딥러닝 기반 군중 집계 기술 동향

Technology Trends on Crowd Counting Based on Deep Learning

목차

안덕호 (D.H. An, dhan2006@naver.com) 세종대학교 컴퓨터공학과

Abstract

군중 계수 및 군중 밀도 추정 방법은 공공 보안 분야에서 매우 중요하다. 단일 이미지 또는 비디오 프레임에서 군중 밀도를 추정하고 계산하는 것은 다양한 시나리오에서 컴퓨터 비전 시스템의 필수 부분이 되었다. 군중 계산 및 밀도 추정에 대한 최근 연구 기술을 확인하고 종합적으로 검토한다. 먼저 군중 계산의 배경 및 다양한 설명을 진행 후 군중 집계 기술에 대한 소개를 진행한다. 마지막으로 군중 집계 및 밀도의 유망한 미래 방향을 제시한다.

I. 서론

군중 계산(Crowd Counting)은 객체 계산(Object Counting)의 일부에 해당하며, 객체 계산은 단일 이미지 또는 비디오에서 개체 인스턴스의 수를 계산하는 것으로 이미지의 개체 수를 계산하는 방법이다. 감시, 미생물학, 혼잡도 추정, 제품 계수 및 교통 흐름 모니터링 등 많은 부분에서 응용이 가능하며, 객체 계산 기술의 유형으로 감지 기반 객체 계산과 회의 기반 객체 계산 방법으로 구성되어 있다.

군중 계산은 이미지로부터 사람의 수를 계산하거나 추정하는 기술이다. 물체 감지(Object Detection)과 달리 희박하고 어수선한 장면을 포함한 다양한 상황에서 임의의 크기의 표적을 동시에 인식하는 것을 목표로 하고 있다. 단일 이미지에서 객체의 정확한 수를 추정하는 것은 어려울 수 있으나, 의미가 있으며 사회 보장 및 개발에 대한 중요성으로 인해 도시 계획 및 공공 안전과 같은 많은 응용 분야에 적용되고 있다.[1]

군중 계산은 VisDrone(Vision Meets Drones) [2]과 NWPU-Crowd Counting[3]과 같은 대회를 통해 기술 개발 및 성능 평가가 이루어져 왔다. VisDrone 의 경우 RGBT 이미지 쌍이 있는 데이터 세트를 제공하고, 평균절대오차 및 평균제공오차 등의 평가 지수를 통해 검증하는 대회이며, NWPU-Crowd Counting 대회는 The NWPU-Crowd Dataset 을 사용하여 모델을 비교하는 대회이다. The NWPU-Crowd Dataset 은 5,109 개의 이미지로 구성되어 있다.

다양한 대회를 통해 군중 계산 분야의 연구는 발전하고 있으며, 2 절에서는 발전된 기술 동향 및 발전 방향을 설명하여 다양한 군중 계산 방법을 검토한다.

II. 군중 계산 기술 동향

군중 계산(Crowd Counting)은 이미지로 사람을 세는 작업으로 ShanghaiTech, QNRF 등의 군중 이미지를 통해 사람의 수를 계산하거나 추정한다. 군중 계산을 사용했던 다양한 기술 동향을 설명한다.

A. *DataSet*

군중 계산에서 사용하는 일반적인 데이터셋의 정보는 표 1 과 같다.

TABLE I. DATASET STATISTICS

데이터	년도	속성	건수
RGBT-CC	2021	혼잡	2,030
NWPU-Crowd	2020	혼잡, 현지화	5,109
JHU-CROWD++	2020	혼잡	4,372
UCF-QNRF	2018	혼잡	1,535
ShanghaiTech	2016	혼잡	482
UCF_CC_50	203	혼잡	50

B. *URC*

각 이미지의 부분 주석만 훈련 데이터로 사용하는 군중 계산 방법이다. 주석이 있는 영역과 주석이 없는 영역의 반복적인 패턴과 그 사이의 패턴에서 영감을 받아 주석이 없는 영역을 처리하기 위해 세 가지 구성요소가 있는 네트워크를 설계하였다.

먼저 URC(Unannotated Regions Characterization) 모듈에서 메모리 뱅크를 사용하여 주석이 추가된 기능만 저장하고 주석이 추가된 영역에서 추출된 시각적 기능이 주석이 없는 영역으로 이동하는 데 도움이 될 수 있다. 두 번째로 각 이미지에 대해 FDC(Feature Distribution Consistency)는 주석이 달린 머리 부분과 주석이 없는 머리 부분의 특징 분포가 일정하도록 정규화를 하였다. 마지막으로 교차 회귀자 일관성 정규화(CCR) 모듈은 주석이 없는 영역의 시각적 기능을 자체 감독 스타일로 학습하도록 설계하였다.

URC 의 모델 구조는 아래의 그림 1 을 통해 확인할 수 있다.

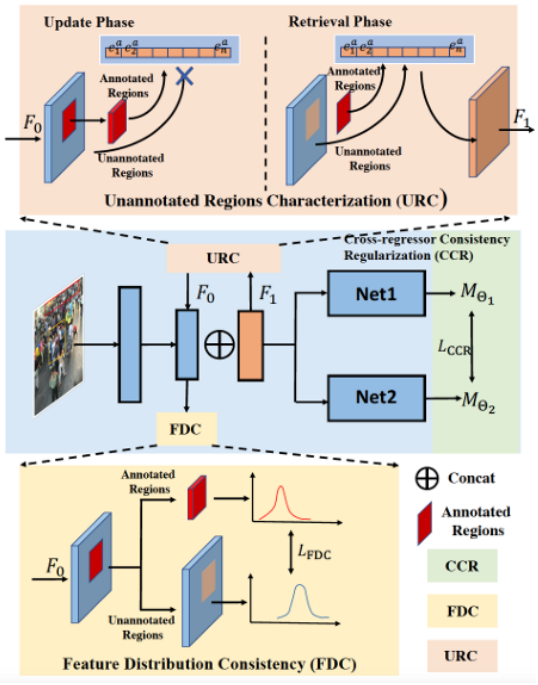

Fig. 1. URC 모델 구조

[출처] Xu, Y., Zhong, Z., Lian, D., Li, J., Li, Z., Xu, X., & Gao, S. (2021). Crowd Counting With Partial Annotations in an Image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15570-15579)

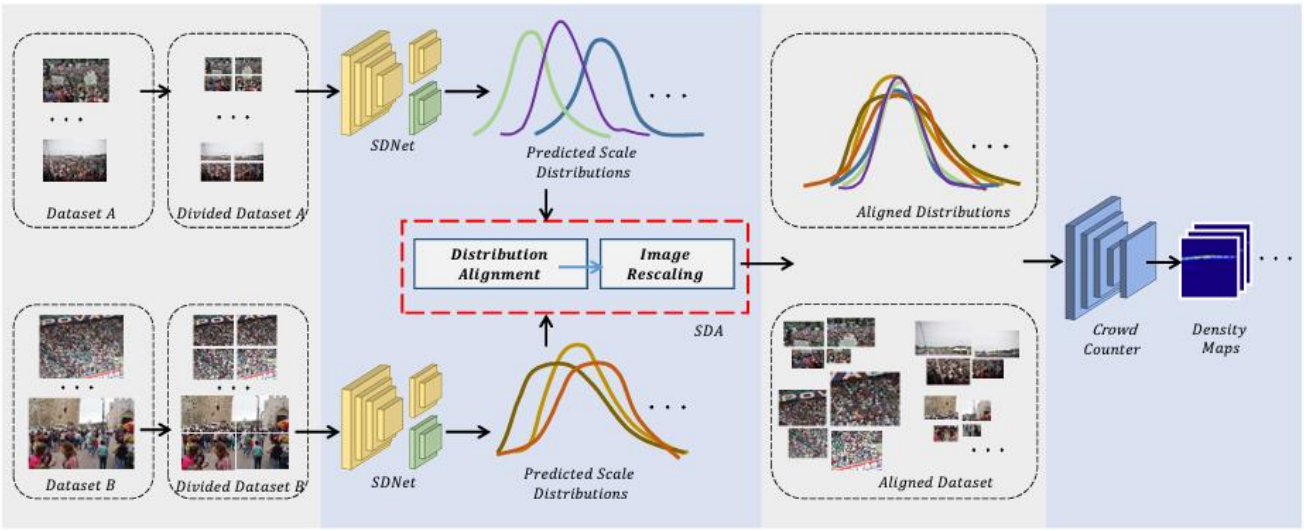

Fig. 2. SDNet 의 구조도

[1] [출처] Ma, Z., Hong, X., Wei, X., Qiu, Y., & Gong, Y. (2021). Towards a Universal Model for Cross-Dataset Crowd Counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3205-3214)

C. SDNet

장면과 데이터 세트에 걸쳐 군중 계산을 위한 보편적인 모델을 학습하는 실제 문제를 처리할 것을 제안하였다.

다른 장면 레이아웃 및 이미지 해상도와 같은 요인으로 인해 발생하는 규모 이동에 대해 군중 계산의 치명적인 요소로 보았으며, 다양한 장면에 적용할 수 있는 보편적인 모델을 훈련시키기 어렵다고 판단하여 주요 모듈로 규모 정렬하는 방법을 설계하였다. 스케일 분포 사이의 거리를 최소화하여 정렬을 위한 최적의 이미지 크기 조정 계수를 얻기 위한 폐쇄형 솔루션을 도출하였다. 또한 스케일 분포 추정을 위해 효율적인 슬라이스된 Wasserstein 거리에 기반한 손실 함수와 함께 새로운 신경망도 추가로 제안하였다.

각 데이터 세트에 대해 특별히 미세 조정된 최첨단 모델보다 성능이 우수한 여러 데이터 세트에서 일반적으로 잘 작동하는 범용 모델을 구축할 수 있었으며 보이지 않는 장면에 대한 부분에서 기존 모델보다 더 괜찮은 가능성을 보였다.

그림 2 는 SDNet 의 구조를 나타낸다.

D. UEPNet

군중 계산에서 부정확한 학습 목표 문제가 주목받고 있으며 해당 모델은 카운트 값 자체 대신 미리 정의된 카운트 간격 구간의 인덱스를 예측하는 방법의 카운트 오류 기여도가 매우 불균형하여 카운트 성능이 저하되는 위험을 최소화하기 위해 모든 간격에 대해 예상되는 계산 오류 기여도를 항상 동일하게 유지하는 UEP(Uniform Error Partition)라는 새로운 카운트 간격 분할 기준을 제안하였다. 또한 카운트 양자화 프로세스에서 불가피하게 도입된 이산화 오류를 완화하기 위해 MCP(Mean Count Proxies)라는 또 다른 기준을 추가로 제안하였다. 해당 모델의 경우 MCP 기준은 추론하는 동안 카운트 값을 나타내기 위해 각 간격에 대한 최상의 카운트 프록시를 선택하므로 이미지의 전체 예상 이산화 오류를 거의 무시할 수 있다.

UEPNet (Uniform Error Partition Network)은 위의 두 제안을 이용하여 설계된 모델로 해당 모델은 다양한 까다로운 데이터세트에도 최적화된 성능을 보이고 있다.

그림 3 은 UEPNet 의 구조를 나타낸다.

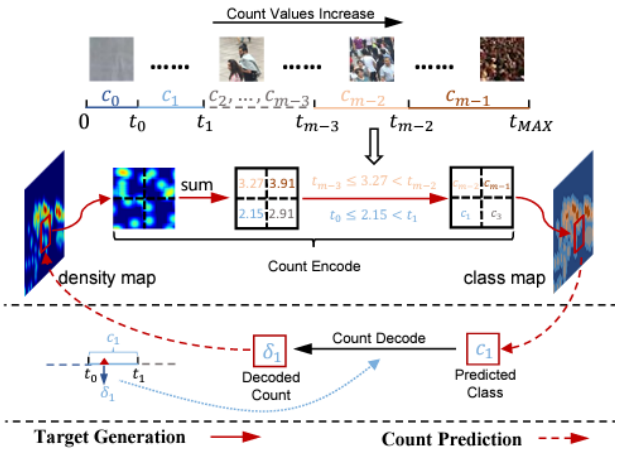

Fig. 3. UEPNet 모델 구조

[출처] Wang, C., Song, Q., Zhang, B., Wang, Y., Tai, Y., Hu, X., ... & Wu, Y. (2021). Uniformity in Heterogeneity: Diving Deep into Count Interval Partition for Crowd Counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3234-3242)

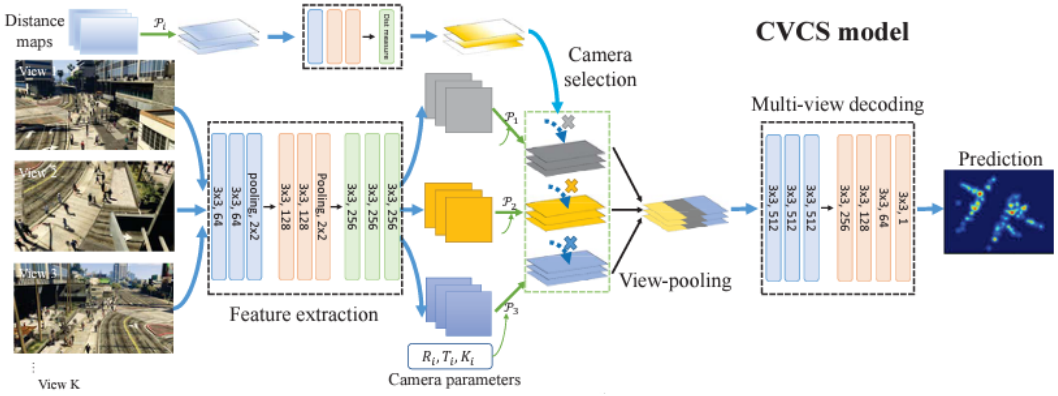

Fig. 5. CVCS 모델 구조

[출처] Zhang, Q., Lin, W., & Chan, A. B. (2021). Cross-View Cross-Scene Multi-View Crowd Counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 557-567).

E. *BM-Count*

BM-Count 는 가우시안을 포인트 주석에 부과하면 일반화 성능이 저하된다는 것이 입증한 것을 기준으로 구성된 다양한 방법들의 불편한 부분인 노이즈 주석을 해결하고자 포인트 감독(BM-Count)만 있는 군중 계산을 위한 이분 매칭 기반 방법을 제안하였다.

먼저 예측된 밀도 맵에서 가장 유사한 픽셀의 하위 집합을 선택하여 bipartite matching 을 통해 주석이 달린 픽셀과 일치시킨다. 그런 다음 일치하는 쌍을 기반으로 손실 함수를 정의하여 잘못된 위치에 주석이 달린 점으로 인한 나쁜 영향을 완화하였다.

그림 4 는 BM-Count 의 구조를 나타낸다.

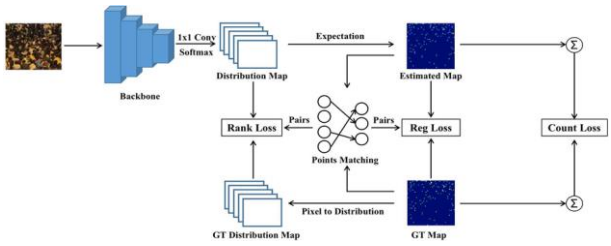

Fig. 4. BM-Count 모델 구조

[출처] Liu, H., Zhao, Q., Ma, Y., & Dai, F. Bipartite Matching for Crowd Counting with Point Supervision.

F. *CVCS*

다중 뷰 군중 계산은 이전에 다중 카메라를 활용하여 단일 카메라의 시야를 확장하고 장면에서 더 많은 사람을 포착하고 가려진 사람이나 저해상도의 사람들에 대한 계산 성능을 개선하기 위해 제안되었으나 동일한 단일 장면과 카메라 뷰에 대해 교육 및 테스트하므로 실제 적용에 제한이 있다.

임의의 카메라 레이아웃으로 다른 장면에서 훈련 및 테스트가 발생하는 CVCS(cross-view cross-scene) 다중 뷰 군중 계산 패러다임을 제안한다. CVCS 는 카메라 보정 오류 또는 잘못된 기능으로 인한 장면 및 카메라 레이아웃 변경 및 비대응 노이즈에서 최적의 뷰 융합 문제를 동적으로 처리하기 위해 카메라 레이아웃 기하학 및 노이즈를 사용하여 여러 뷰를 함께 신중하게 선택하고 융합하는 모델로 일치하지 않는 오류를 처리하도록 모델을 훈련하는 보기 정규화 방법. 또한 가능한 많은 변형을 캡처하기 위해 많은 수의 장면과 카메라 뷰가 있는 대규모 합성 다중 카메라 군중 계산 데이터 세트를 생성하여 이러한 큰 실제 데이터 세트를 수집하고 주석을 추가하는 어려움을 피할 수 있다.

그림 5 는 CVCS 의 구조를 나타낸다.

G. *CFANet*

군중 영역에 더 잘 초점을 맞추기 위해 주의 지도를 통합하여 고품질 군중 밀도 지도와 사람 수 추정을 생성하기 위한 새로운 방법인 CFANet(Coarse-and Fine grained Attention Network)을 제안하였다. CFANet 의 경우 CRR(Crowd Region Recognizer) 및 DLE(Density Level Estimator) 분기를 통합하여 미세한 점진적인 주의 메커니즘으로 구성되어 있다. 또한 CFANet 은 그라디언트의 역전파를 지원하고 과적합을 줄이기 위해 다단계 감독 메커니즘을 사용하고 있다.

그림 6 는 CFANet 의 구조를 나타낸다

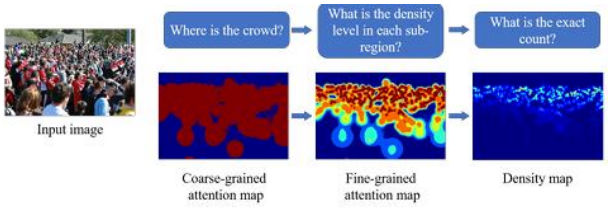

Fig. 6. CFANet 모델 설명

[출처] Rong, L., & Li, C. (2021). Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3675-3684)

H. NLT

Cross-domain crowd counting (CDCC) 은 공공 안전의 중요성으로 인해 뜨거운 주제이며, CDCC 는 소스 도메인과 대상 도메인 간의 이동을 완화하는 목표를 가지고 있다. 하지만 특정 작업의 경우 도메인 이동이 모델의 매개변수 차이에 반영된다는 것을 발견하였다. NLT(Neuron Linear Transformation)는 도메인 요인과 편향 가중치를 활용하여 도메인 이동을 학습하는 방법으로, NLT 는 특정 뉴런에 대해 도메인 이동 매개변수를 학습하기 위해 레이블이 지정된 대상 데이터를 거의 활용하지 않으며, 선형 변환을 통해 대상 뉴런을 생성하고 있다.

6 개의 실제 데이터세트를 사용하여 모델의 실험 및 분석을 진행하였으며, 감독 및 미세조정 훈련보다 더 강력하고 효과적임을 보였다.

그림 7 은 NLT 의 구조를 나타낸다.

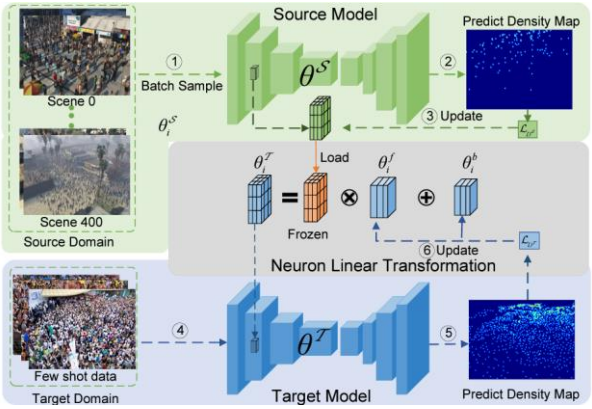

Fig. 7. NLT 모델 구조

[출처] Wang, Q., Han, T., Gao, J., & Yuan, Y. (2021). Neuron linear transformation: Modeling the domain shift for crowd counting. IEEE Transactions on Neural Networks and Learning Systems

I. DACC

군중 계산의 경우 수동으로 레이블이 지정된 데이터에 의존하고 있으며, 다른 대안으로 수동 레이블 없이 실제 데이터로 이동하는 방법을 이용할 수 있다. DACC(DomainAdaptive Crowd Counting)은 실제 데이터로 지식을 전송하는 과정에서 발생하는 도메인 갭을 효과적으로 억제하고 정교한 밀도 맵을 출력하는 부분을 고품질 이미지 번역 및 밀도 맵 재구성을 이용하여 해결하는 모델을 제안하였다.

도메인 공유 및 독립 기능을 분리하고 콘텐츠 인식 일관성 손실을 설계하는 과정을 통해 번역의 품질을 촉진시켰으며, 실제 장면에 의사 레이블을 생성하여 예측 품질을 향상시켰다.

그림 8 은 DACC 구조를 나타낸다.

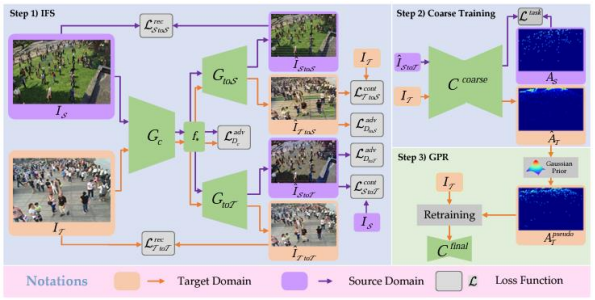

Fig. 8. DACC 모델 구조

[출처] Gao, J., Han, T., Yuan, Y., & Wang, Q. (2021). Domain-Adaptive Crowd Counting via High-Quality Image Translation and Density Reconstruction. IEEE Transactions on Neural Networks and Learning Systems.

J. STDNet

Conv3D 로 인한 모델 크기가 급격하게 증가하는 부분을 완화시키기 위해 3D convolution 과 3D spatiotemporal dilated density convolution 을 포함한 STDNet(SpatioTemporal convolutional Dense Network)을 제안하였다.

다중 스케일 특징을 추출하기 때문에 dilated convolution 을 채널의 주 블록과 결합하여 특징의 표현을 향상시켰다. 또한 새로운 패치 방식 회귀 손실인 PRL(patch-wise regression loss) 방식을 제안하여 군중 레이블 지정의 어려움으로 인해 발생하는 정확하지 않거나 표준이 일치하지 않는 불량을 해결하였다.

그림 9 는 STDNet 구조를 나타낸다.

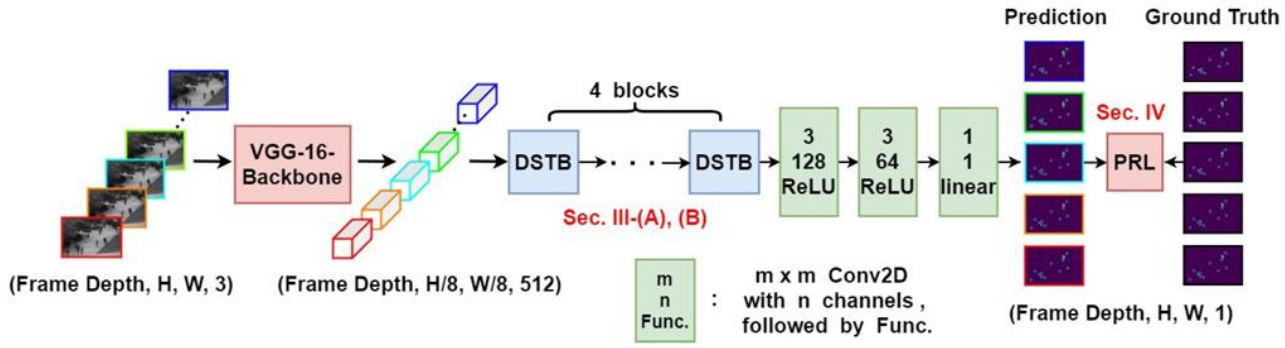

Fig. 9. STDNet 모델 구조

[출처] Ma, Y. J., Shuai, H. H., & Cheng, W. H. (2021). Spatiotemporal Dilated Convolution with Uncertain Matching for Video-based Crowd Estimation. IEEE Transactions on Multimedia

K. *AdaCrowd*

대상 장면에서 하나 이상의 레이블이 지정되지 않은 데이터를 사용하는 모델을 제안하였다. 해당 제안은 레이블이 지정되지 않은 장면 적응형 군중 계산이라는 새로운 유형이다.

AdaCrowd 는 군중 계산 네트워크와 안내 네트워크 총 2 개의 네트워크로 구성되어 있으며, 가이딩 네트워크는 특정 장면의 레이블이 지정되지 않은 이미지를 기반으로 군중 계산 네트워크의 일부 매개변수를 예측한다.

그림 10 은 AdaCrowd 의 모델 구조를 나타낸다.

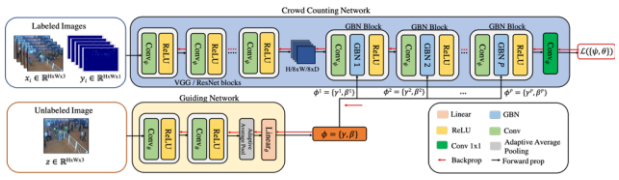

Fig. 10. AdaCrowd 모델 구조

[출처] Reddy, M. K. K., Rochan, M., Lu, Y., & Wang, Y. (2020). AdaCrowd: Unlabeled Scene Adaptation for Crowd Counting. arXiv preprint arXiv:2010.12141.

L. *PDANet*

군중 밀도의 광대한 규모 변화 등의 문제를 해결하기 위해 PDANet(Pyramid Density-Aware Attention-based networ)를 제안하였다.

밀도 인식 군중 계산을 위해 2 개의 분기 디코더 모듈을 사용하였으며 해당 모듈을 사용하여 다양한 스케일 기능을 추출하고 오해가 있는 부분을 억제하며, 독점적인 DAD(Density-Aware Decoder)를 사용하여 다양한 이미지 간의 혼잡도 수준의 변화를 해결한다. 입력 기능의 밀도 수준을 분류기에서 평가 후 높고 낮은 DAD 모듈로 전달하여 저밀도 맵과 고밀도 맵의 합을 공간적 관심을 고려하여 전체 밀도 맵을 생성한다. 정확한 믿로 맵을 생성하기 위해 두 가지의 손실 함수를 사용하였으며, 밀도 맵의 정확도 측면에서 우수함을 보여주었다.

그림 11 은 PDANet 의 모델 구조를 나타낸다

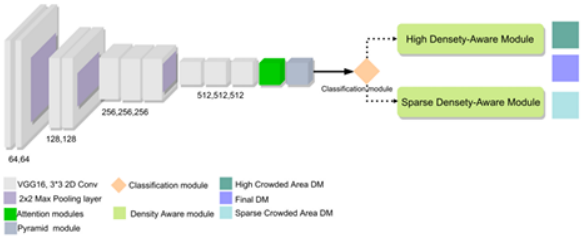

Fig. 11. PDANet 모델 구조

[출처] Amirgholipour, S., He, X., Jia, W., Wang, D., & Liu, L. (2020). Pdanet: Pyramid density-aware attention net for accurate crowd counting. arXiv preprint arXiv:2001.05643

M. *ScSiNet*

다중 열 구조의 문제점을 해결하기 위한 방법으로 ScSiNet(Single column Scale-invariant Network)을 제안하였다.

층간 다중 스케일의 통합과 새로운 레이어 내에서 스케일 불변 변환(scale-invariant transformation) 조합을 통해 정교한 스케일 분변 기능을 추출한다. 또한 밀도의 다양성을 확대하기 위해 무작위로 적분된 손실을 제공하고 있다.

III. 결론

군중 계산(Crowd Counting)을 하기 위한 방법으로 기존의 얼굴을 검출하여 군중의 수를 집계하는 것이 아닌 이미지의 밀도를 이용하여 추정하는 연구들이 이루어지고 있다. 빠르게 전송하고 오류의 손실을 줄이는 과정을 진행하며 점차적으로 군중 계산을 추정 또는 예측하는 결과가 정확해지고 있다. 현재의

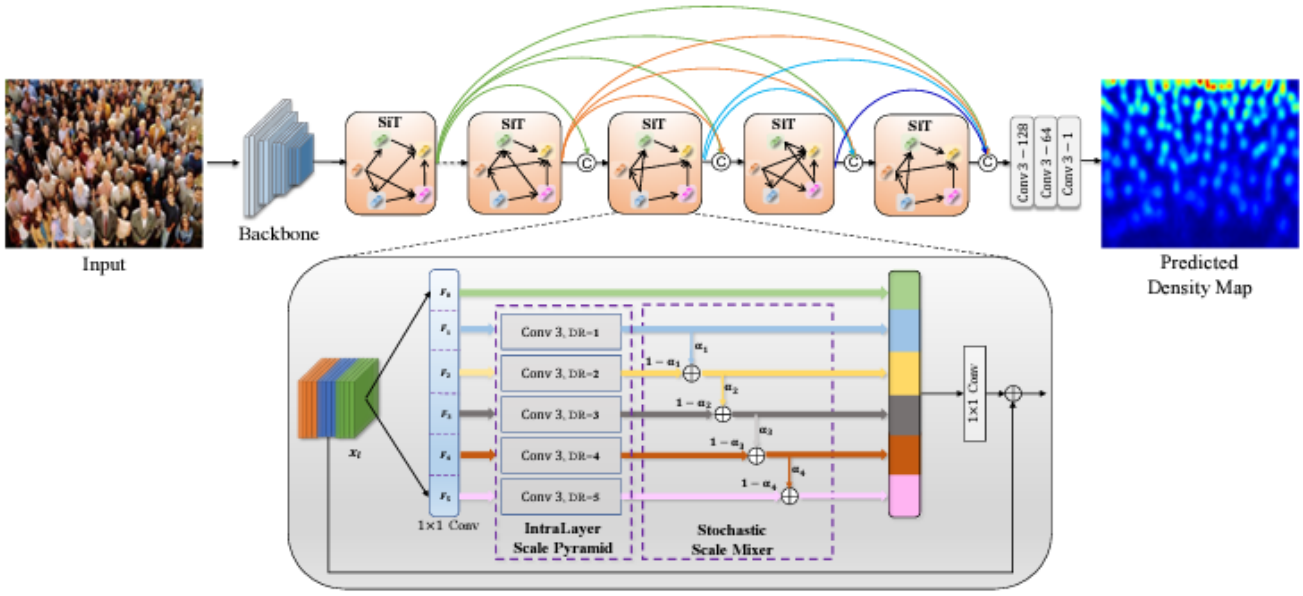

Fig. 12. ScSiNet 모델 구조

[2] [출처] Wang, M., Cai, H., Zhou, J., & Gong, M. (2021). Interlayer and intralayer scale aggregation for scale-invariant crowd counting. Neurocomputing, 441, 128-137

연구에서는 모델의 정확도 뿐만 아니라 모델을 학습하기 위한 전 단계인 전처리 과정을 좀 더 정밀하고 정확하게 하는 단계로 접어들고 있다.

또한 현재의 군중 계산의 경우 크기가 큰 군중 이미지를 대상으로 진행하는 경우가 많아 고성능의 규모를 구성할 필요가 있다. 해당 부분에 대한 문제를 해결하기 위한 방법의 연구가 진행되어야 한다.

참고문헌

[1] https://www.analyticsvidhya.com/blog/2021/06/crowd-counting-using-deep-learning/

[2] http://aiskyeye.com/challenge_2021/crowd-counting-2/

[3] https://www.crowdbenchmark.com/nwpucrowd.html

[4] Xu, Y., Zhong, Z., Lian, D., Li, J., Li, Z., Xu, X., & Gao, S. (2021). Crowd Counting With Partial Annotations in an Image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15570-15579)

[5] Ma, Z., Hong, X., Wei, X., Qiu, Y., & Gong, Y. (2021). Towards a Universal Model for Cross-Dataset Crowd Counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3205-3214)

[6] Wang, C., Song, Q., Zhang, B., Wang, Y., Tai, Y., Hu, X., ... & Wu, Y. (2021). Uniformity in Heterogeneity: Diving Deep into Count Interval Partition for Crowd Counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3234-3242)

[7] Liu, H., Zhao, Q., Ma, Y., & Dai, F. Bipartite Matching for Crowd Counting with Point Supervision

[8] Zhang, Q., Lin, W., & Chan, A. B. (2021). Cross-View Cross-Scene Multi-View Crowd Counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 557-567)

[9] Rong, L., & Li, C. (2021). Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3675-3684)

[10] Wang, Q., Han, T., Gao, J., & Yuan, Y. (2021). Neuron linear transformation: Modeling the domain shift for crowd counting. IEEE Transactions on Neural Networks and Learning Systems

[11] Gao, J., Han, T., Yuan, Y., & Wang, Q. (2021). Domain-Adaptive Crowd Counting via High-Quality Image Translation and Density Reconstruction. IEEE Transactions on Neural Networks and Learning Systems

[12] Ma, Y. J., Shuai, H. H., & Cheng, W. H. (2021). Spatiotemporal Dilated Convolution with Uncertain Matching for Video-based Crowd Estimation. IEEE Transactions on Multimedia

[13] Reddy, M. K. K., Rochan, M., Lu, Y., & Wang, Y. (2020). AdaCrowd: Unlabeled Scene Adaptation for Crowd Counting. arXiv preprint arXiv:2010.12141

[14] Amirgholipour, S., He, X., Jia, W., Wang, D., & Liu, L. (2020). Pdanet: Pyramid density-aware attention net for accurate crowd counting. arXiv preprint arXiv:2001.05643

[15] Wang, M., Cai, H., Zhou, J., & Gong, M. (2021). Interlayer and intralayer scale aggregation for scale-invariant crowd counting. Neurocomputing, 441, 128-137.