●○ 피복지도 및 산림수종

# 산림수종 이미지 데이터(강원 및 충청)

## ●○ 데이터 구축 개요: 강원충청권 산림수종 AI 학습데이터셋이란?

| 과제명 | 주요 내용 | 데이터 수집 방법 | 데이터 구축량 | 데이터 형식 |
|---|---|---|---|---|
| 산림수종 이미지 데이터 (강원·충청권) | · 환경변화탐지 AI기술 개발을 위한 소나무, 낙엽송, 침엽수, 활엽수 4종의 항공영상 및 침엽수, 활엽수 2종 모사영상 데이터 구축 | · 영상기반으로 저작도구 등을 활용한 annotation<br>· 업무효율화를 위해 fine/coarse annotation<br>· 크라우드 소싱을 통한 학습용 데이터 구축 병행 | · fine annotation : 항공이미지16,830건 모사이미지 170건<br>· coarse anntation : 항공이미지32,670건 모사이미지 330건 | · 원본영상 Tiff<br>· GT데이터 Tiff, Json<br>· 항공이미지 1024×1024px 크기<br>· 모사이미지 512×512px 크기 |

- 산림수종 AI 개발을 위해, 항공이미지 및 모사이미지를 사용하여 영상의 객체를 분할하는 학습데이터를 구축
  - AI 학습용 데이터는 Fine annotation 1만7천건, Coarse annotation 3만3천건 구축

| 구분 | | | 수량 | 단위 | 총계 |
|---|---|---|---|---|---|
| 어노테이션 | 산림 Fine annotation | 항공이미지 1024 * 1024 | 32,670 | 장 | 33,000 |
| | | 모사이미지 512 * 512 | 330 | 장 | |
| | 산림 Coarse annotation | 항공이미지 1024 * 1024 | 66,330 | 장 | 67,000 |
| | | 모사이미지 512 * 512 | 670 | 장 | |

  - 25cm 항공이미지을 활용하여 산림수종 4종의 Fine/Coarse 객체 분할 학습데이터(1024*1024 픽셀)을 구축
  - 5m 모사이미지를 활용하여 산림수종 2종의 Fine/Coarse 객체 분할 학습데이터(512*512 픽셀)을 구축
- AI 학습용 데이터 구축 저작도구 개발 및 활용으로 크라우드소싱을 통한 학습데이터 구축
  - 크라우드소싱 추진 시, AI 학습용 데이터의 원활한 구축을 위해 객체를 구획하고, 라벨이 가능한 저작도구 개발·제공하며, 산림수종별 데이터셋 구축 지침을 제공

## ●○ 데이터셋의 구성

- 데이터셋의 어노테이션 정보는 gray scale의 Tiff 이미지 형태로 제공하고, 원천영상, 이미지 크기 등 데이터에 대한 정보를 Json 포맷으로 제공
- 메타데이터 항목

| No | 항목 | | 타입 | 필수 여부 | 예시 |
|---|---|---|---|---|---|
| | 영문명 | 한글명 | | | |
| 1 | Image | 이미지 정보 | | | |
| 1-1 | img_id | 이미지 식별자 | String | Y | LC_AP_00000000_001 |
| 1-2 | img_width | 이미지 너비 | Number | Y | 512 |
| 1-3 | img_height | 이미지 높이 | Number | Y | 512 |
| 1-4 | img_type | 이미지 종류 | String | Y | 항공영상 |
| 1-5 | img_coordinate | 이미지 좌표계 | String | Y | EPSG:5186 |
| 1-6 | coordinates | 좌상단 픽셀 중심 좌표 | String | Y | 000, 000 |
| 1-7 | img_resolution | 이미지 해상도 | Number | Y | 0.51 |
| 1-8 | img_winter | 겨울영상 여부 | Number | Y | 1, 0 |
| 2 | annotations | 어노테이션 정보 | | | |
| 2-1 | ann_id | 어노테이션 식별자 | String | Y | LC_AP_00000000_001_FGT |
| 2-2 | ann_type | 어노테이션 타입 | String | Y | polygon |
| 2-3 | ann_file_type | 어노테이션 파일 유형 | String | Y | tif |

  - 이미지 식별자는 이미지 파일명을 나타냄
  - 이미지 종류, 좌표계, 해상도는 어노테이션에 사용한 원천 영상의 종류, 좌표계, 해상도를 나타냄
  - 좌상단 좌표는 학습용 이미지 데이터의 좌상단 X, Y 좌표를 나타냄
  - 겨울영상 여부는 원천영상 촬영계절의 겨울여부를 나타냄
  - 이미지 너비 및 높이 정보는 학습을 위한 데이터의 크기를 나타냄
  - 어노테이션 식별자는 어노테이션 파일명을 나타냄
  - 어노테이션 타입은 어노테이션 툴에서 지원하는 모양 중 사용한 모양을 나냄

## ●○ 데이터셋의 설계 기준과 분포

### | 효율적인 산림 훼손 모니터링 AI 기술 개발을 위한 학습용 데이터 분류항목 |

- 학습용 데이터는 산림 훼손 모니터링을 기반으로 임상도 제작, 영상 자동분류 분야, 임상 자동 구획 등 다양한 활용분야를 고려하여 4종의 수종에 대해 설계

- 소나무, 낙엽송, 기타침엽수, 활엽수 4종의 산림수종에 대한 항목은 산림 훼손 모니터링을 위한 주요 항목으로, 학습용 데이터 구축 및 AI 적용 시 임상도 현행화 업무에 활용이 용이함

• 학습용 데이터 분류항목

| 수종 | 상세 수종 |
|---|---|
| 소나무 | 소나무 |
| 낙엽송 | 잎갈나무, 일본잎갈 |
| 기타침엽수 | 잣나무, 전나무, 편백나무, 삼나무, 가문비나무, 비자나무, 은행나무 등 |
| 활엽수 | 상수리나무, 신갈나무, 굴참나무, 오리나무, 고로쇠나무, 자작나무, 박달나무, 밤나무 등 |

• 산림수종 이미지 데이터 대상


〈소나무〉


〈일본잎갈〉


〈편백나무〉


〈상수리나무〉

## ●○ 데이터셋 증식·확산 추진을 위한 학습용 데이터 구조 설계

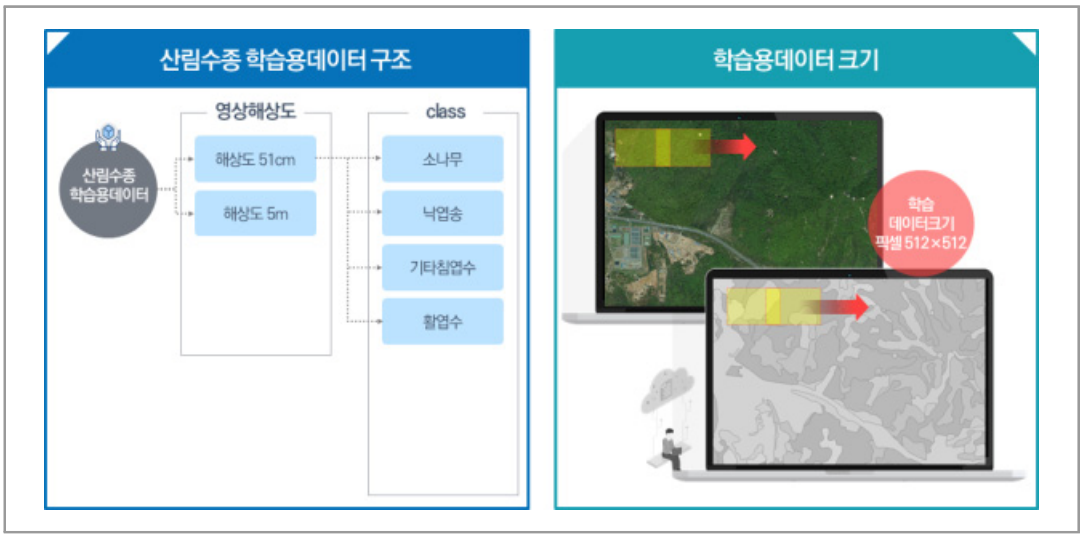
• 산림수종 학습용 데이터는 항공영상(해상도 51cm), 산림위성 대비 영상(해상도 5m)기반으로 각 해상도별 4개 클래스로 분류

- 학습용 데이터는 해상도 51cm, 5m 영상파일과 해당 영상 기반 4개 항목으로 annotation된 이미지 파일(포맷:8bit, Tiff)로 구성

- 영상해상도 2종(51cm, 5m), 클래스 4종(소나무, 낙엽송, 기타침엽수, 활엽수)을 포함하며, 파일명은 '해상도_class_image.Tiff', '해상도_class_annotation.Tiff'로 구성

- 원천데이터로 사용되는 항공영상이나 산림위성 대비 영상은 학습용 데이터로 사용하기에는 큰 사이즈이므로 활용에 적합한 크기로 재단해야함
  - annotation 시 활용한 영상데이터를 각 해상도별 픽셀 크기 512*512, 임의의 중복률로 재단하여 학습용 데이터로 작성
  - annotation 결과를 영상데이터와 동일한 영역 및 중복률로 각 해상도별 픽셀 512*512 크기로 재단하여 학습용 데이터로 작성

- 산림수종 학습용 데이터 구조 및 크기




## 데이터 구조

- 산림수종 항공사진
  - 국토지리정보원에서 제작·배포하는 2018~2019년 촬영 영상으로 선정
  - 해상도 25cm, 사이즈 1024*1024

- 산림수종 모사영상
  - 구름과 눈이 덮이지 않은 2019~2020년 촬영 영상으로 선정
  - 해상도 10m, 사이즈 512*512

- 데이터 형태 및 규모
  - 데이터 분류 : 인공지능 학습용 데이터 (강원충청 산림수종 이미지 AI데이터)
  - 데이터 개수 : 50,000 장

## ●○ 데이터 예시

- 산림수종 이미지 총 구축 데이터에 대한 폴더구조 및 파일명 규칙
- 50,000장의 산림수종 이미지 데이터를 한 세트로 하여 폴더 구조 생성

### | File ID체계 |

| 구분(2) | 영상명(2) | 도엽번호(8), 영상일련번호(4) | 일련번호(3) | GT 구분 |
|---|---|---|---|---|
| FR:산림수종 | 항공 : AP | 00000000 | 001,002,.. | Fine : FGT<br>Coarse :CGT |
| | 모사이미지 : PI | 0000 | | |

### | 적용형식 및 예시 |

| 파일종류 | 형식 | 예시 |
|---|---|---|
| 원시영상 | FileID.tif | FR_AP_00000000_001.tif<br>FR_PI_0000_001.tif |
| Annotation | FileID_FGT.tif<br>FileID_CGT.tif | FR_AP_00000000_001_FGT.tif<br>FR_AP_00000000_001_CGT.tif<br>FR_PI_0000_FGT.tif<br>FR_PI_0000_CGT.tif |
| JSON | FileID.json | FR_AP_00000000_001_FGT.json<br>FR_AP_00000000_001_CGT.json<br>FR_PI_0000_FGT.json<br>FR_PI_0000_CGT.json |
| SHP | FileID_FGT.shp<br>FileID_CGT.shp | FR_AP_00000000_001_FGT.shp<br>FR_AP_00000000_001_CGT.shp<br>FR_PI_0000_FGT.shp<br>FR_PI_0000_CGT.shp |

※ SHP : annotation 작업 중간산출물

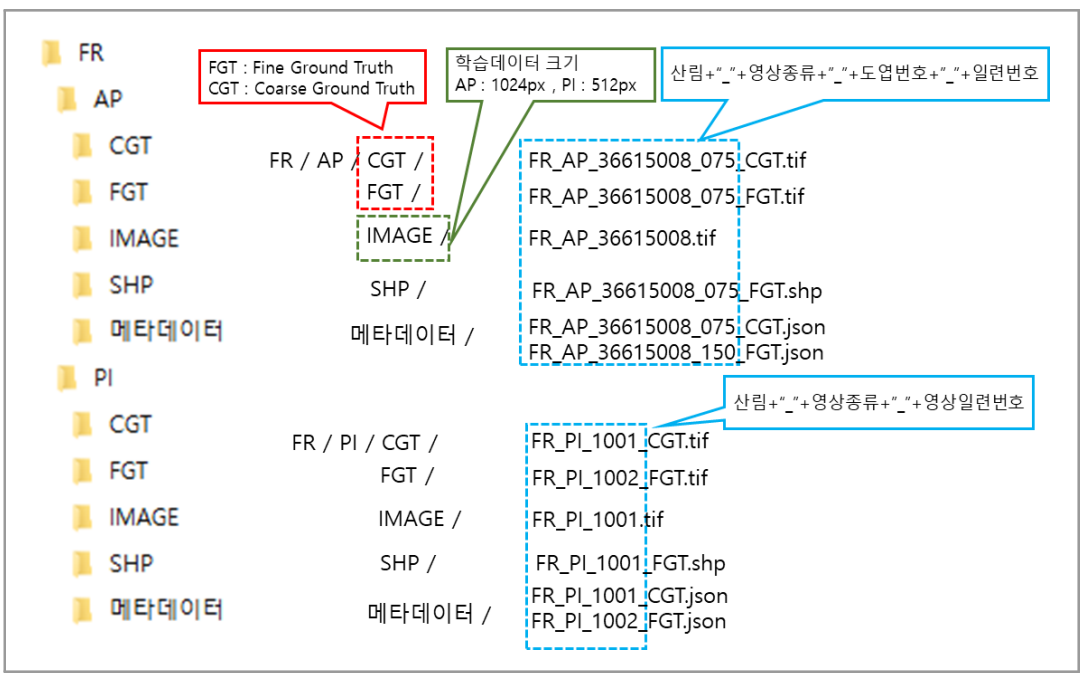
### | SHP 데이터 명세서 |

- 산림청 표준 용어를 준수하여 표준화 실시

| 컬럼명 | 데이터타입(길이) | 컬럼명 설명 |
|---|---|---|
| ANN_CD | NUMERIC(3) | 어노테이션 코드 |
| ANN_NM | VARCHAR2(20) | 어노테이션 명칭 |

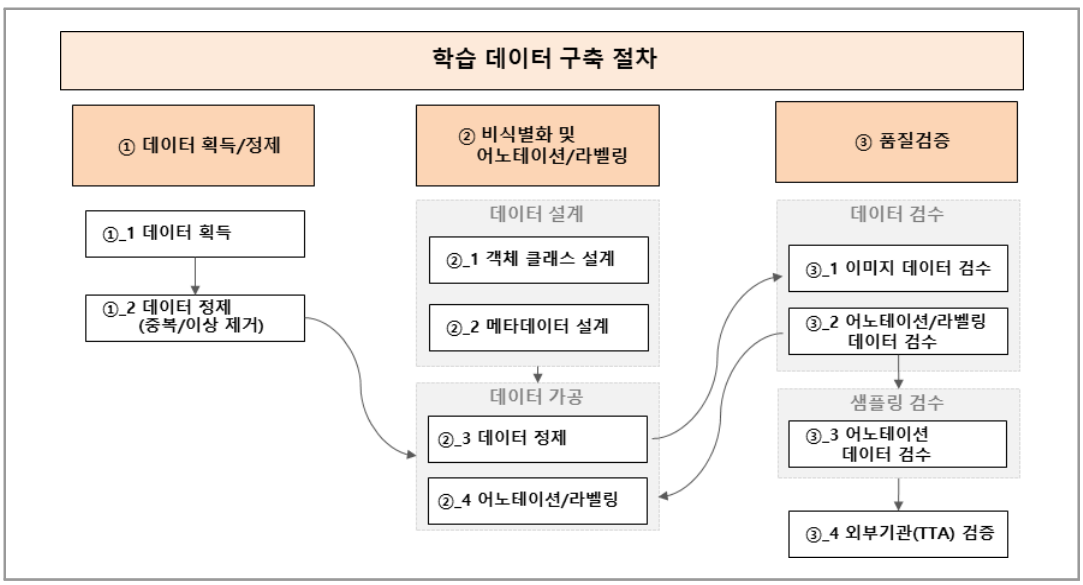
| 컬럼명 | 코드값 | 컬럼명 설명 |
|---|---|---|
| ANN_CD | 110 | 소나무 |
| | 120 | 낙엽송 |
| | 130 | 기타침엽수 |
| | 140 | 활엽수 |
| | 150 | 침엽수 |
| | 180 | 판독불가 |
| | 190 | 비산림 |

## | 파일 저장 디렉토리 구조 |




## ●○ 데이터 구축 과정

- 본 사업의 데이터 구축 과정은 단계별로 구성하며, 각 단계별 세부 절차 내역을 정의함
- 본 과제의 학습데이터는 다수의 사용자에게 공개를 목적으로 구축되는 데이터이므로 학습데이터 내에는 국가보안에 대한 정보를 식별할 수 있는 정보가 제거되어야 함




## | 데이터 획득/정제 |

- 원시데이터는 항공 영상(국토지리정보원) 및 위성 영상 (RapidEye) 데이터를 획득하여 활용함
- 항공 영상 해상도 25cm, 2018~2019년 촬영 영상으로 선정
- 위성영상 ( RapidEye ) 해상도 5m 영상으로 선정

## | 데이터 가공(어노테이션/라벨링) |

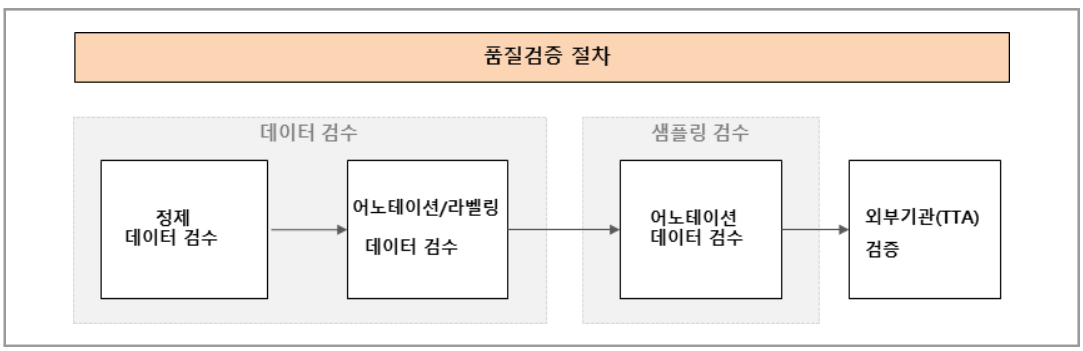
- 객체 클래스 및 메타데이터 설계를 통하여 객체 클래스를 분류하고 코드화 설계함
- 정제 데이터에 학습대상 객체를 어노테이션 하고 객체 분류 및 속성 정보를 입력함

## | 데이터 품질검증 |

- 정제된 데이터를 대상으로 정제 불완전처리 여부를 점검함
- 어노테이션/레이블 완료 데이터를 대상으로 어노테이션 대상 객체가 누락되거나 잘못된 분류/속성이 입력되었는지 여부를 점검함
- 최종적으로 외부기관 (TTA)에 의한 검수를 수행함

## ●○ 검수와 품질 확보

### | 검수 절차 |




**(가) 정제 데이터 검수 (검수작업팀)**

- 정제된 데이터에 대한 검수를 진행
- 정제 데이터 검수 후 오류 리포트를 작성하고 데이터 정제팀에 수정 보완 요청 (어노테이션 검수 결과와 종합)
- 품질검증팀에서는 어노테이션/라벨링 검수와 동시 실시

**(나) 가공(어노테이션) 데이터 검수 (검수작업팀)**

- 최종 가공(어노테이션) 완료된 AI학습용 데이터에 대해 샘플링 검사 실시
  - 일정기간을 정해서 샘플링하여 검수하고 정확도가 목표치에 도달하지 못할 시에는 해당기간에 가공한 데이터를 가공팀에 재송신하여 보완작업을 수행함
- 검증용 View 도구를 이용 육안 검수를 진행
- 검증결과는 가공(Annotation)팀과 공유하여 상호 feedback을 주고 받으면서 목표품질수준에 도달할 때까지 iterative하게 반복(오류 리포트를 통한 주단위/일단위 커뮤니케이션)
  - 데이터 획득·정제팀은 데이터 획득·정제된 영상데이터를 검수할 수 있도록 송신(파일 전달)하며, 송신 데이터는 송신 파일 ID 및 파일명, 영역 (Meta) ID, 각각의 클래스 정의 Data로 항목을 구성함
  - AI학습용 데이터에 대하여 정의된 클래스 오류 CHECK 프로그램에 의한 수작업 확인을 통한 정상 여부 및 오류 여부를 검수함
  - 검수 오류 리포트 및 데이터를 다시 데이터 가공팀에 송신(리포트, 파일)하고 데이터 가공팀은 오류 데이터에 대하여 보완 작업 수행함
  - 오류 데이터 리스트 FILE 수신 → 어노테이션 팀에서 수정 반영 → 데이터 검수팀에 데이터 재송신, 검수팀에서 재검수 실시

### (다) 검수를 통과하지 못한 데이터의 처리 방법

- 검수 통과하지 못한 데이터는 오류대장 작성하고 어노테이션 팀에 전달하여 어노테이션 작업 재실시

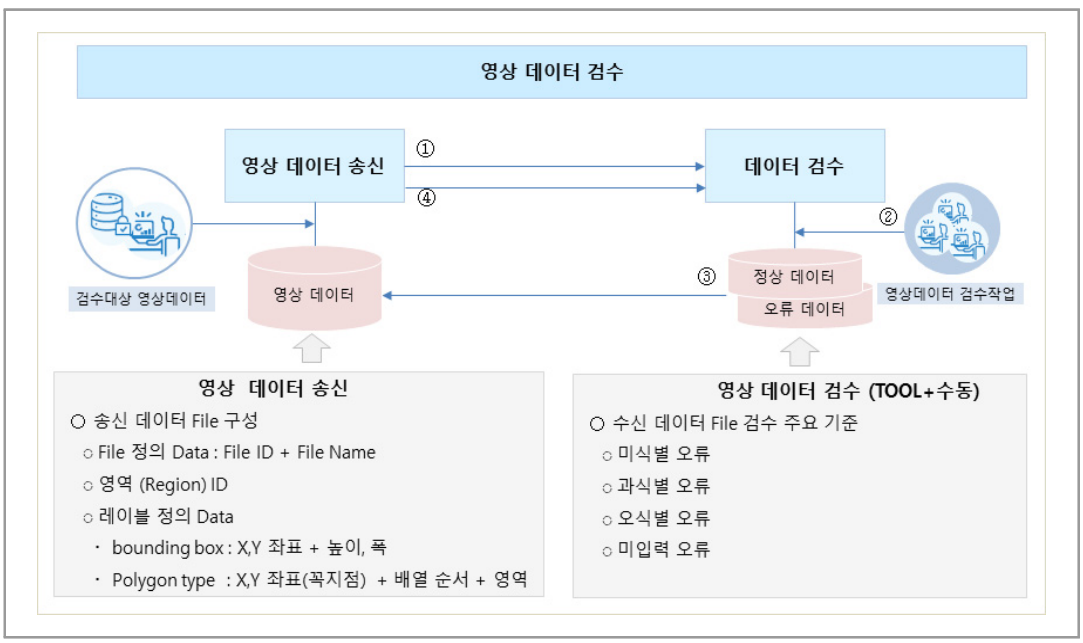



표 | 검수 오류 리포트 예시

검수 오류 리포트(2020.00.00)

| 검수자 | 검수일 | 이미지ID | 어노테이션 ID | 오류유형 | 비고 |
|---|---|---|---|---|---|
| 검수자1 | 2020-08-01 | 1201101005918 | 1609 | 오식별-라벨링 -레이블오류 | |
| 검수자2 | 2020-08-01 | 1201101005918 | | 미식별오류 | 좌상단 객체 |
| 검수자3 | 2020-08-01 | 1201101005918_4 | 1616 | 오식별-어노테이션 -과식별 오류 | |
| … | … | … | … | … | … |

### (라) 외부기관 (TTA) 검증

- 본 과제 결과물은 TTA를 통해 품질검증을 수행하고 품질검증 결과서 획득
- TTA와는 과제 수행 초반에 품질검증팀을 중심으로 협의채널, 협의방식 등을 포함한 상시협업 체계를 구축

- 1단계 각 공정별 작업내용에 대한 표준 가이드라인 초안이 나오면 이를 TTA에 제공하고 TTA의 검토 의견을 받아 가이드에 반영
- 2단계 구축 공정 품질검증 요구사항 기준으로 TTA 현장 검증 수검
- 3단계 데이터셋과 Annotation Tool 등 데이터 정확도/유효성 검증 계획서 기준으로 TTA에 제출하고 검증결과를 feedback받아 최종 데이터셋을 보완
- TTA 품질검증은 3단계에 걸쳐 검증 준비 및 수검

## ●○ 데이터 구축 담당자

수행기관(주관) : ㈜네이버시스템 최일훈 이사
(전화: 070-8821-1178, 이메일: ilhoon74@neighbor21.co.kr)