딥러닝 기반의
스테가노그래피 / 스테그아날리시스

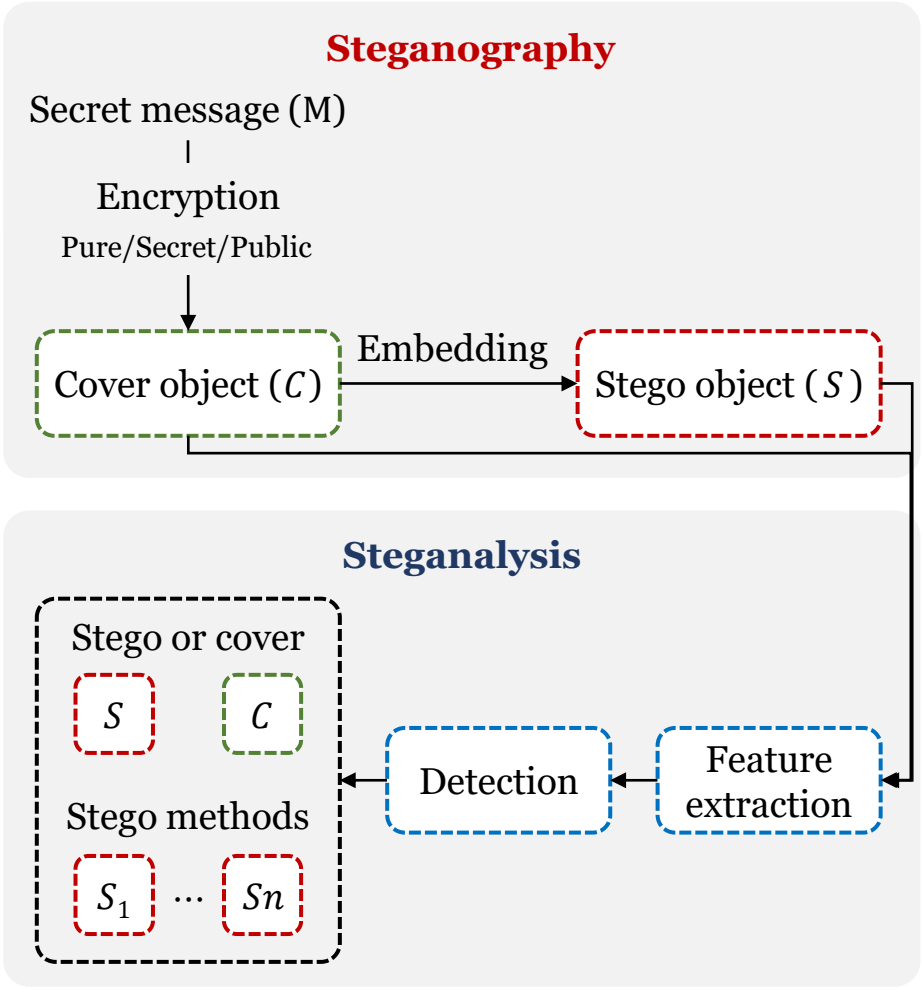
https://youtu.be/1e_etw8FgSY

한성대학교 HANSUNG UNIVERSITY

CryptoCraft LAB
https://crypto.modoo.at

스테가노그래피와 스테그아날리시스

- 중요 정보를 숨기기 위해 텍스트, 이미지, 동영상 등의 미디어 파일에 데이터를 삽입하는 기술
- 스테가노그래피를 통해 스테고 오브젝트를 생성할 수 있음
- 스테고 오브젝트는 스테그아날리시스를 통해 탐지될 수 있음
- 임베딩 방식을 구별할 수도 있으나, 어려운 작업



스테가노그래피

- 스테가노그래피는 크게 **데이터 암호화와 임베딩 과정으로 나뉨**
- **데이터 암호화**에는 다음과 같이 3가지 종류가 있음
  - **일반 :** 암호화 X → 보안성이 임베딩 알고리즘에만 의존하므로 한계 존재
  - **비밀키 및 공개키 :** 암호화 수행 → **보안성 확보**, 하지만 암호화를 거치므로 숨기고자 하는 **데이터의 크기가 증가**할 수 있음
    숨길 수 있는 **메시지의 크기가 감소**, 커버 오브젝트의 용량이 증가하여 비밀 데이터의 **존재 여부가 발각되기 쉬움**
- 이러한 단점을 커버하기 위해 스트림 암호화를 적용한 사례 존재

- **임베딩 과정** : 커버 오브젝트에 비밀 정보를 은닉하는 과정
  - 커버 데이터의 왜곡과 손상을 최소화하여 비밀 데이터가 숨겨져있다는 사실을 숨겨야 함
  - 스테가노그래피의 성능은 다음과 같은 관점에서 평가
    - **임베딩 용량** (숨길 수 있는 비트의 수)
    - **왜곡** (커버와 스테고의 유사성)
    - **보안성** (스테그아날리시스에 대한 저항성)
  - **적응형과 비적응형으로 나뉨**
    - 적응형 : LSB 같이 정해진 위치 및 전체 데이터에 삽입
    - 비적응형 : 데이터 전체가 아닌 삽입이 용이한 (검출이 어려운) 위치에 삽입 → 성능이 더 좋음
  - LSB, PVD, UNIWARD, DCT 등의 여러 방식 존재

스테그아날리시스

- 스테고 오브젝트를 탐지하는 기술
- 지금까지 데이터의 통계적 특성만을 이용한 방식을 사용하였으나, 더 정교한 스테가노그래피 기술들이 등장하면서 최근에는 딥러닝 기반의 스테가노그래피 기술들이 등장
- **특징 추출과 탐지 과정**으로 나뉨
- 특징 추출
  - 데이터가 가지는 다양한 **통계적 특성들을 추출**
  - 대표적인 특징 추출에는 **SPAM, SRM, DCTR** 등이 있음

- **탐지**
  - 위의 과정을 통해 얻은 특징들을 **통계적으로 분석하거나 SVM 또는 딥러닝을 통해 분석**하여 스테고 오브젝트인지 탐지

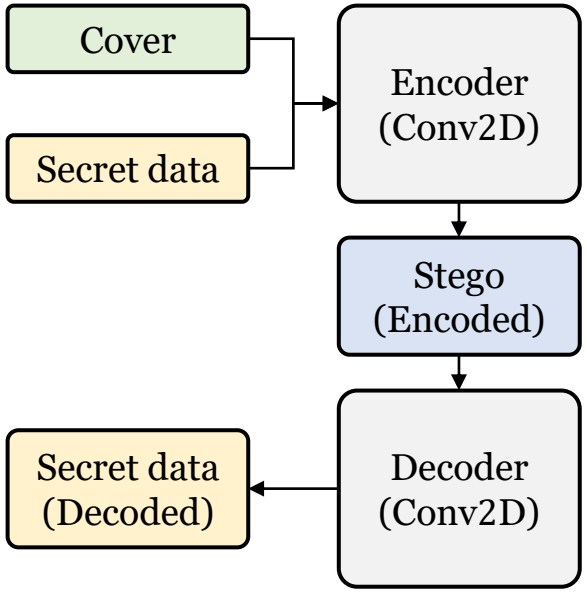
딥러닝 기반의 스테가노그래피

- 기존의 스테가노그래피 기술 (LSB, PVD 등)의 **품질을 보완**하기 위해 제안되기 시작
- **기계학습 기반의 방식**들은 데이터 품질은 향상 시킴, 그러나 **계산 복잡도가 높고 많은 데이터를 임베딩할 수 없음**
- 이러한 한계점을 보완하기 위해 CNN 기반의 모델을 시작으로 하여 더 나은 성능을 위한 GAN 기반의 모델까지 연구됨
- 결과적으로 데이터의 품질 향상 성공

딥러닝 기반의 스테가노그래피 - CNN

- **주로 인코더와 디코더가 결합된 구조 사용**
  - 독립적으로 학습되는 것이 아니라 인코더의 입력부터 디코더의 출력까지 **end-to-end로 학습**

- 일반적인 (공통적인) 동작 과정
  - **커버 오브젝트와 비밀 데이터를 병합한 후, 인코더에 입력**
  - 인코더는 이를 기반으로 커버 오브젝트에 비밀 데이터를 임베딩한 형태의 **스테고데이터 생성**
  - 디코더는 생성된 스테고 오브젝트를 기반으로 다시 커버 오브젝트에 숨겨진 **비밀 데이터 복원**



딥러닝 기반의 스테가노그래피 - CNN

- **연구 사례 1**
  - 비밀 데이터 : 이미지 → 적은 용량이 아닌 동일한 크기의 이미지 숨김

  - **인-디코더 구조 활용**
    - 컨볼루션 레이어 사용
    - Highway / ResNet / Inception 등에서 영향을 받아 **컨볼루션 기반의 잔여 네트워크 구성**
      → 모델이 더 깊어질수록 성능이 저하될 수 있는데 이를 보완하기 위한 구조들을 사용

  - 커버와 스테고 데이터 간의 **히스토그램의 유사성** 측면
    - 해당 기술이 **3-bit LSB 방식보다 더 유사 → 더 강력한 임베딩**

  - 그러나 이미지에서 텍스처가 풍부하지 않은 부분 (그냥 검정색, 흰색 픽셀들로 이루어진 영역)에서는 **약간의 노이즈 발생**

  - 이러한 단점은 VAE (Variational Auto Encoder), GAN과 같은 **더 정교한 생성 능력을 가진 신경망을 통해 극복** 가능할 것으로 예상

딥러닝 기반의 스테가노그래피 - CNN

- **연구 사례 2**
  - 비밀 데이터 : 이미지 → 여러 이미지를 하나의 커버에 숨기는 (**임베딩 용량이 큰**) 경우에 대한 실험도 진행

  - **전처리 + 인-디코더 구조가 end-to-end 형태로 학습**
    - **전처리** : **비밀 데이터** (이미지)의 RGB 채널에 대한 **변환** → 7 채널으로 변경
    - **인코더** : 커버 오브젝트와 변환된 비밀 데이터를 결합하여 인코더에 입력함으로써 **스테고 오브젝트 생성**
    - **디코더** : 스테고 오브젝트로부터 **비밀 데이터 복원 (추출)**

  - **모든 채널에 고르게 퍼져서 은닉**되었음을 확인
    - 이로 인해 단순한 스테그아날리시스로는 알아낼 수 없음

  - 생성 능력은 **컨볼루션의 크기와 구조에 따라 달라짐**을 밝힘

  - **딥러닝 기반의 스테그아날리시스**를 수행한 결과, **LSB 임베딩에 비해 5~9% 더 낮은 탐지 정확도 달성 → 더 강력**

  - 추가로, 비밀 데이터를 난독화 (permutation 함)하여 은닉하는것도 가능하고, 이를 기반으로 비밀 데이터 복구도 가능함을 보임

딥러닝 기반의 스테가노그래피 - GAN

- GAN이 갖는 적대적 학습 구조를 **스테가노그래피와 스테그아날리시스의 적대적 관계에 적용**

- 두 가지 관점 존재
  - **스테가노그래피를 위한 커버 오브젝트 생성**
    → 스테그아날리시스에 덜 민감한 커버 오브젝트를 생성

  - **스테고 오브젝트 생성**
    → 스테고 오브젝트 자체를 생성

딥러닝 기반의 스테가노그래피 - GAN

- **스테가노그래피를 위한 커버 오브젝트를 생성하는 방식 – 연구 사례 1**

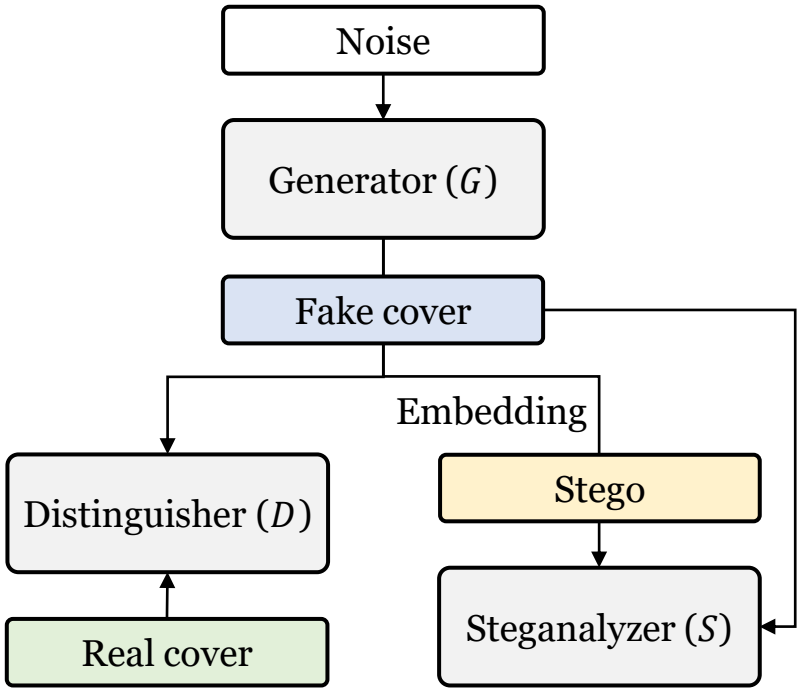
- 3개의 신경망으로 구성
  - G : 랜덤 벡터로부터 실제 커버와 비슷한 **가짜 커버 생성**
  - D : G가 생성한 **가짜 커버와 실제 커버를 구별**
  - S : G가 생성한 **가짜 커버**에 비밀 데이터를 임베딩하여 **스테고 오브젝트를 생성한 후, 해당 스테고 데이터와 가짜 커버를 분류**

- GAN의 기본적인 손실 함수를 기반으로 G와 D의 손실 + G와 S의 손실

- **G와 D 간의 학습**
  - 실제 커버와 유사

- **G와 S 간의 학습**
  - 비밀 데이터가 임베딩 되었는지 알기 어렵도록 하는 커버 생성

- 즉, 실제 커버와 유사하면서 스테그아날리시스에 대한 저항성을 갖는 커버 생성



딥러닝 기반의 스테가노그래피 - GAN

- **스테가노그래피를 위한 커버 오브젝트를 생성하는 방식 – 연구 사례 2**

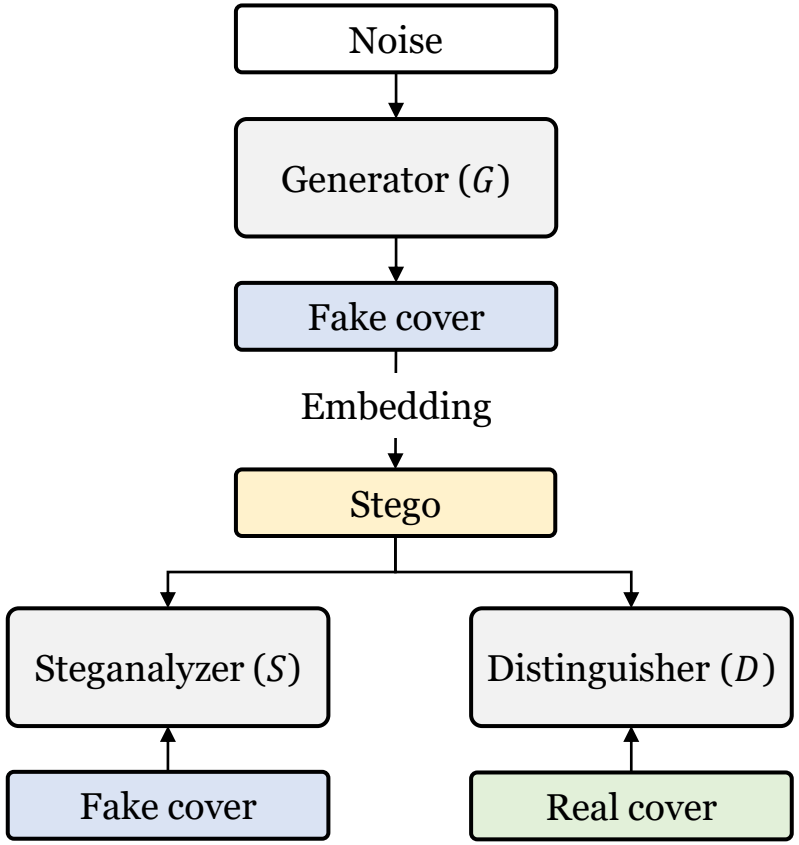
- D의 기능과 전체 구조를 변형하여 약간 다른 관점에서의 커버 데이터 생성

- 3개의 신경망으로 구성
  - G : 랜덤 벡터로부터 **가짜 커버 생성**
  - G가 생성한 가짜 커버를 기반으로 임베딩하여 **스테고 이미지 생성 후, D와 S에 입력**
  - D : **스테고 이미지와 실제 커버를 구별**
  - S : **스테고 이미지와 가짜 커버를 구별**

- **G와 D 간의 학습**
  - **스테고 데이터가 실제 데이터처럼 보이도록 학습**
  - G가 생성한 가짜 커버는 실제 커버와 비슷하지 않음
  - 그러나 **임베딩 (스테가노그래피) 후에 비로소 실제 커버와 비슷**해짐

- **G와 S 간의 학습**
  - 비밀 데이터가 **임베딩 되었는지 알기 어렵도록 하는** 커버 생성

- 즉, 스테가노그래피가 수행되었을 경우 실제와 비슷해지는 커버를 생성
  + 스테그아날리시스에 대한 저항성을 갖는 커버 생성



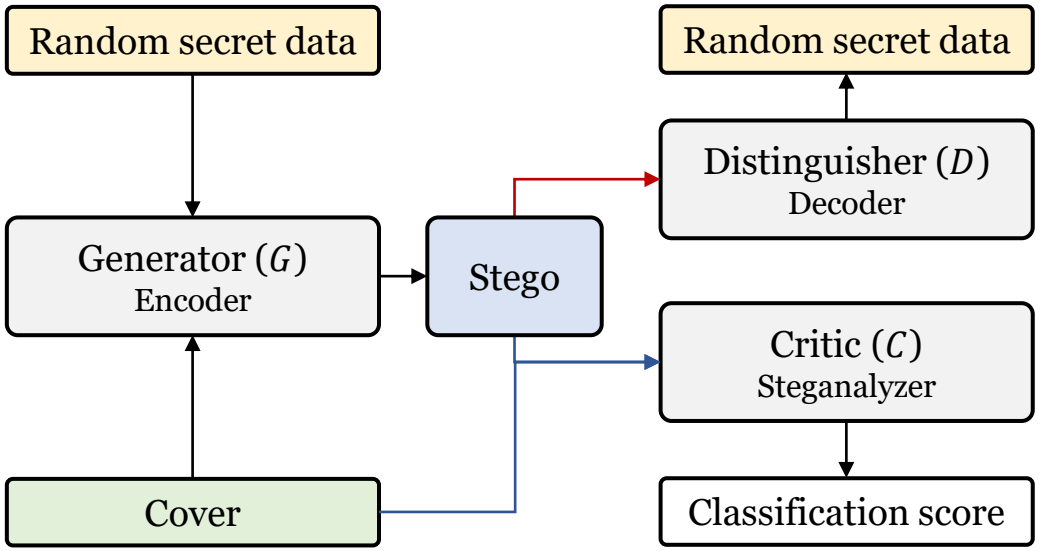
딥러닝 기반의 스테가노그래피 - GAN

- **스테고 오브젝트를 생성하는 방식**

- 3개의 신경망으로 구성
  - **G : 스테고 이미지를 생성**하는 인코더 역할
    - 기존 GAN에서의 **랜덤 벡터가 비밀 데이터로 동작**
    - 기존 GAN과 다르게 랜덤 벡터 뿐만 아니라 커버 데이터를 병합하여 입력 받음
    - CNN 기반의 방식과 동일하게 커버 데이터와 비밀 데이터를 기반으로 스테고 이미지 생성
  - **D :** 숨겨진 **비밀 데이터를 복원**하는 디코더 역할
  - **C :** 스테고 이미지와 실제 커버를 분류하는 **스테그아날리시스** 수행

- **G와 D 간의 학습**
  - **이미지 왜곡이 적은 스테고 이미지 생성**

- **G와 C 간의 학습**
  - **스테그아날리시스에 강력한 스테고 이미지 생성**

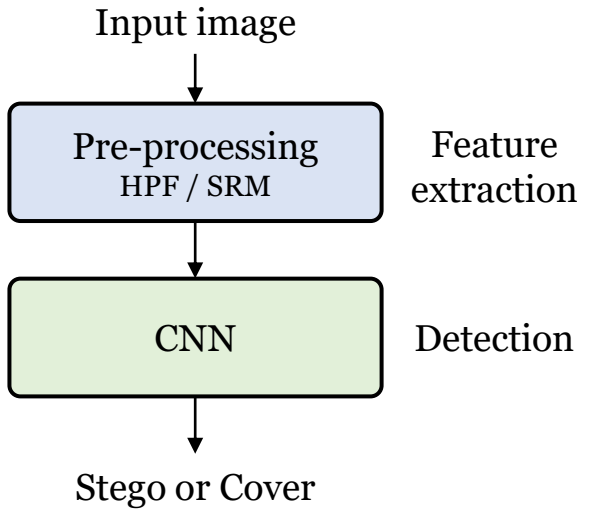


딥러닝 기반의 스테그아날리시스

- 기존의 통계적 방식 및 SVM 기반의 스테그아날리시스와 다르게 특징 추출부터 탐지까지의 과정을 통합하여 학습

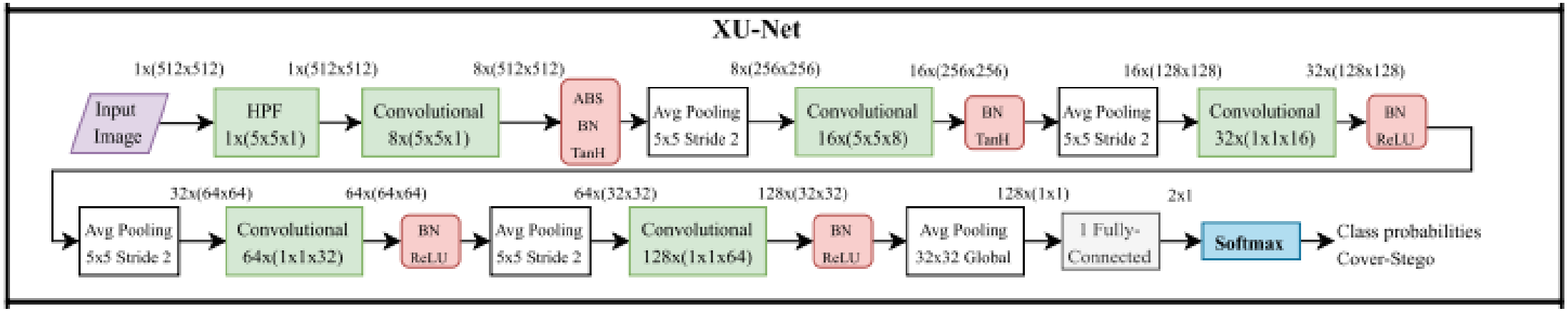
- 대부분의 연구에서 공통적으로 갖는 구조는 오른쪽 그림과 같음
  - **전처리**
    - 입력된 스테고 이미지 또는 커버 이미지에 대해 HPF 및 SRM 적용
    - **이러한 과정이 필요한 이유?**
      - 이미지의 내용이나 의미가 아닌 숨겨진 정보를 분석해야하므로 **일반적인 이미지 분류와 다름**
      - **이미지의 경계선이나, 고/저주파 등의 부분에 대한 특징 추출 및 분석이 가능해야하므로** 이를 강조할 수 있는 필터 필요
      - SRM의 경우 기존의 특징 추출 기술이며, SRM이 갖는 필터 값을 컨볼루션 필터에 적용하는 방식
    - 해당 과정까지 역전파 과정에 포함시키기도 하고, 아닌 경우도 있음
  - **탐지**
    - 딥러닝을 적용한 최초의 연구에서 CNN을 사용하였으며, 이후 모두 CNN을 사용함

- 다음 페이지부터 대표적인 딥러닝 기반 스테그아날리시스 모델들



딥러닝 기반의 스테그아날리시스 – Xu-Net

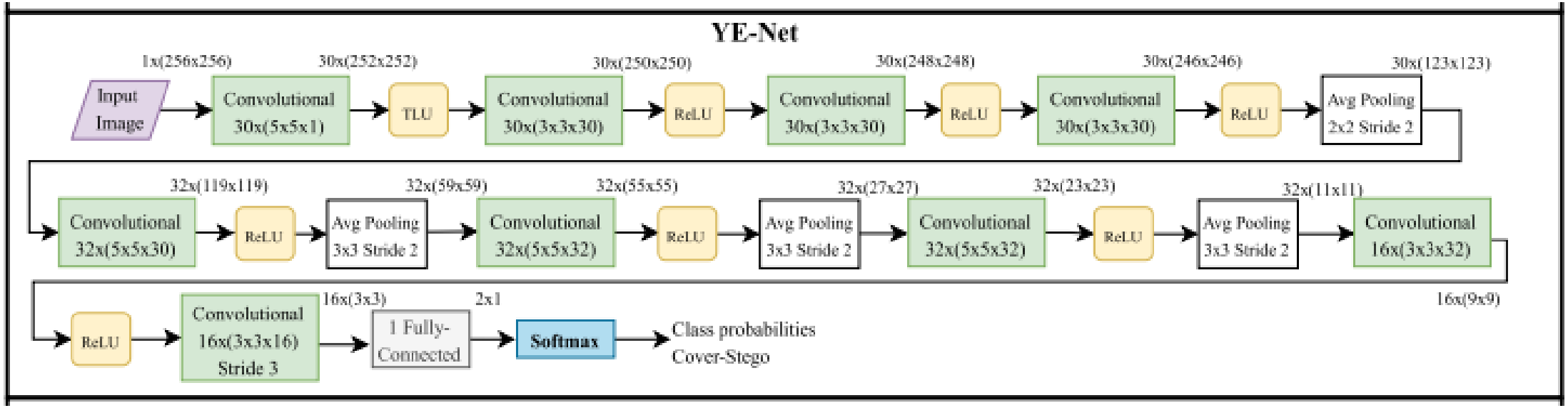
- **전처리** : 5x5 HPF → **잔여 정보** (Residual information)
- **특징 추출 및 탐지 :** 5개의 CNN 레이어
  - **1번째 컨볼루션 레이어 : ABS / BN / Tanh 가 수행**
    - ABS : HPF를 통해 나온 잔여 정보에 **절댓값**을 취함
    - BN : **정규화**
    - Tanh : ABS+BN을 거쳐 나온 값이 0을 중심으로 분포하는데 이러한 분포를 **잘 학습할 수 있는 활성화 함수** 적용
  - **평균 풀링** (Average Pooling) : 뒤에 나올 모델들에서도 사용
    - 노이즈가 많은 스테고 이미지에 대한 학습에서 노이즈 및 작은 변화들에 대한 **불필요한 영향을 감소시키고, 중요 정보 유지** 가능
      → **견고하고 일반화 된 학습 가능**
  - 이후의 레이어들에서는 ABS+BN는 적용하지 않음 (전처리 직후의 잔여 정보에 대한 학습을 위한 과정이므로)



딥러닝 기반의 스테그아날리시스 – Ye-Net

- **전처리** : SRM 필터 (다양한 크기와 방향성을 갖는 커널) 중 30개를 사용
  - SRM 필터가 적용된 컨볼루션 레이어를 **역전파 과정에 포함**
    → 학습에 의해 커널 값이 갱신되도록 함
  - 해당 레이어에는 **TLU** (Truncated linear unit) 활성화 함수 사용
    → 분포를 잘 학습하여 성능 향상 되도록 적용
  - Xu-Net과 다르게 ABS와 BN을 전처리 후의 피처에 적용하지 않음

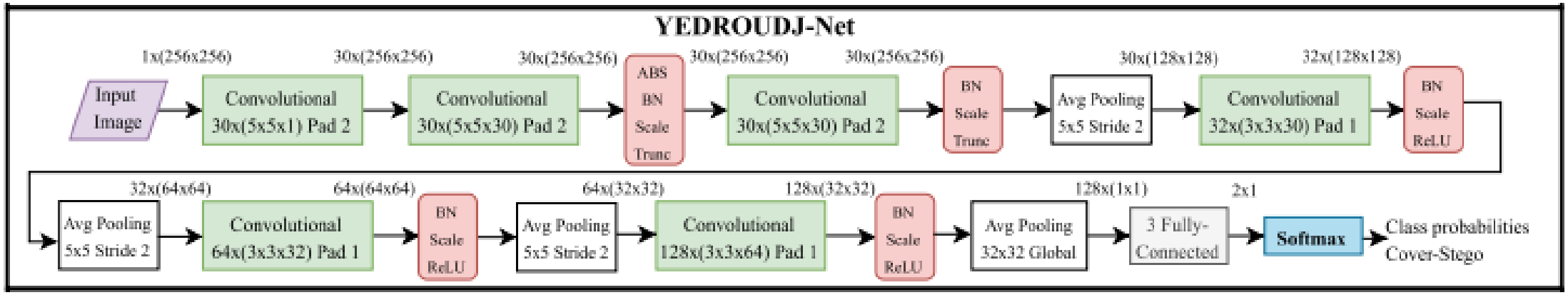
- **특징 추출 및 탐지 :** Xu-Net보다 많은 수의 컨볼루션 레이어 사용 (파라미터 증가)



딥러닝 기반의 스테그아날리시스 – Yedroudj-Net

- **Xu-Net과 Ye-Net의 기법을 활용**
- **전처리** : Ye-Net의 **SRM 필터 + TLU** 활성화 함수 사용

- **특징 추출 및 탐지 :** Xu-Net의 ABS, BN를 사용
  - 첫 번째 컨볼루션 레이어 : **ABS + BN**을 적용하였으며, 활성화 함수는 **TLU**로 대체
  - 이후에는 BN, TLU, ReLu 등을 사용하는 컨볼루션 레이어 4개 거친 후 분류
  - Xu-Net과 Ye-Net에 비해 **성능 향상**



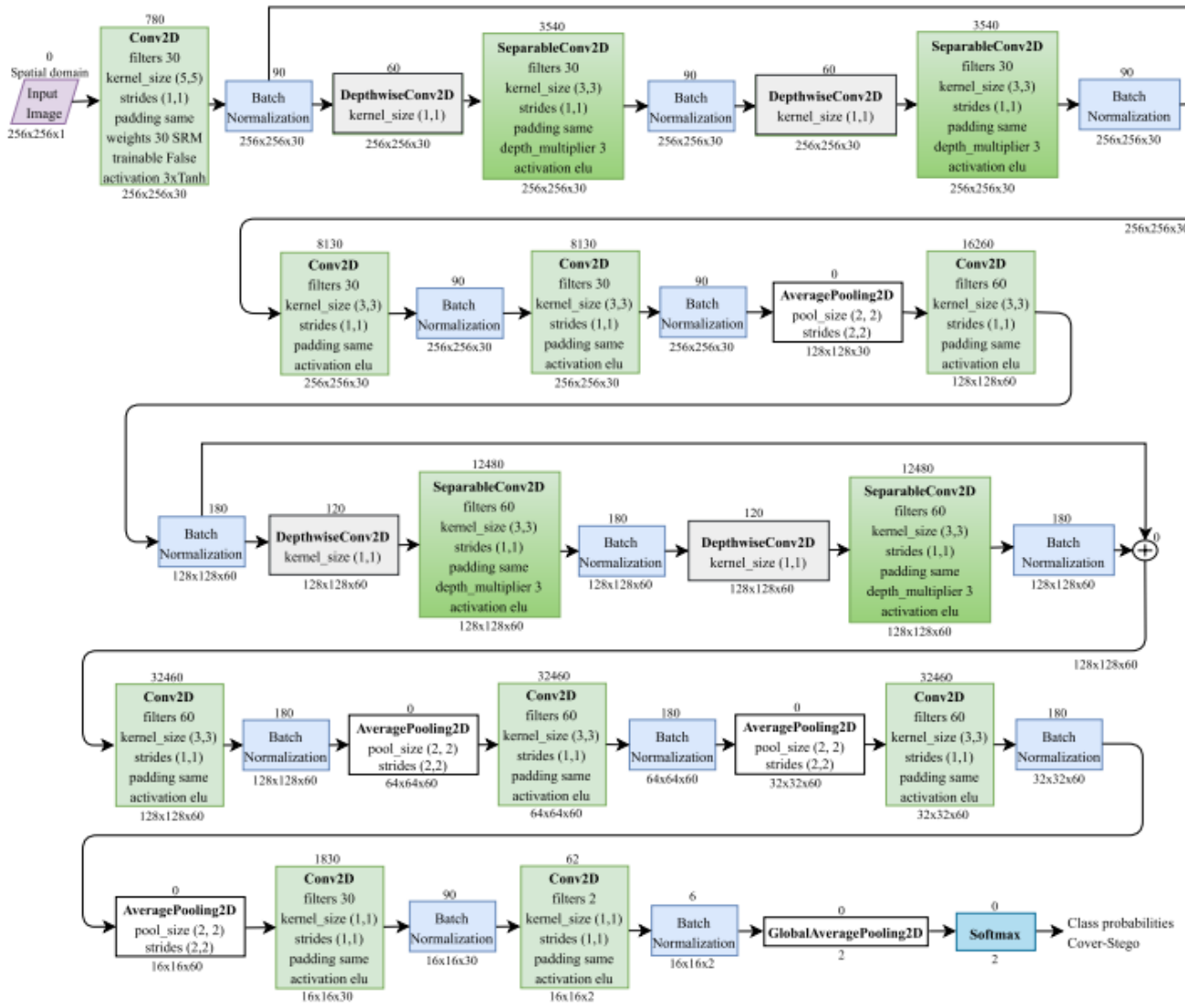
딥러닝 기반의 스테그아날리시스 – GBRAS-Net

- **전처리** : Ye-Net의 **SRM 필터 + 3 Tanh**

- **특징 추출 및 탐지 : 효율적인 CNN 사용**
  - 일반, 깊이별 + 분리 컨볼루션 사용
  - 깊이별 컨볼루션 (Depth-wise Conv)
    - 입력 채널마다 각각 필터를 적용
  - 분리 컨볼루션 (Separable Conv)
    - 하나의 2D 커널을 두 개의 1D 커널로 분해

- 깊이별 + 분리 컨볼루션 통해 성능을 저하시키지 않으면서 파라미터를 줄여 계산 효율성을 증가시킴

- 이러한 구조를 통해 정확도를 조금 더 향상시키면서 파라미터를 줄인 효율적인 스테그아날리시스 모델 구성



딥러닝 기반의 스테그아날리시스 – 성능 비교

- 아래 표는 **대표적인 모델 + 대표적인 스테가노그래피 데이터셋에 관한 성능 비교** 표임

- **정확도 / 딥러닝 네트워크 / 개선 요소 관점**에서 비교한 내용은 다음 페이지부터..

표 x 대표적인 딥러닝 스테그아날리시스 기술 비교

	WOW		S-UNIWARD		Parameters
	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp	
[X]	67.5	79.3	60.9	72.7	87830
[Ye]	66.9	76.7	60.1	68.7	88596
[Yd]	72.3	85.1	63.5	77.4	252459
[G]	80.3	89.8	73.6	87.1	166599

그 스테가노 논문에 넣을 표입니다..

딥러닝 기반의 스테그아날리시스 – 성능 비교

- **정확도**
  - 두 임베딩 알고리즘 모두에 대해 **높은 bpp에 대한 정확도가 높음**
    → 임베딩 용량이 많을수록 스테고 오브젝트로 탐지되기 쉽기 때문에 당연한 결과

  - 스테그아날리시스 모델 기준으로:
    - **GBRAS > Yedroudj > Xu > Ye** 순으로 높은 정확도
    - 그러나, Ye-Net은 다른 데이터셋까지 합치게 되면 성능이 향상되긴 함
    - **GBRAS와 초기에 연구된 Xu 사이의 격차가 상당히 큼**

표 x 대표적인 딥러닝 스테그아날리시스 기술 비교

	WOW		S-UNIWARD		Parameters
	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp	
[X]	67.5	79.3	60.9	72.7	87830
[Ye]	66.9	76.7	60.1	68.7	88596
[Yd]	72.3	85.1	63.5	77.4	252459
[G]	80.3	89.8	73.6	87.1	166599

딥러닝 기반의 스테그아날리시스 – 성능 비교

- **딥러닝 네트워크 관점**
  - 사용된 네트워크는 **기본적인 CNN 구조가 초기**에 주로 사용
  - 이후 컨볼루션 레이어들의 개수가 추가됨에 따라 **계산 복잡도 및 파라미터가 증가**하는 추세를 보임
  - 그러나, GBRAS-Net에서는 **효율적인 CNN** 구조를 사용함으로써 **성능은 향상시키고 파라미터는 감소시킴**

표 x 대표적인 딥러닝 스테그아날리시스 기술 비교

	WOW		S-UNIWARD		Parameters
	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp	
[X]	67.5	79.3	60.9	72.7	87830
[Ye]	66.9	76.7	60.1	68.7	88596
[Yd]	72.3	85.1	63.5	77.4	252459
[G]	80.3	89.8	73.6	87.1	166599

딥러닝 기반의 스테그아날리시스 – 성능 비교

- **성능 향상을 위한 개선 요소**
  - 여러 요소들을 개선시킴
  - **전처리**
    - **HPF / SRM** 사용
    - SRM을 **학습 과정에 포함**시키도록 함
    - SRM을 커널로 사용한 레이어에서는 **활성화 함수를 조절** → 정확도 향상을 위해 전처리 이후의 값을 정제하기 위한 시도

  - 앞에서 살펴보았듯이 절댓값이나 정규화 등을 통해 **잔여 정보를 더 잘 학습할 수 있도록** 하는 기법들도 있음

  - 이외에도 **컨볼루션의 커널 개수나 필터 크기**가 정확도에 영향을 미치는 것으로 파악됨

  - 여러 딥러닝 기술들의 **조합을 통해 성능 향상 및 계산 복잡도 및 모델 크기를 감소 시키는 방향**으로 개선할 수 있을 것으로 보임

표 x 대표적인 딥러닝 스테그아날리시스 기술 비교

	WOW		S-UNIWARD		Parameters
	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp	
[X]	67.5	79.3	60.9	72.7	87830
[Ye]	66.9	76.7	60.1	68.7	88596
[Yd]	72.3	85.1	63.5	77.4	252459
[G]	80.3	89.8	73.6	87.1	166599

감사합니다.