YOLO 학습 성능 향상을 위한 1채널 데이터의 3채널 합성 연구

이승우[0], 유승현[1], 서종웅[1], 백화평[1], 정용화[1]
[0]고려대학교 세종캠퍼스 인공지능사이버보안학과
[1]고려대학교 세종캠퍼스 컴퓨터소프트웨어학과
rainup4632@korea.ac.kr
tidlsld44@korea.ac.kr
qorrns156@korea.ac.kr
seojongwoong@korea.ac.kr
ychungy@korea.ac.kr

요 약

스마트팜과 같은 첨단 시스템은 돼지와 같은 가축의 효율적인 관리를 위해 필수적이며, 이러한 시스템 개발에서 주요 과제 중 하나는 학습 시 관찰되지 않은 농장 환경(Unseen Dataset)에서 탐지 정확도의 저하 문제이다. 본 연구는 Gray 스케일 데이터를 활용하여 Unseen Dataset에서도 높은 탐지 성능을 유지하면서 속도 저하 없이 학습 효율을 극대화할 수 있는 방법을 제안한다. 학습 과정에서 데이터의 텍스처(Texture) 정보에 지나치게 의존하는 문제를 완화하기 위해 Depth Anything과 CLAHE 기법을 적용하여 Edge 및 Texture 데이터를 생성하고, 이를 Gray 스케일 데이터와 결합하여 최적의 채널 조합을 설계하였다. 실험 결과, Gray 스케일 데이터를 B 채널에, Edge 데이터를 G 채널에, Texture 데이터를 R 채널에 배치한 모델이 Baseline 대비 AP50 기준 84.5%에서 90.8%로 6.3% 향상된 성능을 기록하며, 속도 저하 없이 제안된 접근법의 우수성을 입증하였다.

1. 서론

돼지는 전 세계적으로 중요한 가축으로, 인간의 필수적인 영양소 공급원으로서 큰 역할을 하고 있습니다 [1]. 이러한 돼지를 효율적으로 관리하기 위해서는 인간의 직접적인 노력만으로는 한계가 있으며, 스마트팜과 같은 첨단 농장 관리 시스템의 도입이 필요합니다. 그러나, 돼지와 같은 데이터에서는 데이터의 부족으로 인해 컬러 이미지를 제공하지 못하는 경우가 존재하며, 이로 인해 회색조 이미지나 제한된 데이터를 활용해야 하는 상황이 발생합니다. 스마트팜을 도입하기 위해서는 이러한 문제를 포함한 여러 도전을 해결해야 하며, 그중에서도 학습 과정에서 한 번도 관찰되지 않은 돼지 농장 환경(Unseen 데이터셋)에서의 낮은 모델 정확도가 주요 과제로 대두되고 있습니다.

현재 Transformer [2], CNN [3], 그리고 RNN [4]등 다양한 딥러닝 네트워크 구조가 개발되고 있지만, 실제 농장 환경에서 활용되기 위해서는 처리 시간과 정확도를 동시에 고려해야 합니다. 이러한 요구사항을 충족하기 위해, 상대적으로 효율적이고 경량화된 구조를 가진 CNN이 주로 사용되고 있습니다. CNN 네트워크 구조는 RGB(Red, Green, Blue) 이미지 데이터를 입력 영상으로 입력 받은 뒤 해당 영상들을 이용하여 학습한다. 이는 자연 이미지 데이터의 특성을 반영한 것으로, 다양한 시각적 정보와 색상 데이터를 효과적으로 학습하기 위한 기반이 된다.

이러한 네트워크에 단일 채널 데이터(예: Grayscale 영상)를 입력으로 사용할 경우, 기존 3채널 기반 네트워크 설계에 비해 구조적 제약이 발생할 가능성이 있다. 이는 네트워크가 채널 간 상호작용을 학습하도록 설계된 특성을 충분히 활용하지 못하기 때문이며, 단일 채널 데이터의 경우 별도의 전처리 과정이나 맞춤형 네트워크 구조 설계가 요구될 수 있기 때문이다. 특히, 채널 정보가 부족한 상황에서는 네트워크가 데이터의 특성을 충분히 학습하지 못해 성능 저하로 이어질 가능성이 있다 [5].

본 연구에서는 단일 채널 학습 데이터를 효과적으로 활용하기 위한 새로운 접근법을 제안하였다. 다양한 영상 특징(features) 정보을 강조하고, 각 채

널별 정확도와 특징 변화의 영향을 체계적으로 분석하기 위해 여러 영상 처리 기법을 적용하고 실험을 수행하였다. 특히, 학습 네트워크가 텍스처(texture) 정보에 지나치게 의존하는 문제를 완화하기 위해, 외곽선(edge) 정보를 강조한 영상을 생성하여 사용하고 텍스처 정보를 변형하는 방식을 도입하였다. 이를 통해 그림1과 같이 학습 모델이 보다 일반화된 특징을 학습하도록 유도하였다.

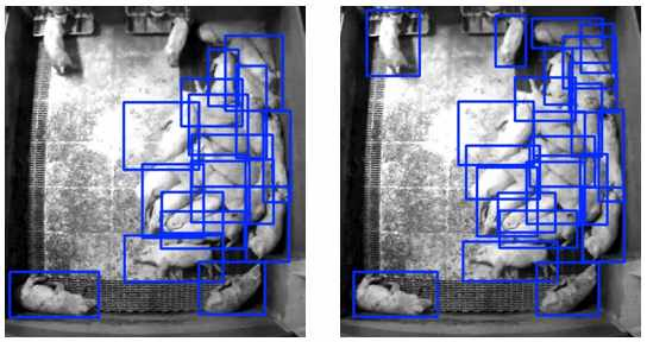

그림 1. 기존 모델과 제안 방법 비교.

2. 연구 방법

본 연구에서는 그림 2와 같이 3채널 기반 탐지기인 CNN(Convolutional Neural Network)의 구조적 특성을 효과적으로 활용하기 위해, 1채널로 구성된 데이터셋의 Gray 스케일 이미지를 데이터 처리 과정을 통해 각 RGB 채널에 매핑하여 입력 데이터를 구성한 후 학습을 진행하였다.

원본 영상은 Gray 스케일 영상으로 설정하였고, 다른 두 가지 영상은 각각 Invert_nv_z와 CLAHE(Contrast Limited Adaptive Histogram Equalization)로 구성하였다. 우선, Edge 강조 영상인 Invert_nv_z를 생성하기 위해 필요한 전경 정보를 구하는 데 Foundation Model로 잘 알려진 **Depth Anything [6]**을 사용하였다. 이 과정에서 획득한 전경 영상을 기반으로 3D 표면의 법선 벡터를 계산한 뒤, z 성분만 추출하여 표면의 수직 기울기를 단일 채널 이미지로 표현하였다. 이를 통해 픽셀 값이 낮은 영역에서 경계선을 강조하는 데이터를 생성하였다. 이후, Edge 데이터의 특징을 더욱 두드러지게 표현하기 위해 Invert 연산을 수행하여 Edge 정보를 강화하였다. 이렇게 생성된 Invert_nv_z는 외부 경계선을 뚜렷하게 표현하고 내부 영역은 흐리게 표현하여, 물체의 외곽과 내부를 분리하는 Edge 데이터로 활용하였다.

또한, 학습에 사용되지 않은 흑돼지나 얼룩돼지와 같은 특수 사례에서 발생하는 텍스처(texture) 차이를 처리하고, 배식통과 돼지를 구분하기 위해

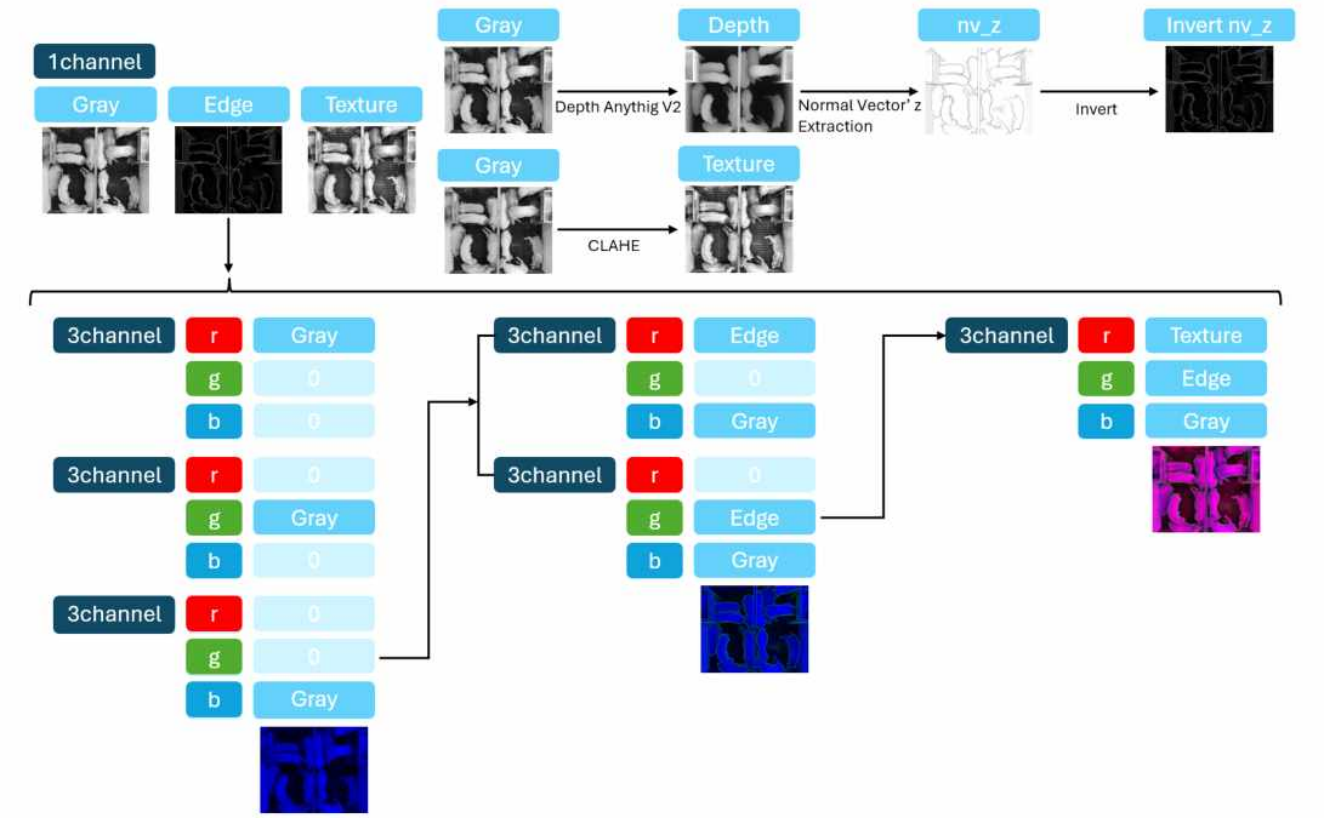

그림 2. 제안된 방법 개요.

학습데이터 구성 방식	Box Precision ↑ (%)	Box Recall ↑ (%)	Box AP_{50} ↑ (%)	F1-score ↑ (%)
Gray 데이터를 단일 채널로 입력한 모델	84.6	76.7	**84.5**	80.4
Gray 데이터를 R채널로만 입력한 모델	82.5	80.1	**87.1**	81.3
Gray 데이터를 G채널로만 입력한 모델	91.3	78.1	**85.0**	79.7
Gray 데이터를 B채널로만 입력한 모델	85.3	79.7	**87.7**	82.4

표 1. Gray 데이터를 R, G, B 채널에 할당한 모델의 성능 지표 비교.

CLAHE 기법을 적용하여 학습 데이터를 보완하였다. 이 과정에서 히스토그램 클리핑 제한 값은 2로, 그리드 크기는 9×9로 설정하여 텍스처 데이터를 생성하였다. 생성된 텍스처 데이터는 돼지와 기타 물체의 질감을 구분하는 데 활용되었다.

3. 실험 결과 및 분석

본 실험에서는 탐지기로 YOLO [7] 모델을 채택하였고 그중에서도 yolov11을 사용하였다, 데이터셋으로 독일 돼지 [8]를 사용하였다. CPU는 13th Gen Intel(R) Core(TM) i7-13700K를 사용하였고, GPU는 GeForce RTX 4090을 사용하였다. 학습 설정은 train = 788 images, valid = 197 images, batch size = 16, epochs = 300으로 구성한 뒤 학습을 진행하였다. 입력 데이터 구성 방식이 모델 학습 성능에 미치는 영향을 체계적으로 분석하기 위해 두 가지 접근 방식을 설계하였다.

3.1 Gray 스케일을 3채널에 독립적으로 할당

각 채널에 할당된 Gray 스케일이 학습 과정에서 어떻게 활용되는지를 분석하기 위해, 그림 3과 같이 Gray 스케일을 R, G, B 채널에 각각 독립적으로 할당하고 나머지 채널은 0으로 설정하여 학습 성능을 비교하였다. 그림 2를 통해 시각적으로 확인할 수 있듯이, Gray 스케일을 B 채널에 할당하여 학습했을 때, R 채널과 G 채널에 할당했을 때보다 false negatives와 false positives가 크게 개선된 것을 확인할 수 있었다.

표 1의 실험 결과에 따르면, Gray 스케일을 단일 채널로 지정한 baseline과 비교했을 때, B 채널에 할당한 모델이 AP50 기준으로 84.5%에서 87.7%로 3.2% 향상된 성능을 보였다. 다양한 채널 구성을 학습 데이터로 사용한 경우, 평가 기준에 따라 최상의 정확도는 다르게 나타났으나, AP50 기준에서는 B 채널 할당이 가장 높은 정확도를 기록하였다. 이는 B 채널이 Gray 스케일의 시각적 표현력을 극대화하여 학습 효율을 효과적으로 높였음을 의미한다.

그림 4, 5는 각각 본 실험에서 사용된 YOLOv11로 학습된 모델들의 3, 4 레이어에서 추출된 features를 시각화한 결과이다. 오른쪽 바와 같이 노란색으로 갈수록 높은 가중치를 부여한 것을, 파란색으로 갈수록 낮은 가중치를 부여한 것을 의미한다. 돼지 하단 영역과 바닥의 구분에서도 B 채널 기반 모델이 가장 명확하고 일관된 경계를 학습했음을 보여준다. 이러한 결과는 B 채널 입력 방식이 Gray 스케일의 특징을 가장 효과적으로 반영했음을 시사한다.

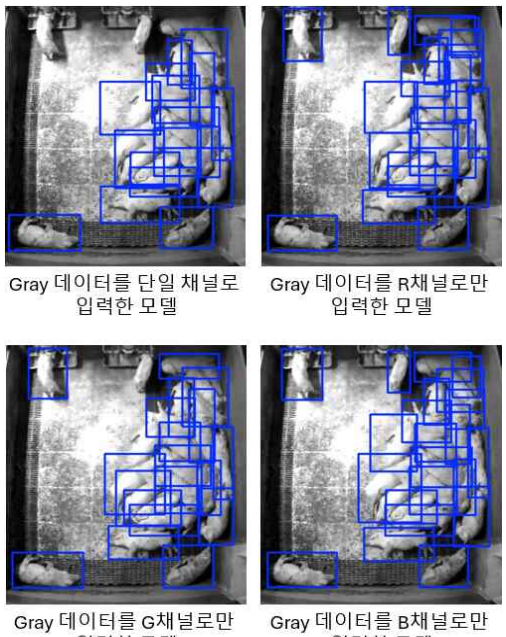

그림 3 . 모델간 차이 시각화.

1 channel		3channels		Box Precision ↑ (%)	Box Recall ↑ (%)	Box AP50 ↑ (%)	F1-score ↑ (%)
Gray	Blue	Green	Red				
O (Baseline)	X	X	X	84.6	76.7	**84.5**	80.4
X	Gray	0	0	85.3	79.7	**87.7**	82.4
		Edge	0	88.5	78.7	**88.7**	83.3
			Texture	89.7	83.8	**90.8**	86.6

표 2. baseline 대비 정확도 향상 비교.

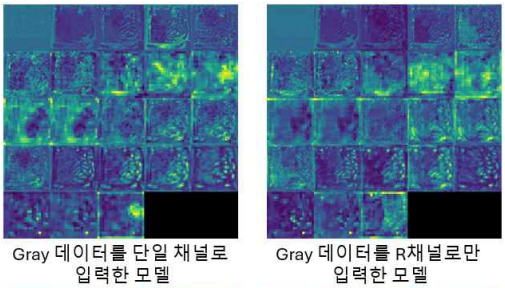

그림 4. 특징추출.

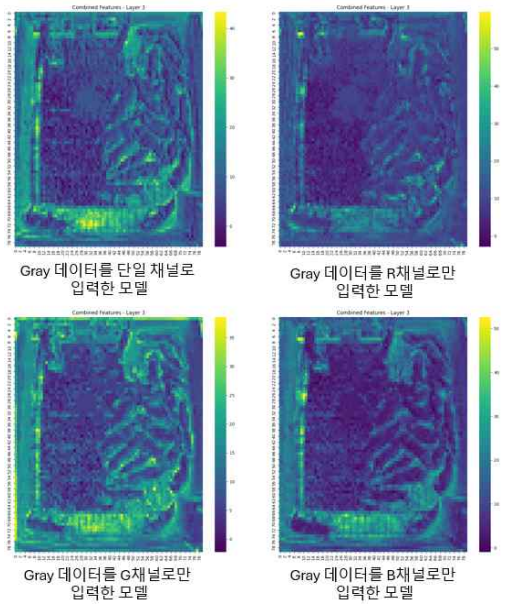

그림 4. 특징 추출 003 layer.

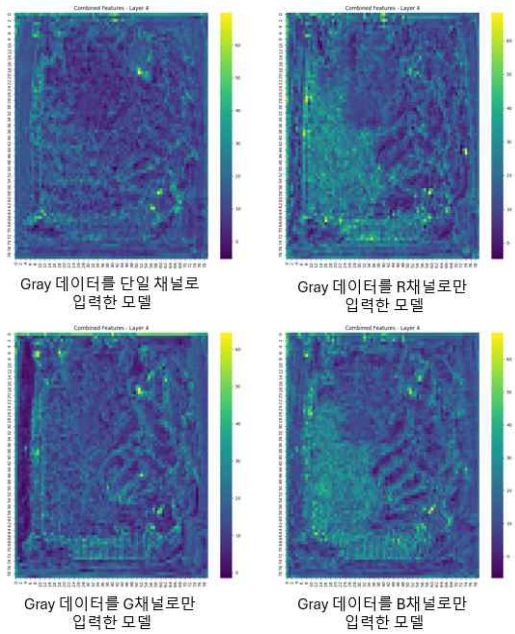

그림 5. 특징 추출 004 layer.

3.2 채널 조합 최적화를 통한 텍스처 의존성 완화

앞선 실험에서는 B 채널을 Gray 스케일 기반으로 설정하고, 텍스처(texture) 정보에 지나치게 의존하는 문제를 완화하기 위해 나머지 두 채널인 R 채널과 G 채널에 대해 최적의 3채널 조합(edge, texture)을 탐색하였다. 여기서 edge 데이터는 Gray 스케일을 기반으로 깊이(Depth)와 법선(normal) 정보를 강조하여 생성된 데이터이며, texture 데이터는 이미지의 질감을 부각시키기 위해 대비를 향상시켜 생성된 데이터이다.

표 2의 실험 결과에 따르면, G 채널에 Edge 데이터를 할당할 경우 B 채널에 Gray 스케일만 할당

한 모델과 비교하여 Precision은 3.2%, AP50은 1%, F1-score는 0.9% 향상되었다. 그러나 Recall이 1% 감소하면서 FP가 증가하고 TP가 감소하는 단점이 관찰되었다. 반면, R 채널에 Texture 데이터를 할당한 경우 이전 모든 실험 결과와 비교했을 때 모든 평가 기준에서 성능이 상승하였으며, 최대 6.9%의 성능 향상이 나타났다.

이러한 결과는 Edge와 Texture 데이터를 활용한 채널 배치 방식이 단순 Gray 스케일 입력보다 훨씬 풍부한 특징을 학습할 수 있도록 모델을 지원함을 보여준다. 특히, Texture 데이터를 R 채널에 배치하고 Edge 데이터를 G 채널에 배치하며 Gray 스케일을 B 채널에 고정한 방식은 입력 데이터의 다양성과 특성을 최대한 활용하여 모델의 학습 성능을 극대화한 최적의 조합임을 시사한다.

4. 결론

본 연구에서는 단일 Gray 스케일을 입력 데이터로 활용할 때 최적의 채널 배치를 탐색하였다. 첫 번째 실험에서는 Gray 스케일을 R, G, B 채널 중 하나에 독립적으로 할당한 결과, B 채널에 데이터를 입력했을 때 AP50 지표가 87.7%, F1-score가 82.4%로 가장 우수한 성능을 기록하며, 특정 채널에 데이터를 적절히 배치하는 것이 효과적임을 확인하였다.

두 번째 실험에서는 B 채널에 Gray 스케일을 고정한 상태에서 G 채널에 Edge 데이터를, R 채널에 Texture 데이터를 추가한 최적의 조합을 통해 AP50 지표를 90.8%, F1-score를 86.6%로 향상시키며, 단순 Gray 스케일 입력 방식 대비 각각 6.3%와 6.2%의 성능 향상을 기록하였다.

이 결과는 단일 Gray 스케일을 활용할 때 채널 배치와 추가 특성 조합이 모델 학습 성능을 극대화할 수 있음을 시사하며, 향후 단일 채널 기반 데이터 활용과 모델 설계에 새로운 가능성을 제공할 것이다.

감사의 글

본 연구(결과물)는 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신사업(2021RIS-004)의 일환으로 이루어졌으며, 또한 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 지역혁신클러스터육성사업(P0024177)의 지원을 받아 진행되었습니다.

참고문헌

[1] H. C. J. Godfrayet al., "Meat consumption, health, and the environment," Science, vol. 361, no. 6399, p. eaam5324, 2018.

[2] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[4] R. M. Schmidt, "Recurrent neural networks (rnns): A gentle introduction and overview," arXiv preprint arXiv:1912.05911, 2019.

[5] I. Ahmad and S. Shin, "An approach to run pre-trained deep learning models on grayscale images," in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021: IEEE, pp. 177-180.

[6] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10371-10381.

[7] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," arXiv preprint arXiv:2410.17725, 2024.

[8] M. Riekert, A. Klein, F. Adrion, C. Hoffmann, and E. Gallmann, "Automatically detecting pig position and posture by 2D camera imaging and deep learning," Computers and Electronics in Agriculture, vol. 174, p. 105391, 2020.