열화상 영상의 의미적 분할을 위한 상호 학습 기반의 비지도 도메인 적응

권석준 ○, 신정민, 한대찬, 최유경†
세종대학교 Robotics and Computer Vision (RCV) 연구실
{sjkwon, jmshin, dchan, ykchoi}@rcv.sejong.ac.kr

요 약

열화상 분할 기술은 낮은 조도, 급격한 조도 변화에 취약한 컬러 센서의 단점을 해결하기 위해 열화상 센서를 사용해서 의미적 분할 예측을 수행하는 기술이다. 열화상 분할 모델을 학습시키기 위해서는 큰 규모의 열화상 라벨링 데이터셋이 필요하지만 이는 매우 부족한 실정이며 열화상 센서의 부족한 질감 정보 등의 고유한 문제로 인해 의미적 분할 예측의 성능이 저하된다. 해당 문제들을 해결하고자 대규모 컬러 데이터셋에서 열화상 데이터셋으로 도메인 적응을 통한 지식 전이를 수행하는 멀티 스펙트럴 도메인 적응 방법론들이 제안되어 많은 주목을 받고 있지만 두 도메인의 고유한 차이 때문에 성공적인 도메인 적응에는 한계가 있다. 이에 우리는 두 도메인이 각각 예측한 의미적 분할 지도를 단순하게 단 방향으로 전달해서 손실을 계산하는 것이 아닌, 상호 학습을 통해 예측한 의미적 분할 지도가 각 도메인의 장점을 모두 반영하는 새로운 학습 기법을 제안한다. 또한 어텐션 메커니즘을 적용하여 특징을 추출하기에 효과적인 영역에 가중치를 부여하고, 두 도메인에서 생성된 어텐션 지도 사이의 손실을 추가하여 한층 더 효과적인 도메인 적응을 통한 지식 전이가 일어나도록 하는 새로운 목적 함수를 설계한다. 우리가 제안한 방식은 MF 데이터셋에서 최첨단의 성능을 달성한다.

1. 서론

열화상 분할(Thermal Segmentation)기술은 열화상 센서를 사용해서 의미적(Semantic) 분할 예측을 수행하는 기술이다. 열화상 센서는 낮은 조도, 급격한 조도 변화에 취약한 컬러 센서의 문제점을 보완하고, 다양한 환경 변화에 강인한 의미적 분할 예측이 가능하다[3],[10]-[13]. 하지만 열화상 영상은 컬러 영상에 비해 라벨링 데이터셋이 부족하므로 모델을 학습시키는데 어려움이 있다. 또한 열화상 영상의 부족한 질감 정보, 흐린 가장자리 등의 요인으로 인해 의미적 분할 예측의 성능이 저하된다.

이러한 문제를 해결하고자 MS-UDA[1]는 대용량 컬러 데이터 셋으로 사전 학습된 전문(expert) 모델 및 멀티 스펙트럴(multi-spectral) 데이터셋 속 컬러 데이터 셋으로 학습하는 준전문(semi expert) 모델을 활용하여 서로 다른 컬러 데이터셋 사이의 도메인 차이를 줄였으며, 동시에 준전문 모델로부터 멀티 스펙트럴 데이터셋 속 열화상 데이터 셋으로 학습하는 열화상 분할 모델을 학습시키는 새로운 학습 과정을 제안했다. 하지만 학습이 끝나지 않은 준전문 모델의 예측 값을 열화상 분할 모델에게 전이(transfer)를 하는 단 방향 학습 방식은 부정확한 정답을 제공할 뿐만 아니라, 모델이 열화상 영상으로부터 취득할 수 있는 유의미한 정보를 학습하기 어렵게 만든다. 또한 컬러 영상의 풍부한 정보를 열화상 영상으로 전달하는 과정에서 효과적인 도메인 적응(Domain Adaptation) 또한 해결해야 할 과제이다.

본 논문에서는 컬러 의미적 분할 모델이 열화상 의미적 분할 모델에 일방적으로 지식을 전달하는 기존의 방법과 달리, 두 모델이 함께 학습하는 상호 학습(Mutual Learning) [7]에 영감을 받아 각각의 파장대가 학습에 유의미한 정보를 서로 전이하는 학습 기법인 MML(Multi-spectral Mutual Learning)을 제안한다. 우리가 제안하는 MML기법을 통해 서로 다른 도메인이 각각의 장점을 배우는 방식으로 학습을 진행한다. 또한 우리는 텍스처가 부족한 열화상 영상에서의 의미적 분할 정확도를 향상시키기 위하여 컬러 의미적 분할 모델의 인코더에서 추출된 어텐션 지도(attention map)를 열화상 모델의 어텐션 지도에게 전이하는 새로운 목적 함수인 어텐션 손실을 제안한다. 어텐션 손실을 통해 풍부한 질감, 뚜렷한 가장자리 덕에 특징을 추출하기에 더 효과적인 영역에 가중치를 부여한 컬러 도메인의 어텐션 지도를 열화상 도메인에 전이함으로써 우수한 성능 향상을 달성한다. 결과적으로 우리는 성공적인 도메인 적응 및 지식 전이를 수행함으로써 정답 라벨이 필요 없는 비지도 학습임에도 불구하고 베이스라인인 MS-UDA의 성능을 능가하였을 뿐만 아니라, 지도학습 기반의 멀티 스펙트럴 의미적 분할 방법론들과도 유사한 성능을 보인다.

2. 관련 연구

2.1 비지도 도메인 적응

비지도 도메인 적응(UDA - Unsupervised Domain Adaptation)은 원본(source) 도메인의 라벨을 활용하여 작업별 지식(task-specific)을 전달함으로써 대상(target) 도메인의 데이터 부족을 완화하는 효율적인 방법이다. [14]는 두 도메인의 의미적 분할 출력이 공간(spatial) 레이아웃 및 로컬 컨텍스트와 같은 특정 유사성을 갖는다는 관찰을 기반으로 적대적 훈련을 채택했다. [6]은 네트워크가 명시적으로 도메인 불변(invariant) 특징 및 도메인 특정(specific) 특징을 분리함으로써 두 특징을 모두 학습하도록 했다. 동일한 도메인 간의 간격(gap)을 줄이는 이전의 많은 UDA 방법론들과는 달리 우리가 제안하는 방법론은 서로 다른 도메인(컬러, 열화상) 간의 한층 더 효과적인 도메인 적응에 중점을 둔다.

2.2 전이 학습

모델을 학습하는 새로운 패러다임으로 전이 학습(transfer learning)은 최근 점점 더 많은 관심을 받는다. 전이 학습이란 한 분야의 문제를 해결하기 위해서 얻은 지식을 다른 문제를 푸는데 사용하는 방식을 의미하며, 전이 학습을 수행하지 않은 모델들보다 비교적 빠르고 우수한 정확도를 달성한다. 우리는 서로 다른 두 도메인(컬러, 열화상)이 만든 어텐션 지도 사이의 추가적인 손실 계산을 통해 지식 전이를 성공적으로 수행한다.

3. 본론

열화상 의미 분할은 라벨링된 데이터셋이 상대적으로 부족하다는 한계점 때문에 컬러 영상 및 해당 데이터로 사전 학습된 모델을 함께 활용하는 비지도 도메인 적응 기법이 좋은 대책이 될 수 있다. 본 장에서는 두 스펙트럼 간의 도메인 적응을 보다 효과적으로 할 수 있도록 하는 양방향 지식 전이 기반의 상호 학습 방식인 MML에 대해 소개한다. 추가로, 질감 정보가 부족한 열화상 영상에서 발생할 수 있는 성능 감소를 보완하고자 상대적으로 질감 정보가 풍부한 컬러 영상으로부터 취득한 어텐션 지도를 열화상 영상에서 취득한 어텐션 지도에게 전이하는 새로운 목적 함수인 어텐션 손실을 제안한다.

3.1 사전 지식

본 논문의 방법론을 소개하기 전에 분야 및 방법론의 원활한 이해를 위해 베이스라인 방법론인 MS-UDA에 대하여 먼저 소개한다. 해당 방법론의 핵심 모델은 크게 3가지로, (1)대용량의 컬러 데이터 셋으로 학습한 전문 모델, (2)멀티 스펙트럴 데이터셋 속 컬러 데이터 셋으로 학습한 준전문 모델(컬러 스트림), (3)해당 분야의 최종 목표인 열화상 스트림으로 구성된다. 먼저 학습시키고자 하는 멀티 스펙트럴 데이터셋 속 컬러 영상을 대용량 컬러 데이터셋인 Cityscapes[2]로 사전 학습된 모델의 입력으로 주어 수도 라벨(Pseudo Label)을 생성한다. 생성된 수도 라벨을 컬러 스트림에 제공하여 학습시킴으로써 두 종류의 컬러 데이터셋 사이의 도메인 차이(e.g. Cityscapes 데이터셋 vs MF 데이터셋)를 줄인다. 이와 동시에 열화상 스트림은 컬러 스트림의 예측 값을 통해 학습함으로써 서로 다른 두 스펙트럼 사이에 도메인 차이를 줄인다.

또한 MS-UDA는 밤 컬러 영상의 낮은 조도, 급격한 조도 변화 등과 같은 요인으로 인해 전문 모델이 예측한 분할 지도의 성능 저하를 언급하며, 이를 해결하기 위해 낮 컬러 영상만을 학습 데이터로 사용하여 수도 라벨을 생성하고 컬러 스트림을 학습시킨다. 그리고 열화상 영상 또한 낮과 밤 시간대에 따른 도메인 차이(픽셀 분포값의 차이)로 인해 서로 다른 시각적인 정보를 가진다는 문제점을 언급한다. 따라서 MS-UDA는 이러한 열화상 영상 내 낮과 밤 사이의 정보 차이를 줄이기 위한 도메인 적응을 위해 CycleGAN[5] 모델을 사용하여 거짓 밤 열화상 영상을 만들어서 시간대에 따른 도메인 차이를 학습한다.

3.2 Multi-spectral Mutual Learning

기존 MS-UDA는 사전 학습된 전문 모델에서 준전문 모델인 컬러 스트림으로, 그리고 컬러 스트림에서 열화상 스트림으로 진행되는 두 단계의 단 방향 전이가 진행된다. 해당 과정에서 열화상 스트림은 컬러 스트림에게 일방적으로 지식을 전달받으며 학습을 진행한다. MS-UDA와는 달리 본 논문에서는 각 스트림에서 예측한 의미적 분할 지도에 상호 학습을 적용하여 컬러 스트림과 열화상 스트림을 협력 관계로 설정하는 상호 학습 기법인 MML(Multi-spectral Mutual Learning)을 제안한다. 또한 MML은 사전 학습된 전문 모델이 예측한 수도 라벨을 컬러와 열화상 스트림 각각에 단 방향 전이하는 방식을 채택하여 학습 초반에 각 스트림 사이의 부정확한 상호 학습을 완화하기 위한 가이드라인을 제시한다.

입력 컬러 영상과 열화상 영상을 각각 x_{rgb}와 x_{th}라고 하고, 사전 학습된 전문 모델이 예측한 수도 라벨을 y_{pseudo}라고 한다. 각 도메인 별 스트림에서 예측한 결과를 P_{rgb}, P_{th}라고 할 때, 컬러 스트림과 열화상 스트림에서 각각 예측한 P_{rgb}와 P_{th}가 서로의 유의미한 정보를 양방향으로 전이하기 위해 아래에 정의된 KL-Divergence Loss[15] L_{KL}를 사용한다.

$$L_{KL}(P_s, P_t) = \sum_{m=1}^{M} p_t \log \frac{p_t}{p_s},$$
$$where\ \{s,t\} \in [\{rgb,th\},\{th,rgb\}]. \quad \cdots \quad (1)$$

또한 우리는 수도 라벨을 컬러와 열화상 스트림 각각에 전이함으로써 부정확한 상호 학습을 완화하기

위해 아래의 Knowledge Transfer Loss L_{KT}를 새롭게 제안한다.

$$L_{KT}(P_s, y_{pseudo}) = -\frac{1}{N}\sum_{h,w}\sum_{c=1}^{C} y_{pseudo}^{(h,w,c)}\, log\left(P_s^{(h,w,c)}\right),$$
$$where\ \ s \in [rgb, th]. \quad \cdots \quad (2)$$

위 식에서 N, C는 픽셀의 수, 전체 클래스의 수를 의미하며, 우리의 최종 손실 L_{MML}은 아래와 같다.

$$L_{MML}(P_s, P_t, y_{pseudo}) = L_{KL}(P_s, P_t) + \delta L_{KT}(P_s, y_{pseudo}),$$
$$where\ \ \{s,t\} \in [\{rgb, th\}, \{th, rgb\}]. \quad \cdots \quad (3)$$

δ는 하이퍼 파라미터이다.

우리가 제안한 상호 학습 기법인 MML은 L_{MML}를 통해 컬러 스트림과 열화상 스트림이 서로의 장점을 배우고 수도 라벨과도 유사해지는 방향으로 상호 학습을 진행하며, 우수한 성능 향상을 달성한다.

3.3 어텐션 메커니즘

기존 MS-UDA에서 사용하는 합성곱 신경망(CNN)은 좁은 수용 영역(Receptive Field)으로 인해 중요한 부분에 대한 가중치 부여가 불가능하다는 한계점이 존재한다. 이를 해결하고자 본 논문에서는 MANet[8]을 기반으로 한 어텐션 메커니즘을 활용한다. 본 논문에서 사용하는 Attention Block(그림 2-(우) 참고)은 위치 어텐션을 수행하는 KAM(Kernel Attention Module)과, 채널 어텐션을 수행하는 CAM(Channel Attention Module)으로 구성되며, 이는 각각 채널 장거리 종속성(long-range dependency)과 위치 장거리 종속성을 모델링 함으로써 넓은 영역에서의 컨텍스트 정보를 반영할 뿐만 아니라 상대적으로 중요한 영역에 가중치를 부여한다. 또한 여러 크기(multi scale)에서 Attention Block을 수행함으로써 특징을 계층적으로 통합한다.

KAM C는 입력 특징 지도의 채널 수, H와 W는 각각 입력 특징 지도의 높이와 넓이로 가정한다. 입력으로 들어오는 특징 지도 $X \in R^{C\times H\times W}$로부터 1x1 필터 연산을 통해 Q(query), K(key), V(value)를 생성하고, 각각의 $Q,K,V \in R^{C\times H\times W}$ 를 $Q,K,V \in R^{C\times(H\cdot W)}$ 로 차원을 변경 한 뒤 아래 식을 통해 위치 어텐션 지도를 구한다. 그리고 X와 더해주는 과정을 통해 최종적으로 정제된 특징 지도 K_{map}을 구한다.

$$K_{map}(Q,K,V) = X + \frac{SP(Q)SP(K)^T V}{SP(Q)\sum_j SP(K)_{i,j}^T} \quad \cdots \quad (4)$$

i는 Q의 i번째 특징 q_i^T에서의 i이고, j는 K의 j번째 특징 k_j에서의 j를 의미하며, SP는 $softplus$ 함수를 나타낸다.

CAM KAM과는 달리, 입력으로 들어오는 특징 지도 $X \in R^{C\times H\times W}$에 1x1 필터 연산을 적용하지 않고 구조 변경(reshape) 과정을 통해 $Q \in R^{C\times(H\cdot W)}$, $K \in R^{(H\cdot W)\times C}$, $V \in R^{C\times(H\cdot W)}$를 생성한다. 그리고 그림1에서 볼 수 있듯이, 채널 어텐션 메커니즘을 통해 채널 어텐션 지도가 적용된 특징 지도 C_{map}을 구한다.

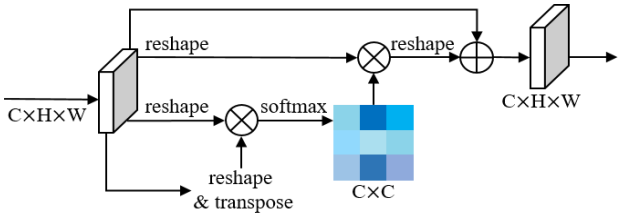

그림 1. CAM 의 동작 과정

각각 위치와 채널에 의해 정제된 위의 두 특징 지도 K_{map}과 C_{map}을 합쳐서 Attention Block의 최종 특징 지도를 구한다.

3.4 어텐션 손실

컬러 영상의 경우 풍부한 질감, 뚜렷한 가장자리 등의 정보 덕에 질감, 가장자리 정보가 부족한 열화상 영상에 비해 모델이 의미적 분할 지도를 예측하기에 수월하고, 컬러 스트림에서의 어텐션 지도 또한 특징을 추출하기에 더 효과적인 영역에 가중치를 두고 있다. 그러므로 컬러 스트림의 어텐션 지도를 열화상 스트림으로 전이하기 위해 어텐션 손실 L_{att}를 새롭게 설계한다. 어텐션 지도 사이의 손실을 계산하는 L_{att}를 적용함으로써 우리의 모델은 열화상 어텐션 지도가 컬러 어텐션 지도를 닮아가는 방향으로 학습을 진행하며, 이를 통해 더 효과적인 지식 전이가 일어날 것이라 기대한다.

컬러 영상 x_{rgb}는 E_{rgb}의 입력으로 들어가서 서로 다른 4가지 크기의 특징 지도를 출력하고, KAM과 CAM을 통해 어텐션 메커니즘을 수행하며, 이때 적용하는 어텐션 지도를 각각 $attK_{rgb_s}$와 $attC_{rgb_s}(s \in [1,2,3,4])$라고 한다. 이를 통해 넓은 영역에서의 컨텍스트 정보를 반영하면서 상대적으로 중요한 영역에 가중치까지 부여한 특징 지도를 구한다. L_{att}는 CAM에서의 어텐션 지도와 KAM에서의 어텐션 지도에 대해 각각 계산되며, 수식은 아래와 같다. L_2는 L2 loss를 의미한다.

$$L_{att}(att_{maps}) = \sum_{s=1}^{S}\left(L_2\left(attC_{th_s}, attC_{rgb_s}\right)\right.$$
$$\left.+L_2(attK_{th_s}, attK_{rgb_s})\right)\ ,where\ \ S \in [1,2,3,4] \quad \cdots \quad (5)$$

3.5 전체 방식

우리의 목표는 정답 라벨이 존재하지 않는 열화상 영상에 대한 의미적 분할을 수행하는 모델을 학습하는 것이다. 이를 수행하기 위해서, 우리 모델은 수도 라벨을 생성하는 사전 학습된 모델과 컬러 스트림, 그리고 열화상 스트림으로 구성되어 있다. MS-UDA와 동일하게 낮 컬러 영상만을 학습

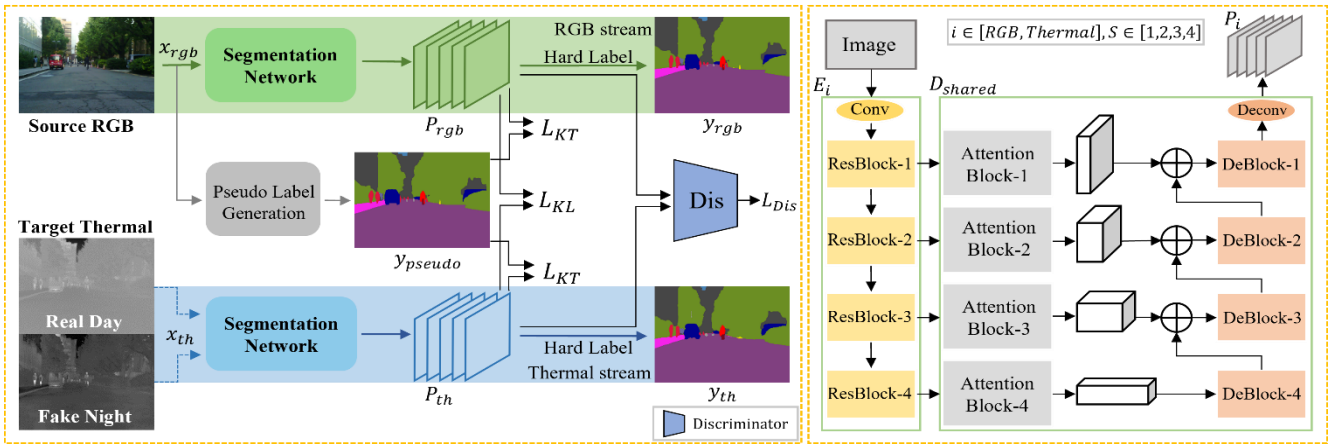

그림 2. (좌) 모델의 전체 구조 (우) 의미적 분할 네트워크(Segmentation Network)의 구조
Segmentation Network 는 입력 도메인에 따른 개별적인 인코더와, 파라미터를 공유하는 디코더로 구성된다.

데이터로 사용하고, 거짓 밤 열화상 영상을 사용하는 방식 또한 채택한다. 그림2-(좌)는 우리의 전체 구조를 나타내고, 컬러와 열화상 스트림은 각각 그림2-(우)의 의미적 분할 네트워크 구조(Segmentation Network)를 가진다.

Cityscapes 데이터셋으로 사전 학습된 HRNet[4]이 예측한 클래스 별 확률을 나타내는 의미적 분할 지도를 사용해 가장 확률이 높은 클래스 만을 선택하여 하드 라벨을 생성하고 이를 컬러 스트림과 열화상 스트림의 수도 라벨 y_{pseudo}로 채택한다. 그림2에서 보는 것과 같이 의미적 분할 네트워크의 인코더는 이미지에서 특징 지도를 추출하는 ResBlock으로 구성되고, 디코더는 어텐션 메커니즘을 통해 특징 지도를 정제하는 Attention Block과 업 샘플링을 수행하는 DeBlock으로 구성된다. ResBlock은 3x3 필터 연산을 2번 수행하며, Attention Block은 위치 어텐션을 수행하는 KAM과 채널 어텐션을 수행하는 CAM으로 구성된다. DeBlock은 간격(stride) 이 2인 3x3 디컨볼루션(deconvolution)으로 구성되며, 앞뒤가 1x1 필터 층(layer)으로 감싸져 있다. 인코더는 입력 도메인에 따라 E_{rgb}와 E_{th}로 구성하고, 디코더는 파라미터를 공유하는 동일한 디코더 D_{shared}로 구성한다. 컬러 영상 x_{rgb}는 E_{rgb}의 입력으로 들어가서 어텐션 메커니즘을 수행해서 위치와 채널에 의해 정제된 특징 지도를 구한다. 그리고 해당 특징 지도는 디코더 D_{shared} 의 입력으로 들어가서 최종적으로 컬러 스트림의 예측인 P_{rgb}를 출력한다. 열화상 영상 x_{th} 또한 위와 동일한 과정을 통해 P_{th}를 출력한다.

인코더 E_{rgb}를 학습시키기 위해 우리가 제안한 상호 학습 손실인 L_{MML}을 사용한다.

$$L_{E_{rgb}}(P_{rgb}, P_{th}, y_{pseudo}) = L_{MML}(P_{rgb}, P_{th}, y_{pseudo}). \quad \cdots \quad (6)$$

열화상 영상을 입력으로 받는 인코더 E_{th}또한 위와 마찬가지로 L_{MML}을 통해 컬러 도메인과 열화상 도메인을 협력 관계로 설정하여 효과적으로 학습한다. 그리고 E_{rgb}와 서로 유사한 임베딩 특징을 추출하는 방향으로 학습하기 위해 추가적으로 적대적 손실 L_{adv}[9]을 사용한다.

$$L_{E_{th}}(P_{rgb}, P_{th}, y_{pseudo}) = L_{MML}(P_{th}, P_{rgb}, y_{pseudo}) + \gamma L_{adv}(P_{th}) \quad \cdots \quad (7)$$

$$L_{adv}(P_{th}) = -\frac{1}{N}\sum_{h,w} log\,\psi(P_{th}^{(h,w)}). \quad \cdots \quad (8)$$

$\psi(\cdot)$는 판별자(discriminator) 함수를 의미하고, γ는 하이퍼 파라미터이다.

E_{rgb}와 E_{th}는 동일한 임베딩 특징을 추출하는 방향으로 학습이 진행되므로 디코더 D_{shared}는 입력 도메인에 관계없이 동일한 파라미터를 가지면서 픽셀별 클래스 분류 결과를 예측한다. D_{shared}를 학습할 때는 L_{MML}뿐만 아니라 어텐션 손실 L_{att}또한 적용한다. 이를 통해 분할 지도 예측을 위한 정보가 부족한 열화상 영상의 단점을 극복함과 동시에 효과적인 지식 전이를 수행한다. D_{shared}를 학습하기 위한 손실 함수 $L_{D_{shared}}$는 아래와 같다. α, β, γ 는 하이퍼 파라미터이다.

$$L_{D_{shared}}(P_{rgb}, P_{th}, P_{pseudo}) = \alpha L_{MML}(P_{rgb}, P_{th}, y_{pseudo}) + \beta L_{MML}(P_{th}, P_{rgb}, y_{pseudo}) + \gamma L_{adv}(P_{th}) + L_{att}(att_{maps}) \quad \cdots \quad (9)$$

마지막으로 판별자 Dis는 의미적 분할 예측 P가 컬러 또는 열화상 스트림 중 어느곳에서 생성되는지 여부를 구별하기 위한 모듈이다. 컬러, 열화상 스트림과 판별자의 경쟁을 통해 열화상 스트림의 예측이 컬러 스트림의 예측과 유사한 방향으로 나아가도록 모델이 학습한다. 판별자를 학습하기 위한 손실 함수 L_{dis}는 아래와 같다.

$$L_{Dis}(P) = -\frac{1}{N}\sum_{h,w}\{(1-z)log\psi(P^{(h,w)}) + zlog\psi(P^{(h,w)})\} \quad \cdots \quad (10)$$

(a) RGB Image (b) Thermal Image (c) Ground-Truth (d) MS-UDA (e) Ours

그림 3. 의미적 분할 결과 시각화

P가 컬러 도메인의 예측일 때 z는 0 이고, 열화상 도메인의 예측일 때 z 는 1 이다.

4. 실험

4.1 구현 정보

우리 실험은 Pytorch를 기반으로 진행하고, 학습을 위해 배치 사이즈(batch size)는 8로 설정하며 300에포크(epoch)의 학습을 수행한다. 분할 네트워크를 학습하기 위해 0.002의 학습률(learning rate), 0.9의 모멘텀(momentum), 0.0005의 가중치 감소(weight decay) 를 가지는 SGD 옵티마이저(optimizer)를 사용하였고, 판별자의 경우 0.002의 학습률을 가지 는 Adam 옵티마이저를 사용하였다. 하이퍼 파라미터 α, β, γ는 각각 1, 0.25, 0.01로 설정하고, 상호 학습 수행에 적용하는 δ는 초기단계에 0.2로 설정 후 에포크에 진행에 따라 1까지 선형적으로 증가시킨다.

학습 단계에서는 멀티 스펙트럴 이미지 쌍이 컬러 스트림과 열화상 스트림에 동시에 들어가고, 추론 단계에서는 열화상 이미지만을 열화상 스트림 모델의 입력으로 활용하여 최종적인 의미적 분할 지도를 예측한다.

4.2 데이터셋 및 평가지표

데이터셋 우리의 모든 실험은 MF 데이터셋[3]에서 진행된다. MF 데이터셋은 9가지의 클래스로 의미적 분할 정답이 라벨링된 640×480 크기인 1569장의 영상으로 구성되어 있다. 우리는 학습 시간 때 낮 영상만을 사용하기 때문에, MS-UDA와 마찬가지로 기존의 MF 데이터셋 중 낮 영상을 학습 영상으로 사용한다.

평가 지표 우리의 모든 실험은 MF 데이터셋에서 진행된다. 수도 라벨을 생성하는 HRNet은 19가지의 클래스를 가지는 Cityscapes 데이터셋으로 사전 학습한다. 우리는 MF 데이터셋과 Cityscapes 데이터셋 중 겹치는 3가지 클래스 (자동차, 사람, 자전거) 각각에 대한 IOU의 평균인 mIOU를 픽셀 단계에서 측정한다. 각 표에서 빨간색 영역으로 표시된 mIOU는 픽셀 단계에서의 클래스 분류 정확도를 나타내는 수치이므로 높을수록 좋은 성능을 의미한다.

표 1. 각 스트림 별 의미적 분할 결과

방법론	스트림	차	사람	자전거	mIOU
MS-UDA	컬러	94.9	76.3	81.6	83.3
	열화상	86.4	59.3	57.7	**67.8**
Ours (w.o L_{att}, L_{KL})	컬러	96.6	80.3	78.7	**85.2**
	열화상	73.4	67.4	42.4	61.1

표 2. MF 데이터셋에서 의미적 분할 결과

	Method	차	사람	자전거	mIOU
Sup	MFNet[3]	65.9	58.9	42.9	55.9
	RTFNet[10]	86.3	67.8	58.2	70.7
	FuseSeg[11]	87.9	71.7	**64.6**	74.7
	CMX[12]	90.1	75.2	64.5	76.7
UDA	HeatNet[13]	56.4	68.8	33.9	53.0
	MS-UDA	86.4	59.3	57.7	67.8
	Ours	**91.2**	**79.1**	64.2	**78.2**

4.3 실험 결과 및 분석

본 논문에서 베이스라인으로 설정한 MS-UDA의 분할 네트워크는 합성곱 연산만으로 이루어진 인코더-디코더 구조를 가진다. 해당 모델은 컬러 스트림의 예측을 열화상 스트림으로 단 방향 전이를 통해 열화상 스트림이 컬러 스트림의 예측을 최대한 모방하는 방향으로 학습을 진행한다. MS-UDA에서 컬러 스트림과 열화상 스트림에 대한 의미적 분할 성능은 각각 83.3%, 67.8% 로 표 1에서 확인할 수 있다. 반면 L_{att}와 L_{KL}의 적용 없이 어텐션 메커니즘만을 적용한 우리 모델의 경우 컬러 스트림과 열화상 스트림에서의 성능은 각각 85.2%, 61.1% 이다. 합성곱 연산에 비해 넓은 영역에서의 컨텍스트 정보를 반영하고 중요한 영역에 가중치를 부여하는 어텐션 메커니즘을 통해 컬러 스트림에서의 성능이 기존 대비 1.9% 향상을 보인다. 하지만 열화상 스트림의 경우 기존 대비 6.7% 하락이 발생한다. 우리는 컬러에서 열화상 스트림으로 전이가 일어날 때 성

표 3. 각 손실 적용에 따른 검증 실험

	L_{att}	L_{KL}	차	사람	자전거	mIOU
Ours	-	-	73.4	67.4	42.4	61.1
	✓	-	81.7	65.7	52.3	66.6
	-	✓	87.8	67.9	55.6	70.4
	✓	✓	**91.2**	**79.1**	**64.2**	**78.2**

능 하락이 발생한 이유를 열화상 스트림에 적용되는 어텐션 메커니즘이 혼자서는 열화상 도메인의 특징으로부터 중요한 영역에 가중치를 부여하지 못하고, 둘 사이의 도메인 적응이 효과적이지 않다고 가정한다. 이를 해결하고자 두 도메인에서 적용되는 어텐션 지도 사이의 손실 L_{att}를 적용해서 풍부한 정보(e.g. 선명한 색상, 뚜렷한 가장자리, 질감)를 기반으로 학습하는 컬러 스트림이 집중해서 보는 영역을 열화상 스트림에 부여하는 과정을 통해 열화상 스트림에서의 성능 향상을 일으킨다. 또한 L_{KL}을 사용한 MML의 적용을 통해 학습 초반에 각 스트림 사이의 부정확한 상호 학습을 완화하기 위한 가이드라인을 제시하며, 두 도메인의 예측이 각각의 장점을 배우는 방향으로 학습하도록 한다.

우리가 제안한 학습 방식의 효과는 표 2에서 입증한다. 우리가 제안한 방식은 모든 클래스에 대해 MS-UDA를 능가하고, 정답 라벨을 사용하는 다른 지도 학습 방법론들의 성능과 유사하거나 이를 넘어선다. 또한 우리는 열화상 영상만을 사용함에도 불구하고 컬러-열화상 융합을 사용하는 기존의 비지도 도메인 적응(UDA) 방식인 HeatNet[13]과 비교해서도 큰 폭의 성능 향상을 보인다. 이를 통해 우리가 제안한 모델은 어텐션 지도 사이의 손실을 사용하여 상대적으로 풍부한 정보를 담고 있는 컬러 영상이 가중치를 부여하는 영역에 초점을 두고 학습할 뿐만 아니라, 예측된 의미적 분할 지도 사이의 상호 학습을 통해 도메인 차이가 존재하는 영상에서 성공적인 도메인 적응을 수행한다.

그림 3은 MF 데이터셋으로 의미적 분할 예측을 시각화 한 결과이다. MS-UDA와 비교할 때 더 정확하고 선명한 가장자리를 가진 예측을 수행한다. 낮의 경우 자전거를 타고 있는 사람을 성공적으로 예측하고, 밤의 경우 사람과 자동차의 형태를 더 정확하게 예측한다. 각각 빨간색-자동차, 파란색-사람, 갈색-자전거를 나타낸다.

4.4 검증 실험

표 3은 우리가 효과적인 도메인 적응과 지식 전이를 위해 적용한 손실 L_{att}, L_{KL}에 대한 검증 실험(Ablation study)이다. 우리가 적용한 손실을 통해 성공적인 도메인 적응과 지식 전이가 수행되고, 특히 L_{KL}의 적용에 대한 향상 폭이 큰 것을 볼 수 있다.

5. 결론

본 논문에서 우리는 비지도 학습에서 도메인 적응을 효과적으로 수행함으로써 성공적으로 의미적 분할 예측을 수행한다. 우리가 적용한 상호 학습과 어텐션 지도 사이의 손실을 통해 서로 다른 두 도메인 사이에서 효과적으로 지식을 전이할 수 있다. 실험으로 우리 모델의 효과적인 학습 방식에 대해 입증한다. 우리의 연구는 컬러-열화상 뿐만 아니라 임의의 서로 다른 두 도메인에도 적용 가능할 것이라고 기대한다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었으며 (IITP-2022-RS-2022-00156345). 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2020M3F6A1109603, NRF-2020R1F1A1076987)

참고문헌

[1] Y. Kim, et al., “MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation,” In RAL, 6(4):6497-6504, 2021.

[2] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” In CVPR, 2017.

[3] Q. Ha et al., “Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes,” In IROS, 2017.

[4] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” In TPAMI, 43(10):3349-3364, 2020.

[5] J. Y. Zhu et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” In ICCV, 2017.

[6] W. L. Chang et al., “All about structure: Adapting structural information across domains for boosting semantic segmentation,” In CVPR, 2019.

[7] Z. Ying et al., “Deep Mutual Learning,” In CVPR, 2018.

[8] R. Li et al., “Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images,” In TGRS, 60:1-13, 2021.

[9] I. Goodfellow et al., “Generative adversarial networks,” In ACM, 63(11):139-144, 2020.

[10] Y. Sun et al., “Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes,” In RAL, 4(3):2576-2583, 2019.

[11] Y. Sun et al., “Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion,” In TASE, 18(3):1000-1011, 2020.

[12] H. Liu et al., “CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers,” In *arXiv:2203. 04838,* 2022.

[13] J. Vertens et al., “Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images,” In IROS, 2020.

[14] Y. H. Tsai et al., “Learning to adapt structured output space for semantic segmentation,” In CVPR, 2018

[15] D. Hendrycks et al., “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” In *arXiv:1601. 02136,* 2016.