데이터 불균형 문제 해결을 위한 생성적 적대 신경망 기반 데이터 증강

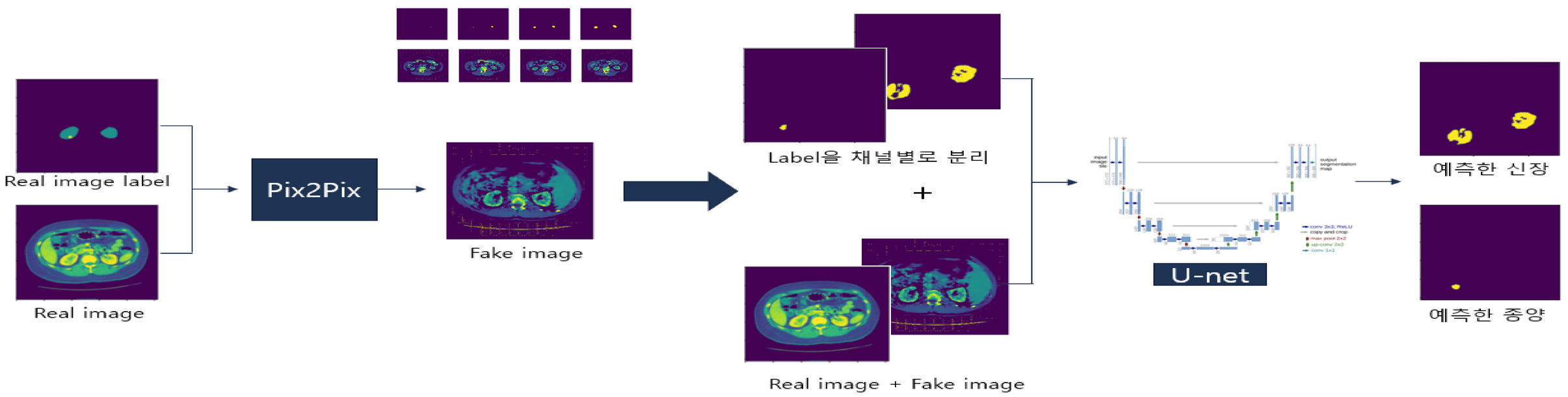
최재홍, 이승리, 서영재, 서원진, 허종욱

한림대학교 소프트웨어학부

배 경

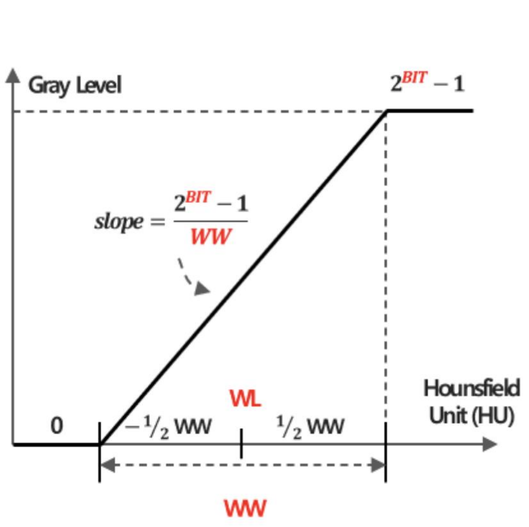
딥러닝 기술 활용에 있어 가장 필수적인 해결 과제는 학습에 필요한 데이터를 충분히 확보하는 것이다. 의료 데이터를 확보하는 과정에서 개인정보 보호 문제나 발병률이 높지 않은 경우에는 충분한 데이터 확보가 어렵다. 딥러닝 학습에서 이미지 데이터가 부족한 경우 원본 데이터를 여러 기하학적인 방법을 통해 데이터를 증가하는 방법이 활용되고 있다. 하지만 이 방법으로 추가적인 정보를 얻기에는 한계가 있다. 이를 바탕으로 데이터 증가에 대한 필요성을 느끼고 Generative adversarial networks(GAN)을 통해 가상의 데이터를 생성하여 데이터 양을 증가시켜 신장과 신장암에 대한 segmentation을 진행하여 궁극적으로 기계학습 성능 향상을 하고자 하는 목적이다.

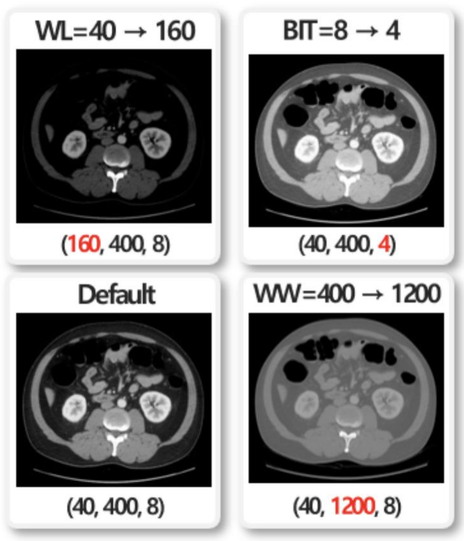
방 법



1) 일반 데이터와 기본적인 augmentation을 적용하여 학습을 진행한다.
2) Pix2Pix를 통해 생성된 fake image를 train dataset에 추가한다.
3) 새로 구성한 데이터셋으로 Segmentation 작업을 수행하며, 정확한 비교를 위해 하이퍼 파라미터의 값은 고정하여 학습을 진행한다.

데이터 전처리

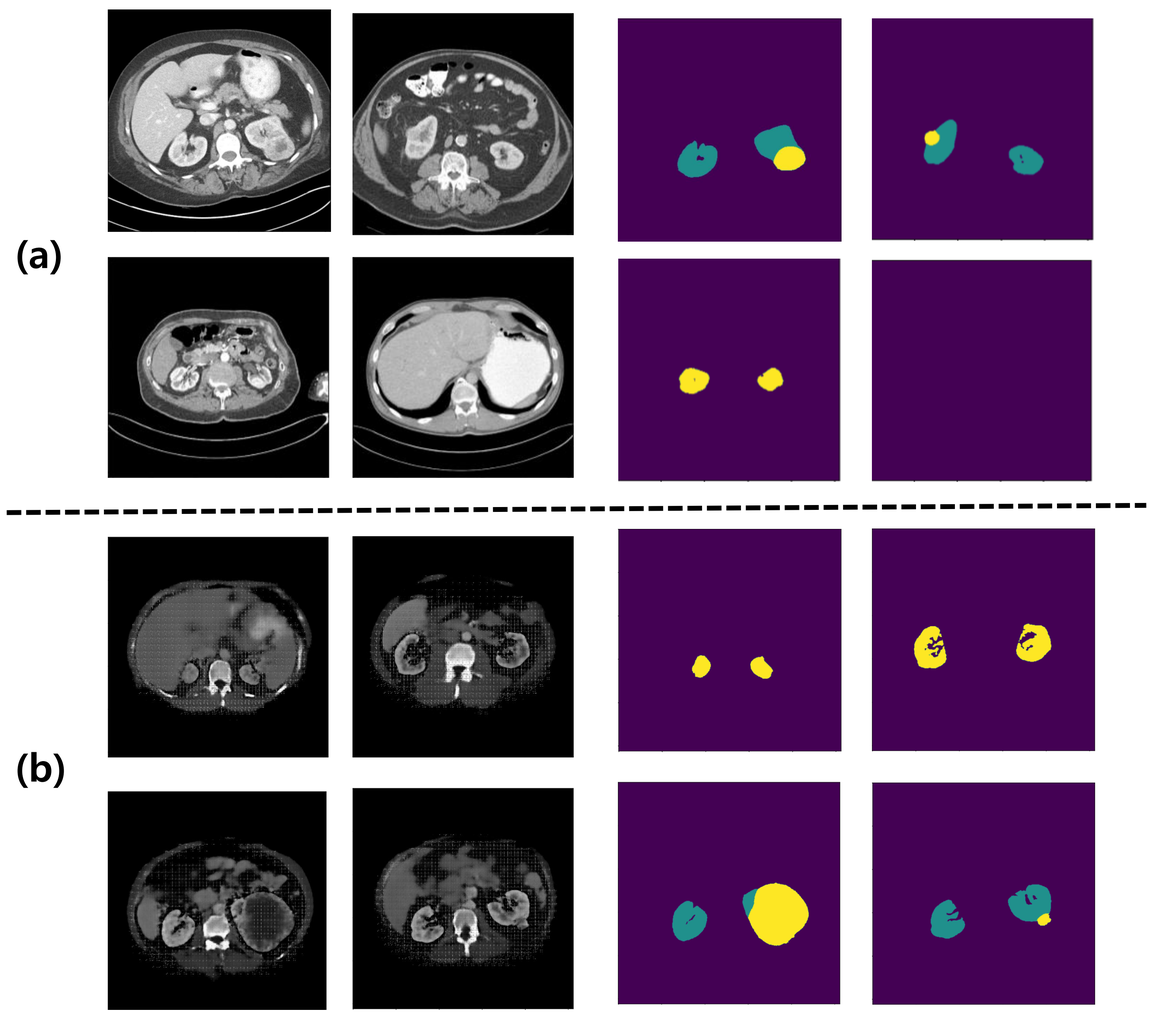




Configurations are expressed in (WL, WW, BIT)

- CT이미지는 신체부위에 따라 픽셀값의 범위(window width)와 기준이 되는 픽셀값(window center)을 조절해 이미지를 사용한다.
- 실험적으로 학습이 가장 잘되는 픽셀값의 범위는 400, 기준이 되는 픽셀값은 0으로 고정했다.

데이터



(a) 실제 환자 데이터 : 6400장
(b) 합성 이미지 : 1333장

결 과

Case	U-Net	U-Net2	FPN
기존 Dataset	58.51	89.61	88.98
기하학적인 변형으로 인한 데이터증강을 했을 때	85.70	88.93	91.28
Train Dataset에 Fake Data 추가	88.70	87.50	89.74
기하학적인 변형으로 인한 증강도 했을 때	**90.73**	**90.76**	**91.41**

결론 및 고찰

합성 데이터는 본래의 데이터 수가 제한된 상황에서 데이터를 늘리는 데 효과적이다. 이를 통해 다양한 데이터 분야에서 개인정보에 제한 없이 데이터를 확보하는 것이 가능해진다. 본 기술이 의료 분야에만 국한되어 적용하는 것이 아닌 상대적으로 데이터가 부족한 산업까지 확장이 가능하다.