 있는지 그리고 이들이 얼마나 상호작용하는지를 수치로 표현하는 방법이라고 할 수 있다. Self-Attention은 Query, Key, Value를 통해 계산된다. Query는 연관성을 파악할 때 기준이 되는 정보이며, key와 value는 query 정보의 각

부분이 갖는 의미나 특징을 나타낸다. 이때 key와 value는 같은 값이지만, 서로 수행하는 역할이 다르다. 쉽게 말해 입력 문장에서 각 단어 간의 연관성 파악을 위해 Self-Attention 계산을 수행한다고 할 때, query는 "이 단어와 관련된 정보는 무엇입니까?"라는 질문에 해당한다면 key는 각 단어가 가진 의미나 특징을 나타내는 정보이다. Query에 해당하는 질문을 풀기 위해서는 query와 일치하는 key를 찾아야 한다. 이를 위해 query와 key 간의 내적을 계산하여 서로 간의 연관성을 알아낸다. 이 값을 'Attention Score'이라고 한다. 이 score가 높을수록 해당 단어 간의 연관성이 높아 문장을 이해하는 데에 큰 영향을 미친다고 할 수 있다. 그다음, Attention Score를 0과 1 사이의 확률 분포로 변환하기 위해 Softmax 함수를 적용한다. 이는 각 단어에 대한 중요도를 확률로 재표현한 것이다. 이 값을 'Attention Weight'라고 한다. 마지막으로 Attention Weight와 남은 value를 곱하여 Self-Attention의 출력을 계산한다. 이 최종 출력은 문장에서 각 단어 간의 맥락적 연관성과 중요도를 나타낸다.

Cross-Attention[11]은 Self-Attention과 비슷한 mechanism을 지니지만, query, key, value를 만들어내는 값이 두 개의 입력으로부터 비롯될 때 쓰인다. Self-Attention이 하나의 입력으로부터 query, key, value를 생성하였다면, Cross-Attention은 두 개의 입력을 받아 하나의 입력으로부터는 query를, 그리고 나머지 하나의 입력으로부터는 key와 value를 생성한다. Cross-Attention 계산은 Self-Attention 계산 방법과 동일한 과정으로 진행된다. Cross-Attention은 주로 Multimodal과 같은 서로 다른 두 개의 task 간의 상관관계를 모델링 할 때 사용된다. 본 프레임워크에서는 서로 다른 역할을 수행하는 SISR과 SSSR 브랜치 간의 상관관계를 알아내야 하므로 Self-Attention이 아닌 Cross-Attention mechanism을 이용해 DSRL 프레임워크[2] 개선하였다. DSRL 프레임워크를 구체적으로 어떠한 방식으로 개선했는지는 **Sec. 3**을 통해 확인할 수 있다.

2-3. Interpolation

DSRL 논문에서는 up-sampling과 down-sampling 시 bilinear interpolation만을 사용하였다. Bilinear interpolation은 linear interpolation을 두 번 적용하여 2차원 데이터를 보간하는 방법이다. 보간하고자 하는 곳과 인접한 네 개의 지점을 대상으로 x축 방향으로 두 지점의 거리비를 이용하여 linear interpolation을 통해 새로운 두 개의 값을 구한다. 그 다음, 이 두 값을 y축 방향으로 linear interpolation 하여 원하는 지점에 대한 최종 보간 값을 얻는다.

Bicubic interpolation[12]은 cubic interpolation[12]을 두 번 적용하여 2차원 데이터를 보간하는 방법이다. Bilinear interpolation보다 더 많은 주변 데이터를 바탕으로 3차원 함수를 이용하여 보간을 한다. 인접한 16개의 지점을 대상으로 x축 방향으로 cubic interpolation[12]을 수행해서 나온 값을 y축 방향으로 cubic interpolation[12] 하여 원하는 지점에 대한 보간 값을 얻는다.

DSRL 논문에는 bilinear interpolation 사용에 대한 타당한 이유가 드러나 있지 않다. 대게 bicubic interpolation[12]이 bilinear interpolation보다 더 정밀한 보간을 수행함에도 말이다. 그리하여 본 논문에서는 추가적으로 개선된 DSRL 프레임워크[2]의 up-sampling과 down-sampling 단계에 bicubic interpolation[12]을 적용했을 때의 결과가 bilinear interpolation를 적용했을 때보다 mIoU 성능이 높아질 거라고 예상한다.

본 연구에서 살펴볼 것들은 다음과 같다.

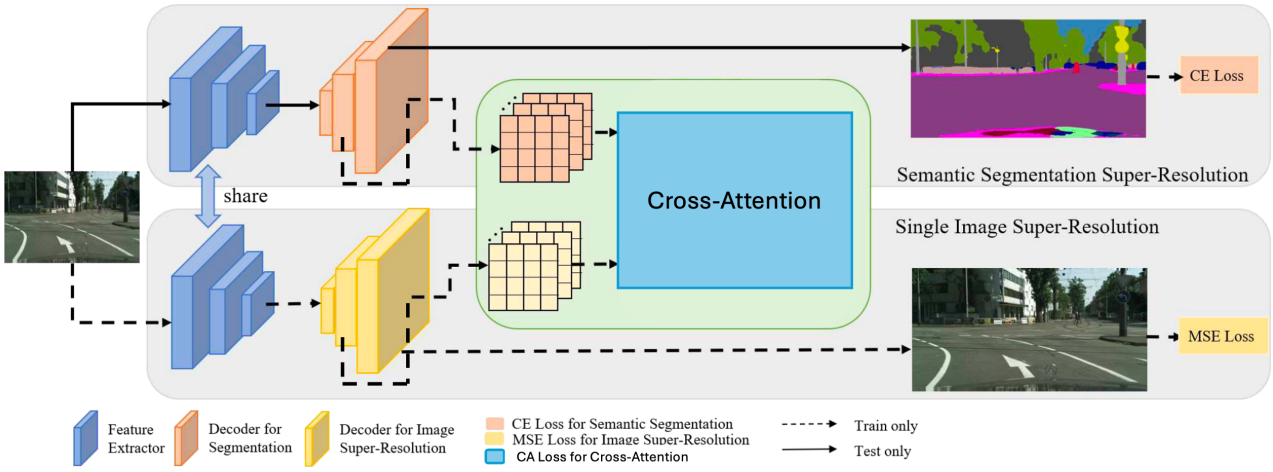

Figure 4. Advanced DSRL 프레임워크의 전체적인 구조를 나타낸 그림이다.

(1) Cross-Attention mechanism을 사용한 개선된 DSRL 프레임워크는 기존 DSRL 프레임워크[2]보다 Semantic Segmentation task를 더 잘 수행한다고 가정하고, Semantic Segmentation에서의 Attention mechanism 적용에 대한 타당성 및 효과를 다시 한번 입증한다.

(2) 개선된 DSRL 프레임워크에 bicubic interpolation[12]을 적용했을 때의 성능이 bilinear interpolation을 적용했을 때보다 더 높을 것이라고 가정하고, 각각 성능 비교를 통해 어떠한 interpolation을 적용하는 것이 더 적절한지 추가적으로 알아본다.

3. Advanced DSRL Framework

개선된 DSRL 프레임워크와 기존 DSRL 프레임워크[2]를 비교했을 때, 가장 크게 바뀐 곳은 FA 모듈이다. 본 논문은 이 FA 모듈에는 모호함과 한계점이 존재한다고 지적한다. 첫 번째로 FA 모듈에서 사용하는 feature들의 기준이 모호하다는 점이다. SISR과 SSSR의 마지막 브랜치에서 나온 feature를 기준으로 삼는데 이에 대한 이유나 설명은 나와있지 않다. 두 번째로 FA 모듈에서는 이 feature들이 갖고 있는 정보를 모두 사용하지 않고, 높은 memory overhead를 이유로 전체 픽셀 중 1/8개의 픽셀만 subsampling 하여 해당 픽셀 간의 정보만을 사용한다. 여기에는 픽셀을 sampling 하는 기준이 없을뿐더러 어떤 픽셀이 더 중요한 정보를 담고 있는지 혹은 어떤 픽셀이 더 필요한 픽셀인지에 대한 디테일한 요소들은 전혀 고려하지 않고 있다. 본 논문에서는 이러한 FA 모듈의 불명확성과 한계를 지적하며 FA 모듈 대신 Cross-Attention 모듈을 통해서도 SISR과 SSSR 간의 정보 공유를 더 효과적으로 진행할 수 있다고 생각하였으며, 실제로도 그렇게 될 것이라고 가정한다.

개선된 DSRL 프레임워크[2]는 **Figure 2**와 같이 동작한다. FA 모듈 대신 Cross-Attention 모듈이 사용되었다. Cross-Attention 모듈의 입력이 되는 feature는 FA 모듈에서 사용한 feature와는 다르다. FA 모듈에서는 SISR과 SSSR dcoder의 마지막 feature map을 각각 채널이 3인 feature map으로 변환하여 FA 모듈의 입력으로 사용하였다. 반면, Cross-Attention 모듈에서는 SISR과 SSSR decoder에서 마지막으로 실행되는 up-sampling layer의 이전 feature map을 입력으로 사용한다. 이 입력 feature은 Cross-Attention 계산을

수행하는 데에 사용된다. Query는 SSSR의 feature map으로, key와 value는 SISR의 feature map으로 지정하였다. 궁극적으로는 Semantic Segmentation task를 수행하는 것이 이 프레임워크의 목적이기 때문에 SSSR의 feature map을 기준이 되는 정보로 삼아 이것을 query로 지정하였다. 이렇게 지정된 query, key, value를 이용하여 Cross-Attention 계산이 수행된다. 이 과정을 통해 SISR feature의 high-resolution인 세밀하고 정교한 이미지 정보와 semantic 정보가 담긴 SSSR feature가 결합됨으로써 SSSR 브랜치에 SISR 브랜치의 정보가 반영되고, 두 브랜치 간의 연관성이 표현된 feature가 만들어진다. 나아가, 해당 feature는 Cross-Attention 모듈의 입력 중 하나인 SSSR feature와 concatenate 되어 Cross-Attention의 출력이 반영된 새로운 SSSR feature를 생성한다. 이 새로운 SSSR feature는 SSSR 브랜치의 decoder로부터 생성된 feature와는 다르며, SISR의 정보가 담긴 feature라고 할 수 있다.

FA 모듈이 Cross-Attention 모듈로 변경됨에 따라 FA loss function 역시 바뀌었다. Cross-Attention 모듈이 Semantic Segmentation을 얼마나 잘 수행하는지 알아보기 위해 Semantic Segmentation target mask의 분포와 새로 생성된 SSSR feature 분포 간의 차이를 학습하는 Cross-Attention loss function을 추가하였다. Cross-Attention loss function은 **Equation 1**과 같다.

$$L_{ca} = CrossEntropyLoss(pred_{ca} - target_{seg})$$

Equation 1. Cross-Attention loss function

4. Optimization

전체 loss function은 **Equation 2**와 같다. SSSR 브랜치에서의 Semantic Segmentation 수행을 위한 multi-class Cross-Entropy loss function, SISR 브랜치에서의 Mean Squared Error (MSE) loss function 그리고 Cross-Attention 모듈에서 새로 생성된 SSSR feature 분포와 Semantic Segmentation mask 간의 Cross-Entropy loss function으로 구성된다. SSSR 브랜치를 대상으로 하는 loss function과 SISR 브랜치를 대상으로 하는 loss function은 각각 **Equation 3**과 **Equation 4**와 같다.

$$L = L_{ce} + w_1 L_{mse} + w_2 L_{ca}$$

Equation 2. 개선된 DSRL 프레임워크의 전체 loss function이다.

$$L_{ce} = \frac{1}{N}\sum_{i=1}^{N} -y_i log(p_i)$$

Equation 3. SSSR 브랜치에서의 multi-class Cross-Entropy loss 수식[2]이다.

$$L_{mse} = \frac{1}{N}\sum_{i=1}^{N} ||SISR(X_i) - Y_i||^2$$

Equation 4. SISR 브랜치에서의 Mean Squared Error (MSE) loss 수식[2]이다.

w_1과 w_2는 loss 값의 범위를 비슷하게 조정해 주는 가중치로 기존 DSRL 논문과 같이 각각 0.1과 1.0으로 설정하였다.

5. Experiments

5-1. Datasets

학습에 사용된 데이터셋은 CityScapes[3] 데이터셋이다. CityScapes[3] 데이터셋은 도시 길거리 풍경을 시각화 한 이미지로 image와 mask를 모두 제공하기 때문에 주로 Segmentation task 수행에 많이 쓰인다. 총 19 개의

class로 이루어져 있으며 2,975 개의 training, 500 개의 validation, 그리고 1525 개의 test 이미지로 구성되어 있다. 모든 이미지들은 1024 x 2048 resolution를 지닌다.

5-2. Implementation Details

본 실험에서는 backbone 네트워크로 ResNet101[13]을 사용하는 DeepLabv3+[8]를 Segmentation 네트워크로 사용하였다. DeepLabv3+[8]는 encoder-decoder 구조로 이뤄져 있으며, CityScapes[3] 데이터셋에 대해 state-of-the-art (SOTA) 성능을 달성했다. Advanced DSRL 프레임워크에서의 모든 Segmentation 네트워크는 Momentum이 더해진 mini-batch stochastic gradient descent (SGD)를 통해 학습된다. Momentum 0.9, weight decay는 1e-5로 설정했으며, 초기 learning rate는 0.001로 학습을 진행하다가 learning rate schedular는 power를 0.9로 설정된 poly learning rate strategy를 적용해 learning rate를 점차 줄여갔다. Train 시 Epoch는 200, batch size는 4로 학습을 진행하였다. Validation과 Test 시에도 역시 위와 동일한 조건에서 평가를 진행하였다. 또한, CityScapes[3] 데이터셋에서의 semantic segmentation 성능을 더 높이고 다양한 데이터를 만들어내기 위해 train set에 대해 세 가지 Data Augmentation 기법을 적용시켰다. RandomHorizontalFlip, RandomScaleCrop, 그리고 RandomGaussianBlur를 사용했다.

5-3. Evaluation & Analysis

본 실험에서 입력 이미지는 1024 x 2048 resolution을 갖는 원본 이미지에 down-sampling을 한 256 x 512 resolution에 해당한다. 출력 이미지는 512 x 1024 resolution이다. 이를 바탕으로 기존 DSRL 프레임워크[2]로 학습한 결과와 Advanced DSRL 프레임워크로 학습한 결과는 **Figure 3** 표를 통해 확인할 수 있다. 동일한 resolution을 갖는 입력 이미지에서 기존 DSRL 프레임워크로 학습했을 때는 47.46%, Advanced DSRL 프레임워크로 학습했을 때는 61.58% mIoU를 갖는다. 본 논문에서 제안한 Advanced DSRL 프레임워크로 학습했을 때, 기존 대비 14.12% 증가했다는 것을 알 수 있다. 이로써 본 논문은 Cross-Attention 모듈을 사용한 DSRL 프레임워크의 Semantic Segmentation task 수행 능력이 더 뛰어나다는 것을 검증하였으며, Semantic Segmentation에 적용된 Attention mechanism의 효과가 뛰어나다는 것을 다시 한번 증명하였다.

Methods	mIoU(%)
기존 DSRL[2]	47.46
Advanced DSRL	61.58

Figure 3. 256 x 512 resolution에서 기존 DSRL 프레임워크로 학습한 결과와 개선된 Advanced DSRL 프레임워크로 학습한 결과를 비교한 표이다.

추가적으로 bilinear interpolation과 bicubic interpolation[12] 중 개선된 DSRL 프레임워크에 어떠한 interpolation 방법을 적용했을 때 더 좋은 결과를 보이는지 실험한 결과는 **Figure 4**와 같다. 동일 조건에서 bilinear interpolation를 사용했을 때의 mIoU는 61.78%, bicubic interpolation[12]을 사용했을 때의 mIoU 값은 61.58%가 나왔다. 미세한 차이이지만 bilinear interpolation을 적용한 결과가 bicubic interpolation[12]을 적용한 결과보다 mIoU 값이 약 0.2% 더 높다는 것을 알 수 있다. 이에 따라 DSRL 프레임워크[2]에서는 bicubic interpolation[12]보다 bilinear interpolation 방법론이 이미지 화소를 좀 더 잘 보간한다고 결론지을 수 있다.

Methods	mIoU(%)
Bilinear Interpolation	61.78
Bicubic Interpolation[12]	61.58

Figure 4. 입력 resolution이 256 x 512 일 때, 개선된 Advanced DSRL 프레임워크를 통해 interpolation 방법을 다르게 학습시킨 결과를 비교한 표이다.

6. Conclusion

본 논문은 Attention mechanism이 적용된 Cross-Attention 모듈을 사용함으로써 Advanced Dual Super-Resolution Learning 프레임워크를 제안한다. Cross-Attention 모듈은 Single Image Super-Resolution과 Semantic Segmentation Super-Resolution feature의 일부 픽셀만이 아닌 전체 픽셀을 사용하기 때문에 Semantic Segmentation 성능이 기존 Dual Super-Resolution Learning 프레임워크 대비 14.12% 증가하였다. 이를 통해 본 연구는 Dual Super-Resolution 프레임워크에 Feature Affinity 모듈보다 Cross-Attention 모듈을 적용하는 것이 Semantic Segmentation task를 수행하는 데에 더욱 효과적임을 입증할 수 있었다. 반면, Advanced Dual Super-Resolution Learning 프레임워크에 적용되는 interpolation별 성능을 비교한 실험의 경우, bilinear interpolation을 적용한 결과값이 0.2% 더 높은 예상외의 결과를 보였다. 두 값의 차이가 크지는 않지만, DSRL 프레임워크에는 bicubic interpolation보다 bilinear interpolation을 적용하는 것이 조금 더 좋다는 결과를 얻었다. 후속 연구에서는 여러 resolution을 갖는 입력 이미지에 대한 성능 비교 실험과 또 다른 데이터 증강 기법인 FixedResize 적용에 따른 성능 비교를 진행할 것이며, 기존 Dual Super-Resolution 프레임워크와 Advanced Dual Super-Resolution 프레임워크를 사용했을 때의 FLOPs 변화도 각각 구해서 비교해 볼 예정이다. 또한, bilinear interpolation을 적용했을 때의 결과가 bicubic interpolation을 적용했을 때보다 왜 높은 성능을 보였는지 그 이유를 알아내고자 한다.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

[2] LI Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual Super-Resolution Learning for Semantic Segmentation. In *CVPR*, 2020.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. Cityscapes Dataset. https://www.cityscapes-dataset.com.

[4] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishoop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[5] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. In *TPAMI*, 2018

[6] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In ICCV, 2017.

[7] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *CVPR*, 2019.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandrou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[9] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[11] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *ICCV*, 2021.

[12] Robert G.Keys. Cubic convolution interpolation for digital image processing. In *IEEE*, 1981.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.