# End-to-End Multi-Task Learning with Attention

태그

## 1. Introduction

### Multi-Task Learning(MTL)

- 지금까지의 CNN들은 보통 하나의 task 달성을 목표로 함: real-world에서는 여러개의 task들을 동시에 처리할 수 있어야 함

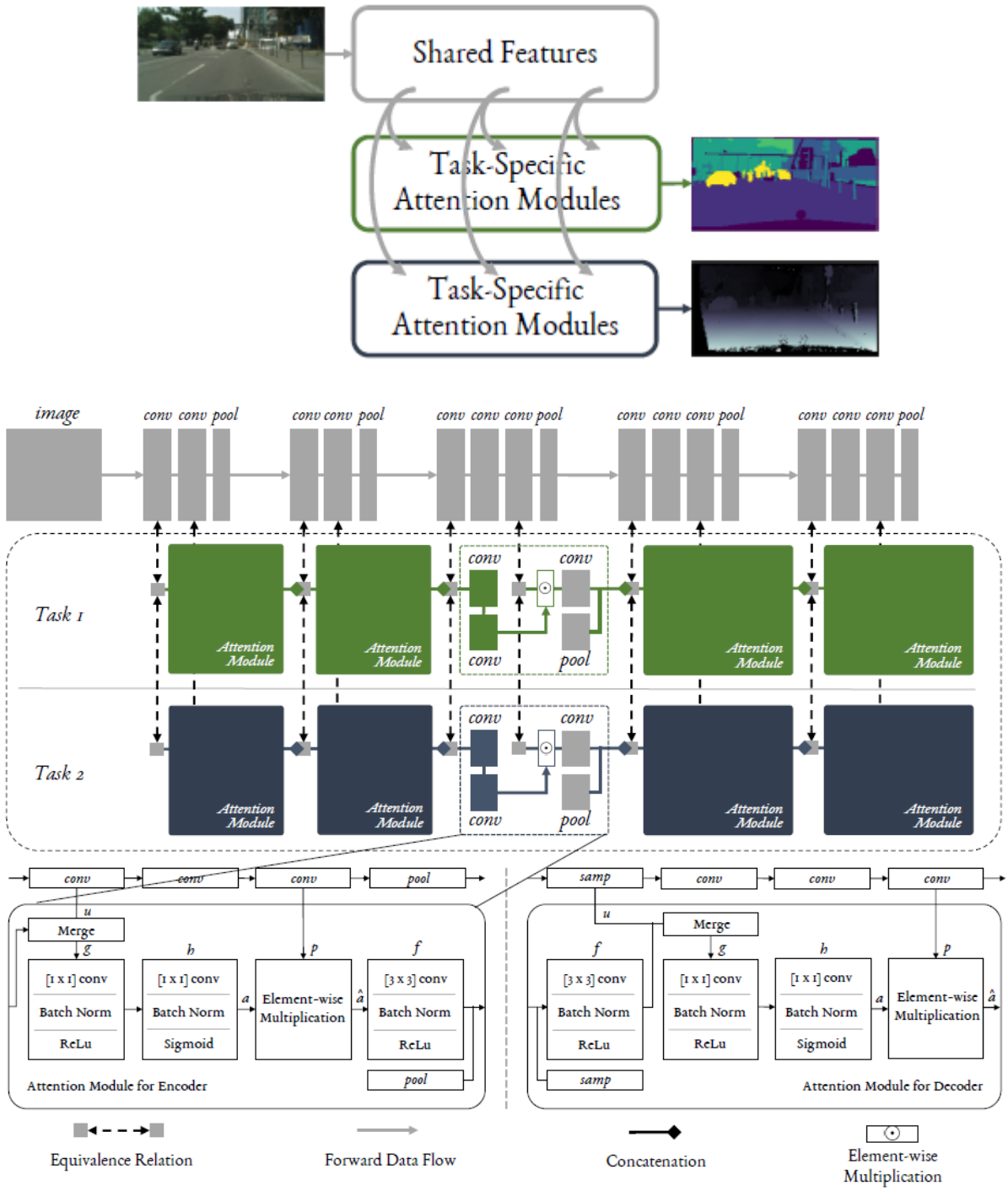
### Key Challenges

1. 네트워크 아키텍처
    - 네트워크 간 각 task가 공유되어야 하고 task별로 특정한 feature 역시 분류할 수 있어야 함
    - over-fitting 방지를 위해 포괄적이어야 함
    - 그와 동시에 under-fitting 방지를 위해 각 task마다 feaature를 학습할 수 있어야 함
2. loss function
    - 각 task들을 동일한 정확도로 학습해야 함
    - 쉬운 일부 task가 나머지를 우세해서는 안됨

- 지금까지의 MTL 접근법은 이 두 가지 중 하나에만 초점을 뒀음

## 3. Multi-Task Attention Network

### 3.1. Architecture Design




- single shared network와 $K$ task-specific attention network로 구성
- encoder-decoder network인 SegNet을 기반으로 구성
- VGG-16에 기반한 SetNet의 encoder와 이에 대칭인 decoder로 구성
- shared network를 통해 여러 task에 대한 공유된 feature의 일반화를 극대화하는 동시에 soft attention mask를 통해 task별 성능 역시 향상시킬 수 있음

# 4. Experiments

## 4.1. Image-to-Image Prediction (One-to-Many)

### 4.1.1. Datasets

- CityScapes
  - 높은 해상도의 거리 이미지 데이터셋
  - task
    1. semantic segmentation
    2. depth estimation
- NYUv2
  - RGB-D의 실내 장면 데이터셋
  - task
    1. semantic segmentation
    2. depth estimation
    3. surface normal

### 4.1.2. Baselines

- **Single-Task, One Task**: vanilla SegNet
- **Single-Task, STAN**: Single-Task Attention Network
- **Multi-Task, Split**: standard multi-task learning. 마지막 layer에서 각 task에 대한 최종 예측값을 분류
  - Wide: convolution feature의 개수 변화
  - Deep: convolution layer의 개수 변화
- **Multi-Task, Dense**: attention module 없이 각 task-specific network들이 서로 하나의 shared network를 구성
- **Multi-Task, Cross-Stitch**: Cross-Stich Network

### 4.1.3. Dynamic Weight Average

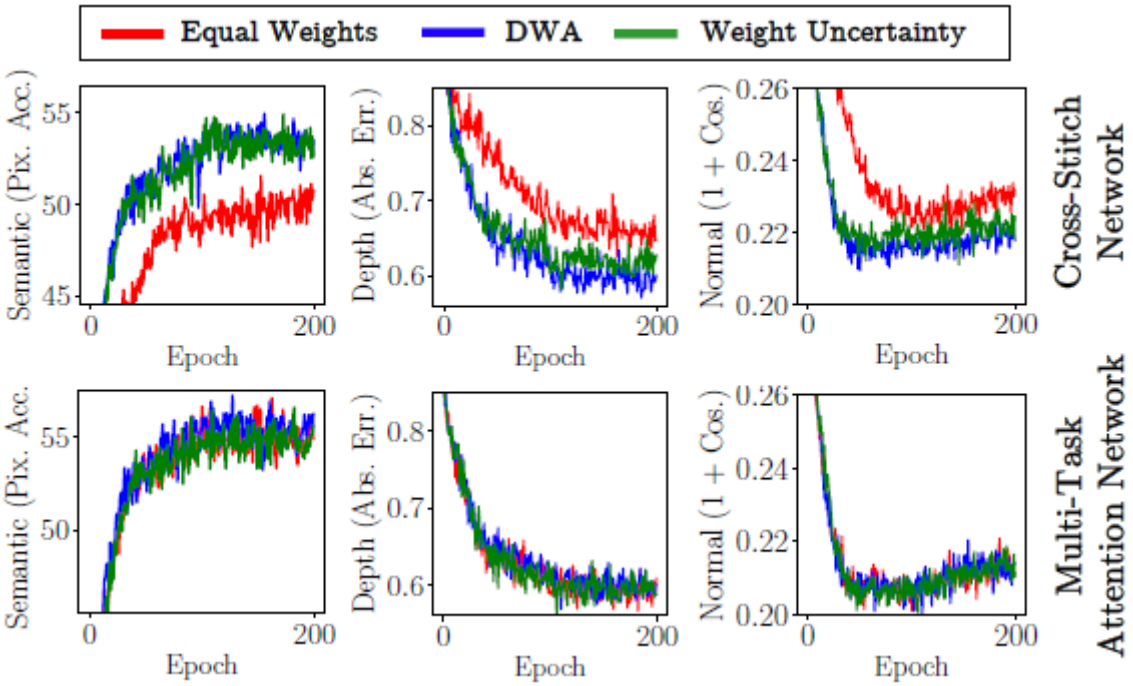
- 시간에 따른 각 task 별 loss 값의 변화율에 따라 task weighting

$$\lambda_k(t) := \frac{K \exp(w_k(t-1)/T)}{\sum_i \exp(w_i(t-1)/T)}, w_k(t-1) = \frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)}, \tag{7}$$

### 4.1.4. Results on Image-to-Image Predictions

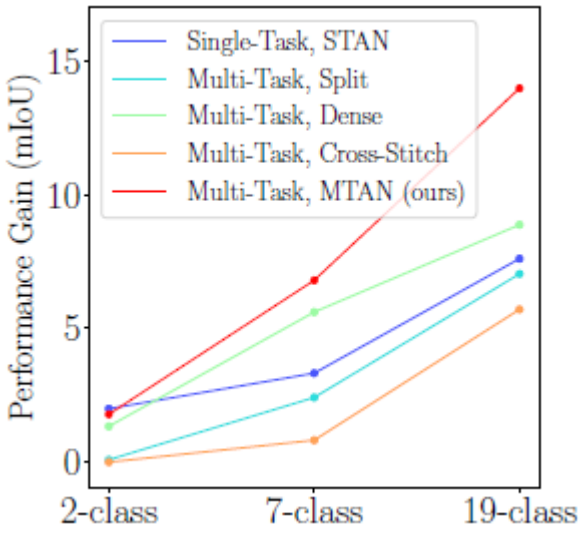
| Type | #P. | Architecture | Weighting | Segmentation (Higher Better) | | Depth (Lower Better) | | Surface Normal: Angle Distance (Lower Better) | | Surface Normal: Within $t^\circ$ (Higher Better) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mIoU | Pix Acc | Abs Err | Rel Err | Mean | Median | 11.25 | 22.5 | 30 |
| Single Task | 3 | One Task | n.a. | 15.10 | 51.54 | 0.7508 | 0.3266 | 31.76 | 25.51 | 22.12 | 45.33 | 57.13 |
| | 4.56 | STAN | n.a. | 15.73 | 52.89 | 0.6935 | 0.2891 | 32.09 | 26.32 | 21.49 | 44.38 | 56.51 |
| Multi Task | 1.75 | Split, Wide | Equal Weights | 15.89 | 51.19 | 0.6494 | 0.2804 | 33.69 | 28.91 | 18.54 | 39.91 | 52.02 |
| | | | Uncert. Weights [14] | 15.86 | 51.12 | **0.6040** | 0.2570 | **32.33** | **26.62** | **21.68** | **43.59** | **55.36** |
| | | | DWA, $T = 2$ | **16.92** | **53.72** | 0.6125 | **0.2546** | 32.34 | 27.10 | 20.69 | 42.73 | 54.74 |
| | 2 | Split, Deep | Equal Weights | 13.03 | 41.47 | 0.7836 | 0.3326 | 38.28 | 36.55 | 9.50 | 27.11 | 39.63 |
| | | | Uncert. Weights [14] | **14.53** | 43.69 | 0.7705 | 0.3340 | **35.14** | **32.13** | **14.69** | **34.52** | **46.94** |
| | | | DWA, $T = 2$ | 13.63 | **44.41** | **0.7581** | **0.3227** | 36.41 | 34.12 | 12.82 | 31.12 | 43.48 |
| | 4.95 | Dense | Equal Weights | 16.06 | 52.73 | 0.6488 | 0.2871 | 33.58 | 28.01 | 20.07 | 41.50 | 53.35 |
| | | | Uncert. Weights [14] | **16.48** | **54.40** | 0.6282 | 0.2761 | **31.68** | **25.68** | **21.73** | **44.58** | **56.65** |
| | | | DWA, $T = 2$ | 16.15 | 54.35 | **0.6059** | **0.2593** | 32.44 | 27.40 | 20.53 | 42.76 | 54.27 |
| | ≈3 | Cross-Stitch [20] | Equal Weights | 14.71 | 50.23 | 0.6481 | 0.2871 | 33.56 | 28.58 | 20.08 | 40.54 | 51.97 |
| | | | Uncert. Weights [14] | 15.69 | 52.60 | 0.6277 | 0.2702 | 32.69 | 27.26 | 21.63 | 42.84 | 54.45 |
| | | | DWA, $T = 2$ | **16.11** | **53.19** | **0.5922** | **0.2611** | **32.34** | **26.91** | **21.81** | **43.14** | **54.92** |
| | 1.77 | MTAN (Ours) | Equal Weights | **17.72** | 55.32 | **0.5906** | 0.2577 | 31.44 | **25.37** | **23.17** | 45.65 | 57.48 |
| | | | Uncert. Weights [14] | 17.67 | **55.61** | 0.5927 | 0.2592 | **31.25** | 25.57 | 22.99 | **45.83** | **57.67** |
| | | | DWA, $T = 2$ | 17.15 | 54.97 | 0.5956 | **0.2569** | 31.60 | 25.46 | 22.48 | 44.86 | 57.24 |

1. 파라미터의 추가 없이 또는 적은 개수의 파라미터 만으로도 우수한 성능 유지
2. 다양한 loss function weighting schmem들에 대해서도 높은 성능을 유지




- Cross-Stich Network보다 MTAN에서 다양한 weighting scheme에 대해서 일정한 학습 추세를 보임

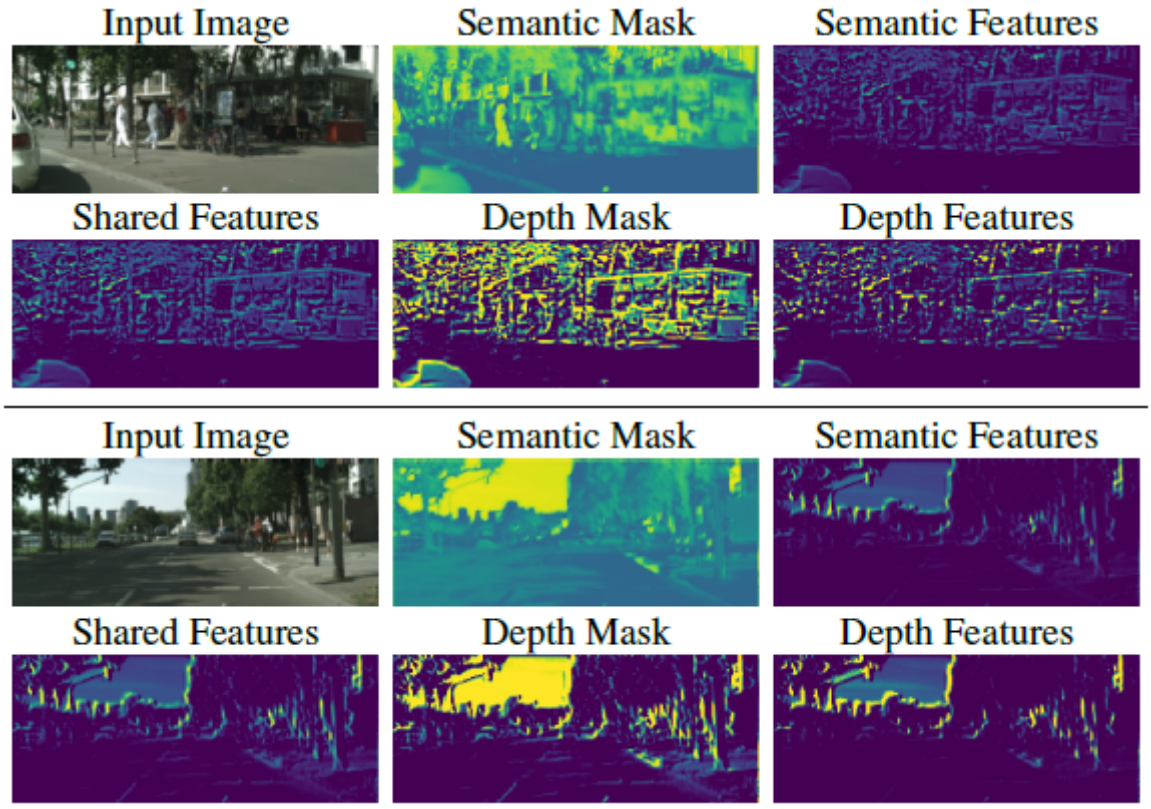
### 4.1.5. Effect of Task Complexity




- 2-class에서는 Single-Task, STAN이 다른 모든 multi-task network들을 능가
  - 간단한 방식으로 네트워크의 파라미터를 완전히 활용 가능하기 때문
- task가 더 복잡해짐에 따라 모든 network에서 성능이 향상됨

- MTAN에서는 더 높은 속도로 성능이 향상됨

### 4.1.6. Attention Masks as Feature Selectors




- attention mask 적용 후 두 task 간 분명한 차이점 확인 가능
- 공유된 feature에서 uninformative한 부분을 가리고 각 task에서 유용한 부분에 초점을 잡는 데 도움을 줌

## 4.2. Visual Decathlon Challenge (ManytoMany)

| Method | #P. | ImNet. | Airc. | C100 | DPed | DTD | GTSR | Flwr | Oglt | SVHN | UCF | Mean | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scratch [23] | 10 | 59.87 | 57.10 | 75.73 | 91.20 | 37.77 | 96.55 | 56.3 | 88.74 | 96.63 | 43.27 | 70.32 | 1625 |
| Finetune [23] | 10 | 59.87 | 60.34 | 82.12 | 92.82 | 55.53 | 97.53 | 81.41 | 87.69 | 96.55 | 51.20 | 76.51 | 2500 |
| Feature [23] | 1 | 59.67 | 23.31 | 63.11 | 80.33 | 45.37 | 68.16 | 73.69 | 58.79 | 43.54 | 26.8 | 54.28 | 544 |
| Res. Adapt.[23] | 2 | 59.67 | 56.68 | 81.20 | 93.88 | 50.85 | 97.05 | 66.24 | 89.62 | 96.13 | 47.45 | 73.88 | 2118 |
| DAN [25] | 2.17 | 57.74 | 64.12 | 80.07 | 91.30 | 56.54 | 98.46 | 86.05 | 89.67 | 96.77 | 49.38 | 77.01 | 2851 |
| Piggyback [19] | 1.28 | 57.69 | 65.29 | 79.87 | 96.99 | 57.45 | 97.27 | 79.09 | 87.63 | 97.24 | 47.48 | 76.60 | 2838 |
| Parallel SVD [24] | 1.5 | 60.32 | 66.04 | 81.86 | 94.23 | 57.82 | 99.24 | 85.74 | 89.25 | 96.62 | 52.50 | 78.36 | 3398 |
| MTAN (Ours) | 1.74 | 63.90 | 61.81 | 81.59 | 91.63 | 56.44 | 98.80 | 81.04 | 89.83 | 96.88 | 50.63 | 77.25 | 2941 |

- MTAN은 대부분의 baseline을 능가함
- DropOut의 적용이나 dataset의 재그룹화, 적응형 가중치 감소 등 또한 복잡한 정규화 전략 없이 최신 모델들과 경쟁력 있는 성능을 보임

# 5. Conclusions

- E2E 방식으로 task-shared feature들과 task-specific feature들을 자동적으로 학습할 수 있음
- MTAN은 다른 모델과 경쟁력 있는 모습을 보였으며 다양한 손실 함수에 대해 견고한 성능을 보임
- attention mask를 통해 가중치를 공유하는 동시에 높은 파라미터 효율성을 유지함