Improving Dog-Length Estimation Framework

24-2 Yonsei DSL Corp. Project

PetNow X Yonsei DSL

Jungwoo Kim, Hyunjin Kim, Chaemin Hwang, Jeongwoo Lee, Joowon Yang, Jongwook Jeon, Minkyu Kim, Youngil Lee, Yingjun Shen, Hyunah Ko

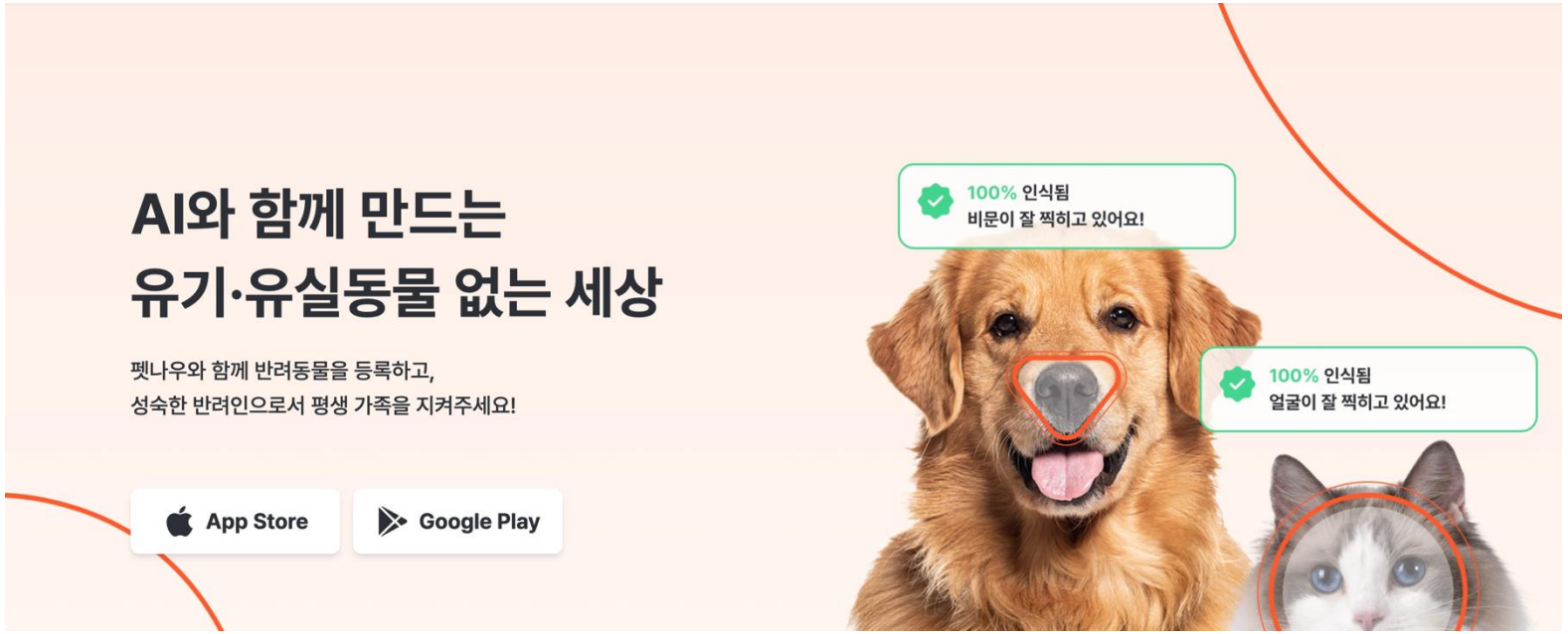
dslab.yonsei@gmail.com

Contents

Problem Statement

PetNow Overview

- AI를 이용한 반려동물 생체인식 (Nose Fingerprint) 및 생체정보 자동취득 기술을 보유.
- 인공지능 연구역량 보유 박사급 1명 석사급 1명 학사급 3명, 인공지능 학습 전용 서버 운용 중 (3 A100)



Problem Statement

Problem Statement

- 기업에선 반려동물 보험 언더라이팅 과정에서 필요한 반려견 체장 정보를 수기로 기록하는 문제를 해결하기 위해 Dog-body length estimation 기술 개발을 진행 중.
- 보다 구체적으로, 해당 기업에서는 'monocular camera set-up'에서 동작하는 모델 개발을 목표로 함.
- 해당 과정에서 3D space에서의 실제 physical distance 측정을 위해 camera parameter를 다루는 과정에서 문제가 발생하고 있음.
- 프로젝트 전반에서 cm 단위의 정확도를 가지는 성능 개선을 최종 목표로 함.

Task Information

- 1차적 목표로는 pixel-space에서 실제 3D space로 적절한 transformation을 위해 필요한 camera parameter의 정의 등의 camera calibration 정보 취득을 목표로 함.
- 기존의 기업 pipeline에 대한 개선 사항 모색 및 전체적 성능 개선을 통해 cm 단위의 모델 개발을 본 프로젝트에서 달성 목표로 함.
- 현재 진행 중인 기업의 데이터 수집 과정에서 추후 프로젝트 진행을 위해 필요한 수집 데이터의 정의 또한 추가적으로 함께 진행함.
- 최종적으로 취득한 parameter를 이용해 다양한 카메라 기종에 대한 적용을 목표로 함.

Data Formulation

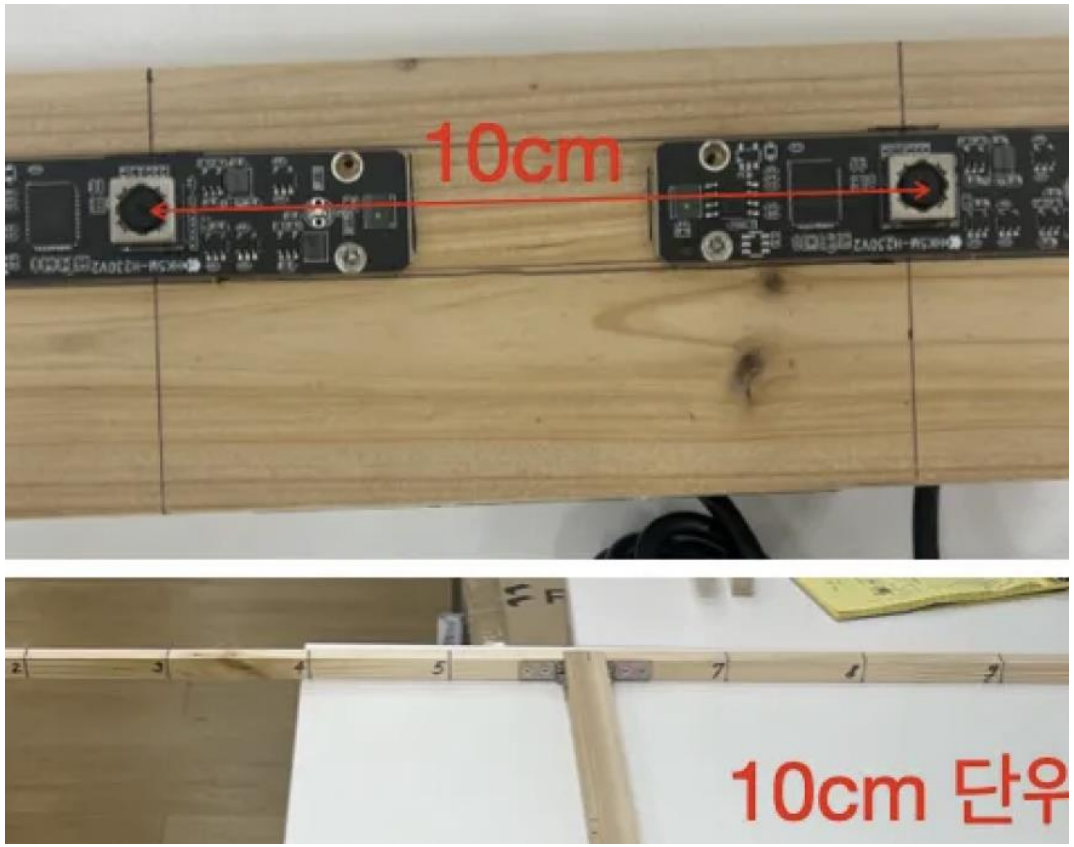


Figure 1. 데이터 수집 환경. 2-cam 환경에서 수집한 데이터 중 하나의 카메라 비디오를 사용. 대략적인 depth 측정을 위한 115cm 바와 개의 체장 측정을 위한 120cm 바로 구성됨.

Data Formulation

Figure 2. PetNow 데이터 예시. 전체 데이터셋은 10-30초의 40-50개의 비디오로 구성됨. 해당 비디오로부터 체장이 잘 측정된 프레임을 추출한 뒤 직접 체장을 라벨링하여 데이터셋 구성. 최종적으로 21개의 이미지를 수집.

Data Formulation

AI-Hub Dataset

- 기업에서 제공한 21개의 test image로는 프로젝트 진행에 어려움이 있어 AI-Hub의 반려견/반려묘 건강 정보 데이터를 추가로 수집하여 정제.
- 20개의 각도에서 취득한 다양한 사진 중 2개의 적절한 사진 2장을 선정하여 데이터셋 구성. (08-우측면 중앙 & 10-우측면우45도)
- 각각의 이미지 정보로는 견종, 나이, 체장정보 (등허리 길이, 목둘레, 흉곽둘레, 체중) 등을 포함하고 있음.
- 최종적으로 10798개의 추가적인 이미지 데이터셋을 구축.

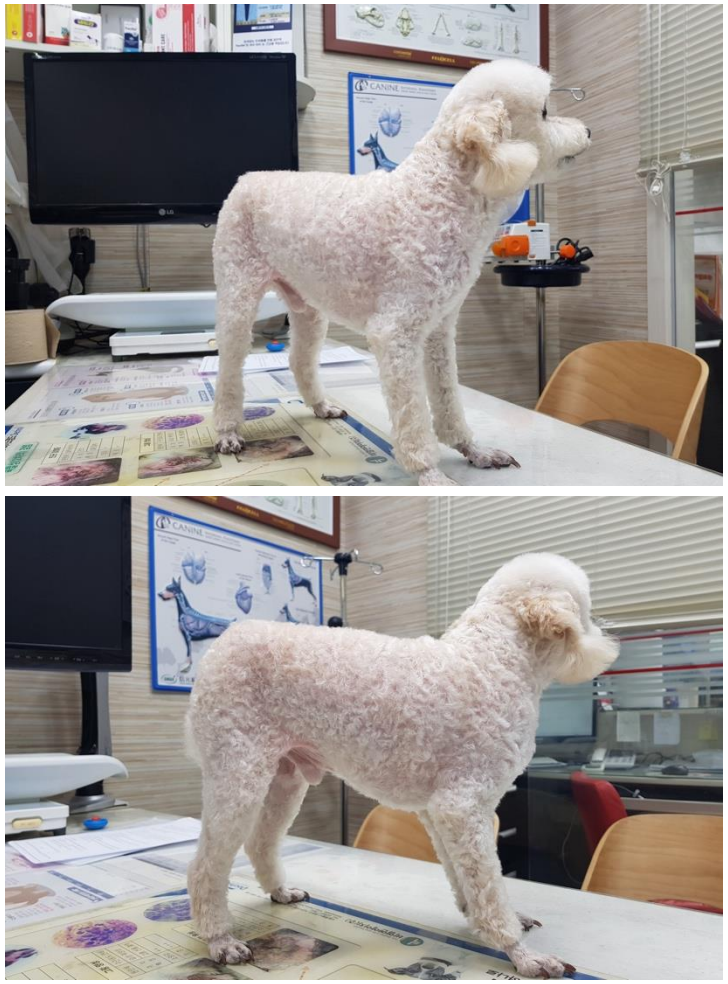


Figure 3. AI-Hub 데이터셋 예시. 우측면우45도와 우측면중앙 사진을 선별하여 활용

Naïve Pipeline

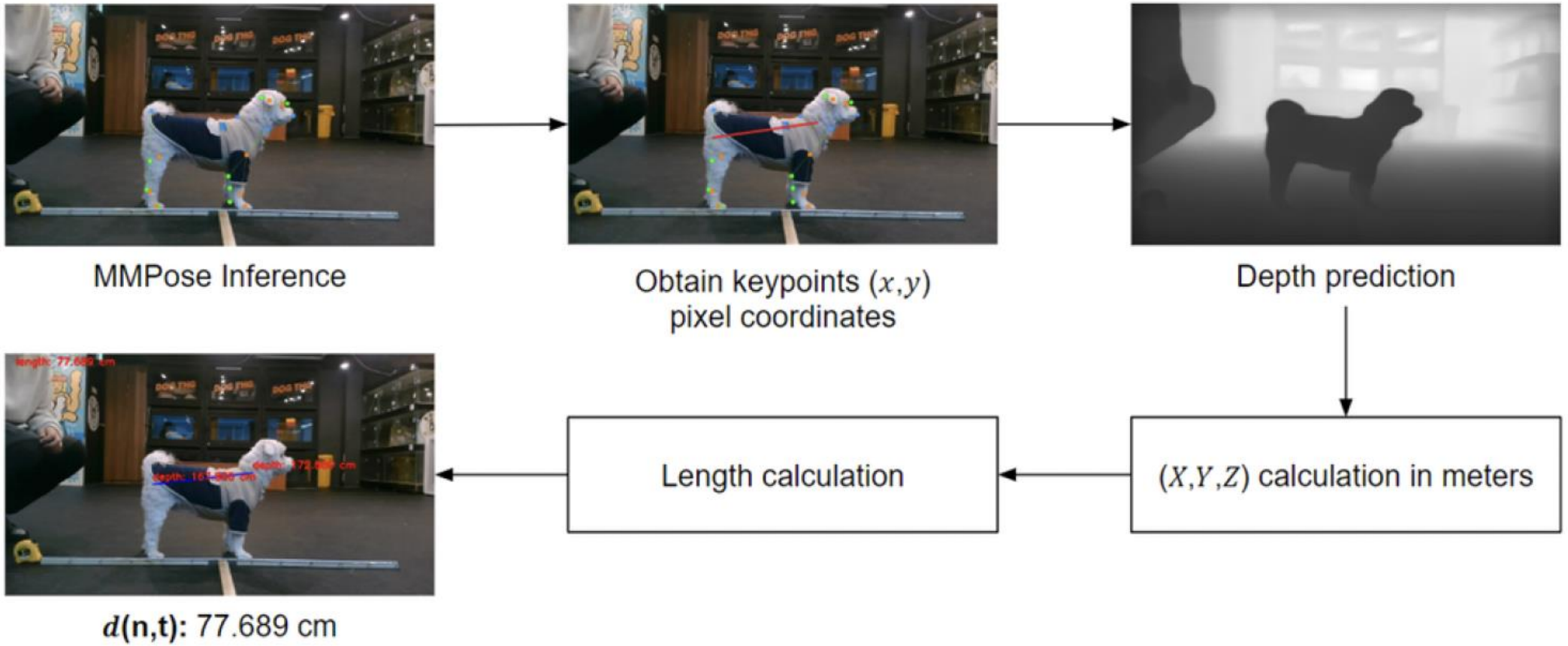
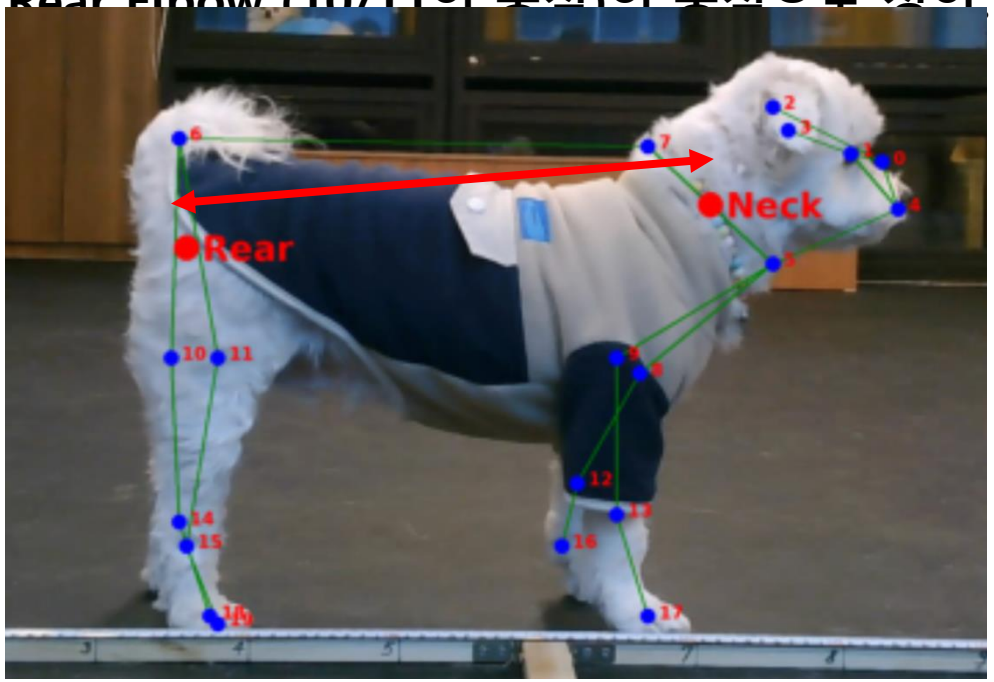


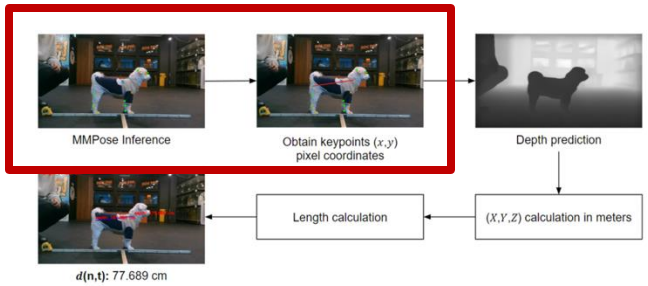
Figure 4. 기존 PetNow에서 사용한 체장 측정 파이프라인. Depth Prediction과 Transformation을 통해 얻은 3차원 좌표를 기반으로 삼각측량법으로 실제 3D 좌표를 연산.

Step 1-2. MMPose Inference and Keypoints Extraction

MMPose [1] Inference & Keypoint Extraction

- Pre-trained model을 활용해 20개의 keypoint를 아래 figure와 같이 추출함
- Neck Point는 Throat (5)와 Withers (7)의 중점, Rear Point는 Tail Start (6)과 Rear Elbow (10/11의 중점)의 중점으로 정의

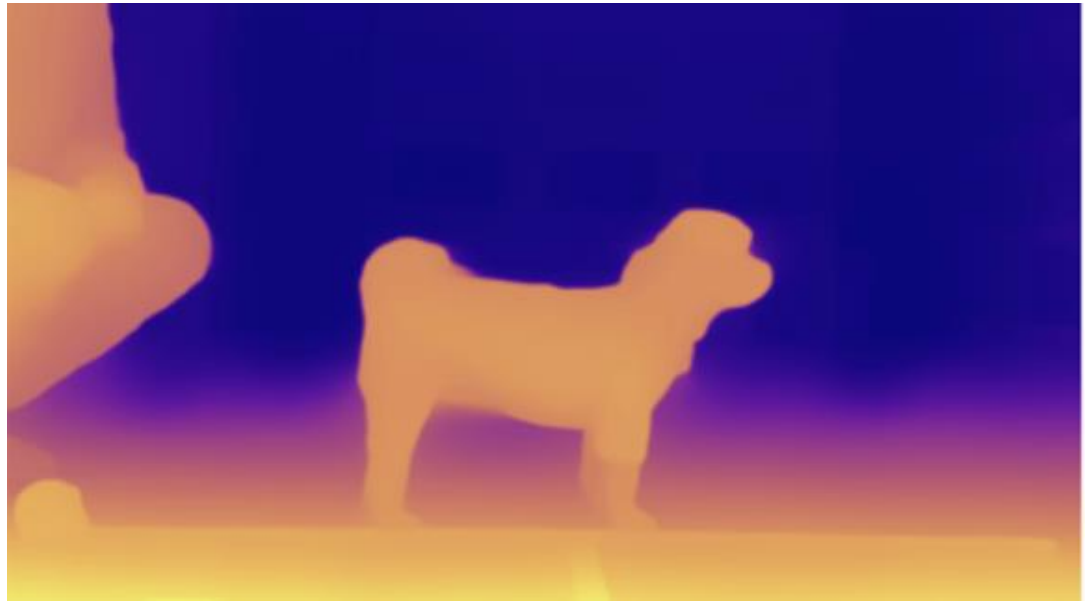


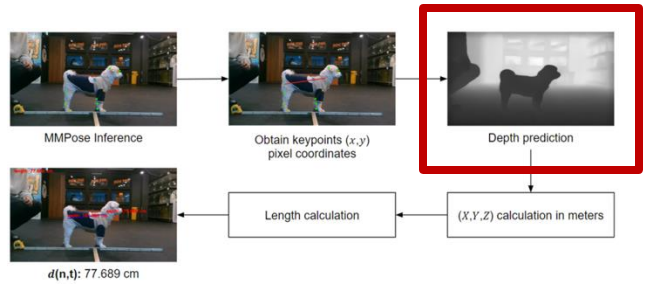


Step 3. Depth Estimation

Depth Estimation

- Pixel 별로 실제 object까지 거리인 Metric Depth를 예측하는 모델을 이용해 거리를 측정.
- 기업 Pipeline에서는 Depth-Anything v2 [2] 모델을 Metric Depth에 맞게 fi ne-tuning한 버전의 모델을 사용.





Step 4-5. 2D-to-3D Transformation

2D-to-3D Transformation

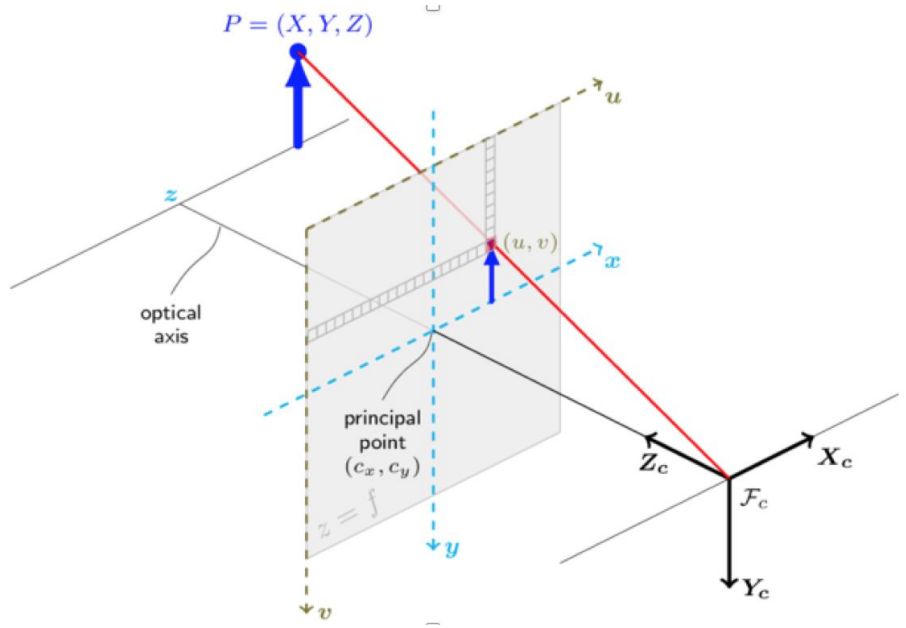
- **Image Plane에서 pixel 단위의 2D로 정의된 좌표를 실제 3D 공간 상에서의 물리적 거리로 변환함.**

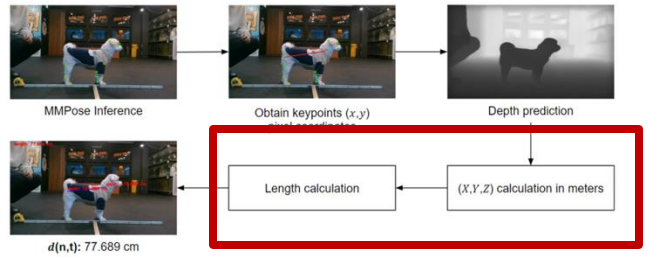
- **Focal Length** (f_x, f_y) 와 Principal Point (p_x, p_y), 그리고 depth value (d_x, d_y)를 이용해서 2D pixel coordinate인 (x, y)를 3D 공간적 좌표인 (X, Y, Z)로 변환함.

$$X = \frac{(x - c_x) \cdot d_x}{f_x}$$

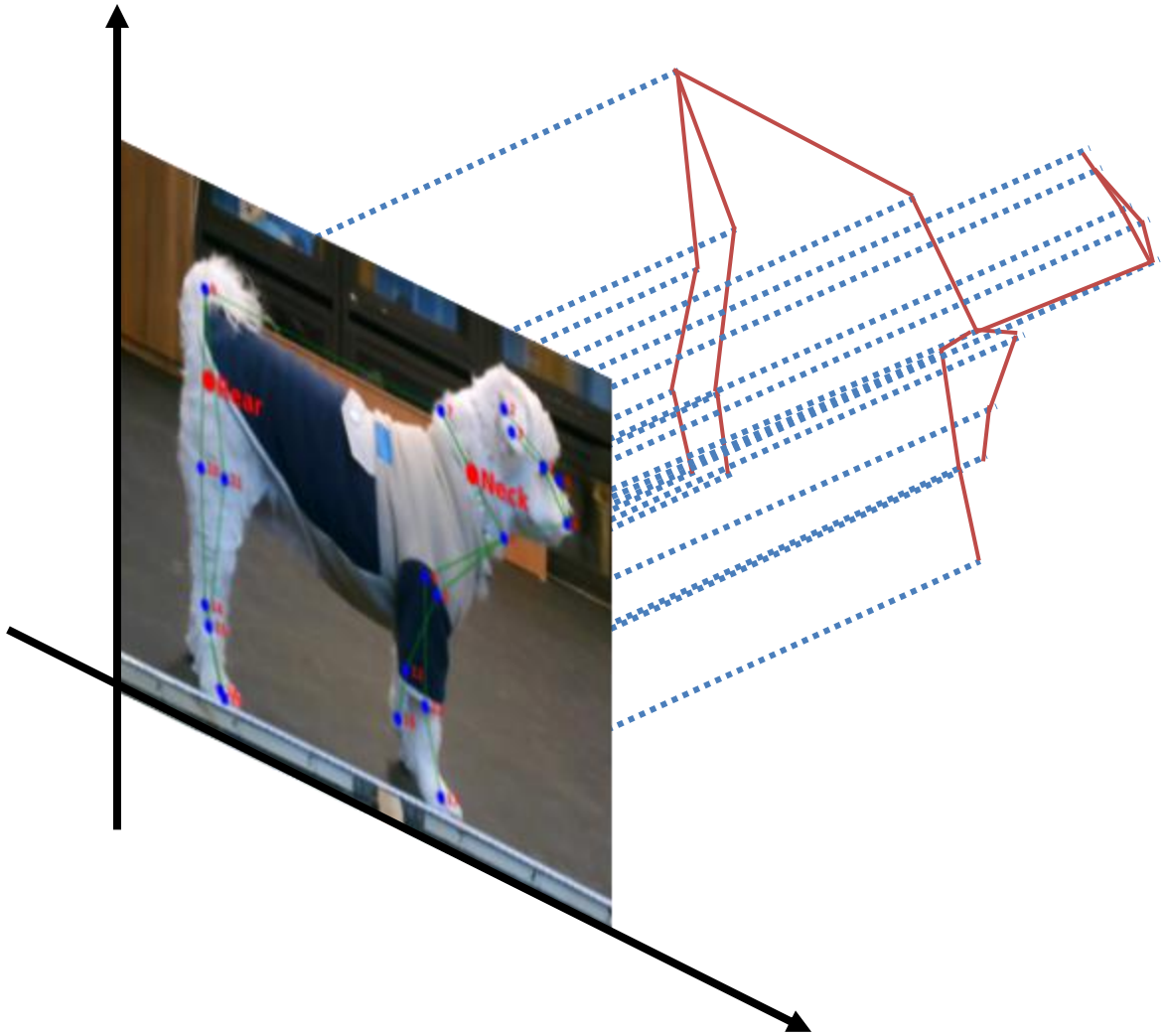
$$Y = \frac{(x - y) \cdot d_y}{f_y}$$

$$Z = d$$

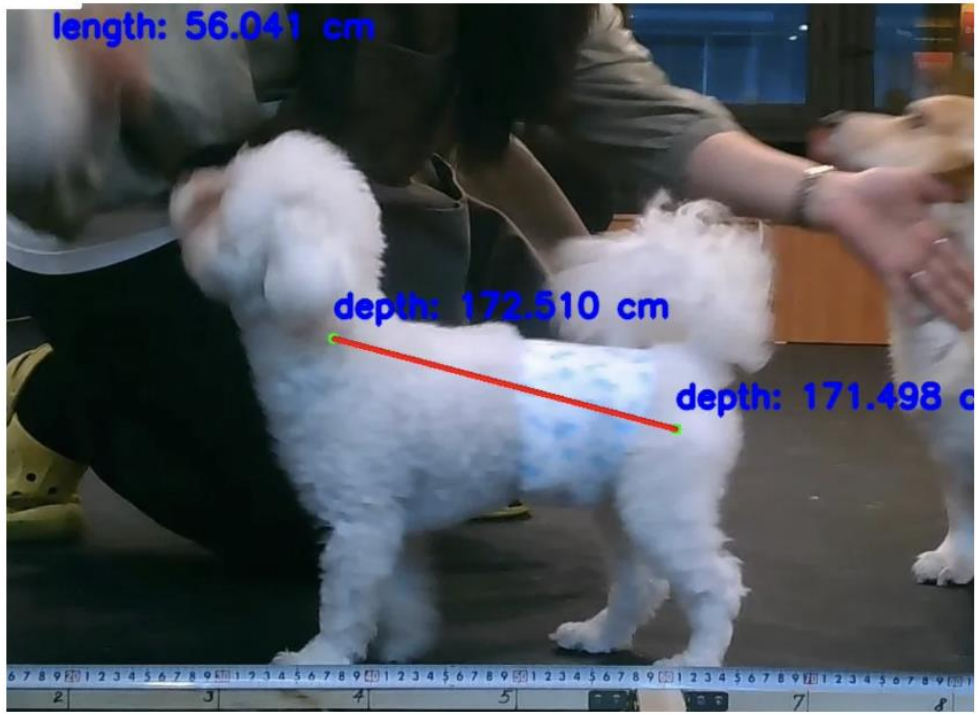




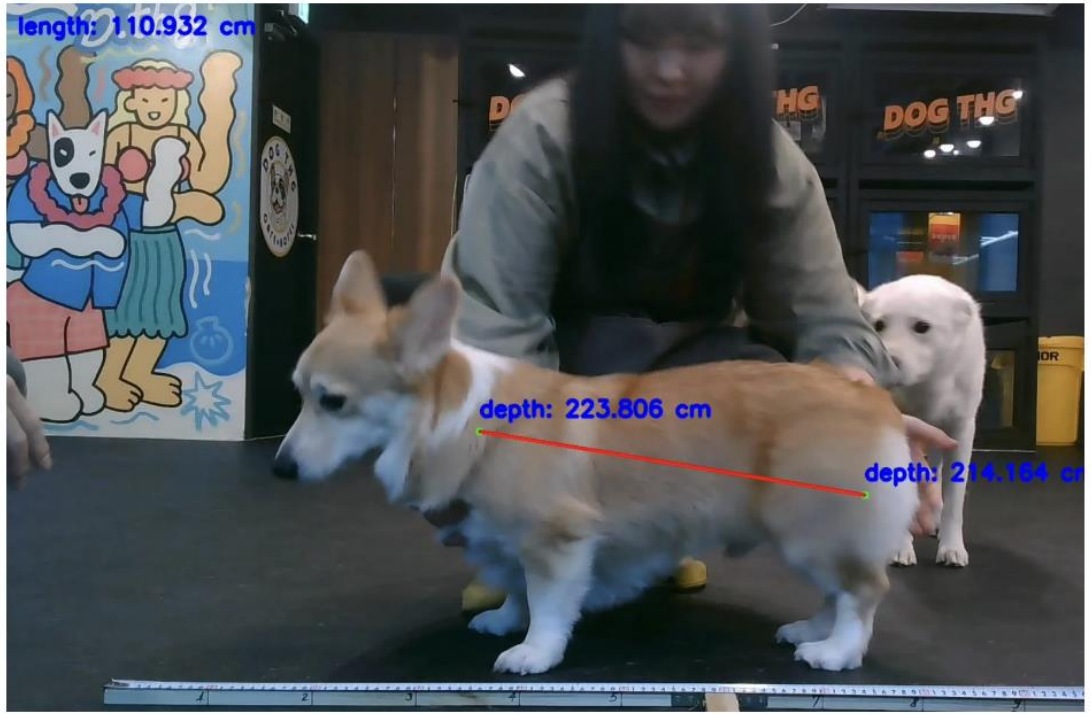
Overall Pipeline Visualization



Limitations in Naïve Pipeline



Model predicted length: 56cm
Actual length ≈ 30cm



Model predicted length: 110cm
Actual length ≈ 45cm

Figure 5. PetNow에서 제안한 Pipeline을 통해 측정한 체장 데이터. 실제 비디오에서 확인한 Ground Truth 체장과 큰 오차를 보임.

Pipeline Improvement

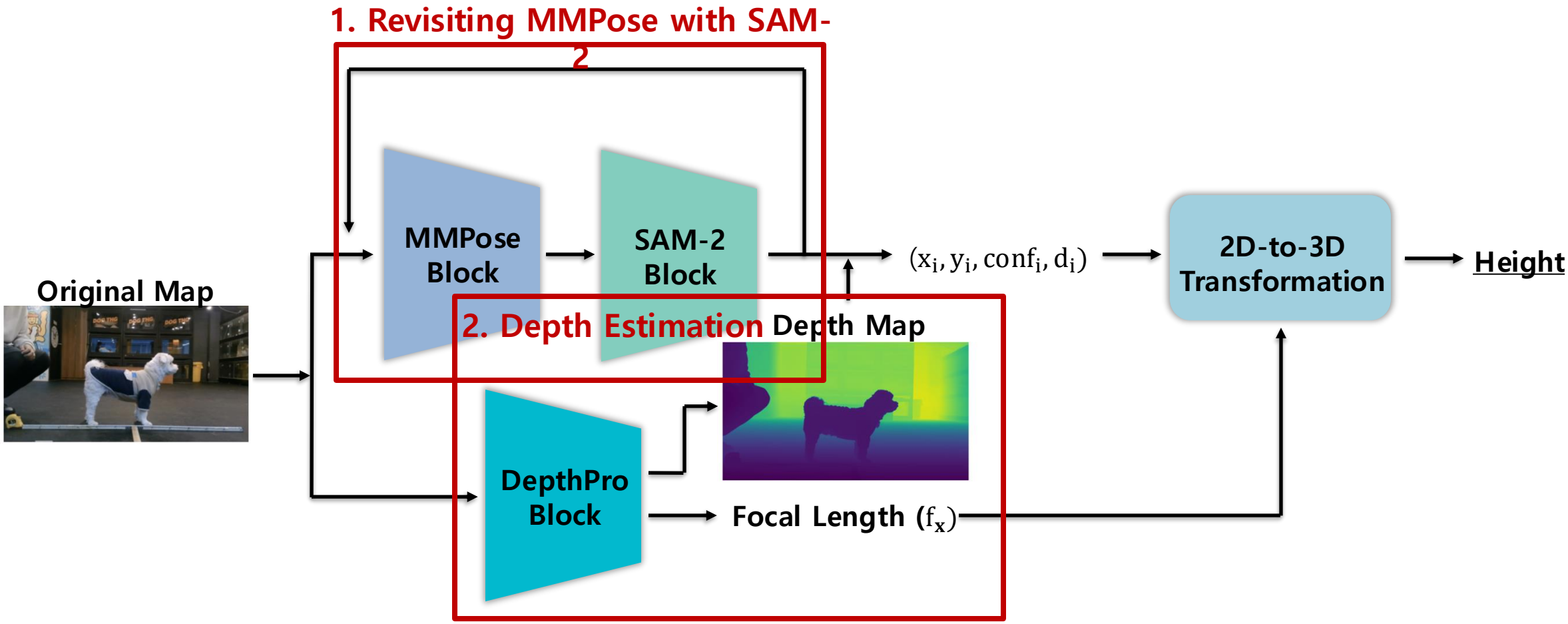


Figure 6. 본 프로젝트에서 제안하는 개선된 Framework. 크게 SAM-2 [3] 를 이용해서 MMPose의 성능 개선을 진행하고, DepthPro [4] 모델을 도입하여 Depth Estimation 모델의 성능 향상을 목표로 함.

Revisiting MMPose with SAM-2

Limitations of MMPose

- MMPose의 경우 다양한 이미지들에서 상당히 낮은 정확도를 보임.
- 사진이 촬영된 각도, 단모/장모견 여부에 따라 실패 비율을 비교해보았을 때, 유의미한 차이를 보이진 않으나 주로 장모견, rotated image에서 더 높은 실패비율을 보임.

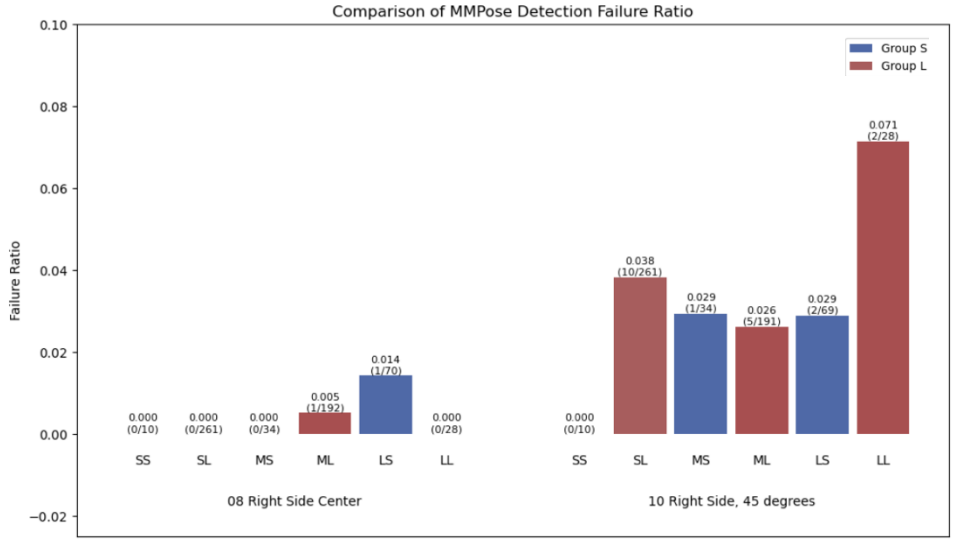


Figure 7. MMPose가 적절히 keypoint extraction에 실패한 예시.

Revisiting MMPose with SAM-2

Refinement via SAM-2

- Segmentation Model인 SAM-2를 이용해서 해당 Dog Segment 외부를 masking 하여 MMPose를 통한 Pose Extraction을 진행.

- 기존에 62.9 (±3.02) %에서 76.8 (±33.5) %로 높아진 Detection ACC를 보임. 다만, keypoint의 경우 72.0 (±20.2) % 에서 86.5 (±25.6) % 로 향상폭이 다소 낮음.

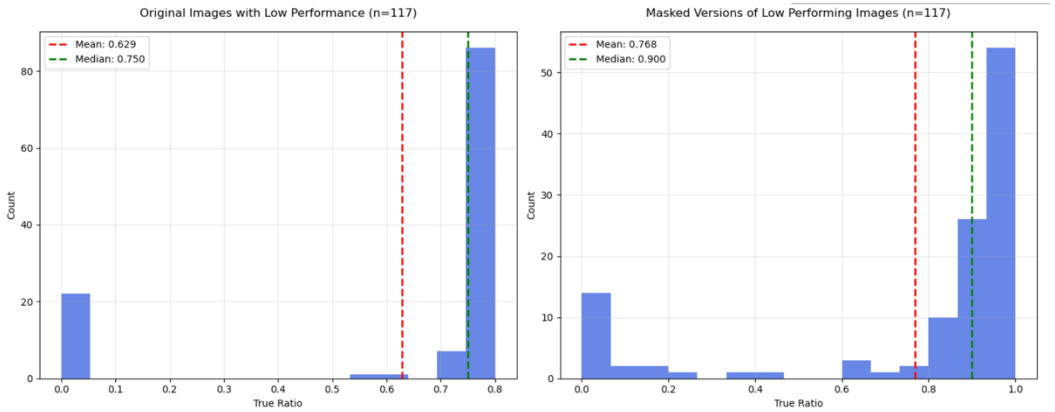
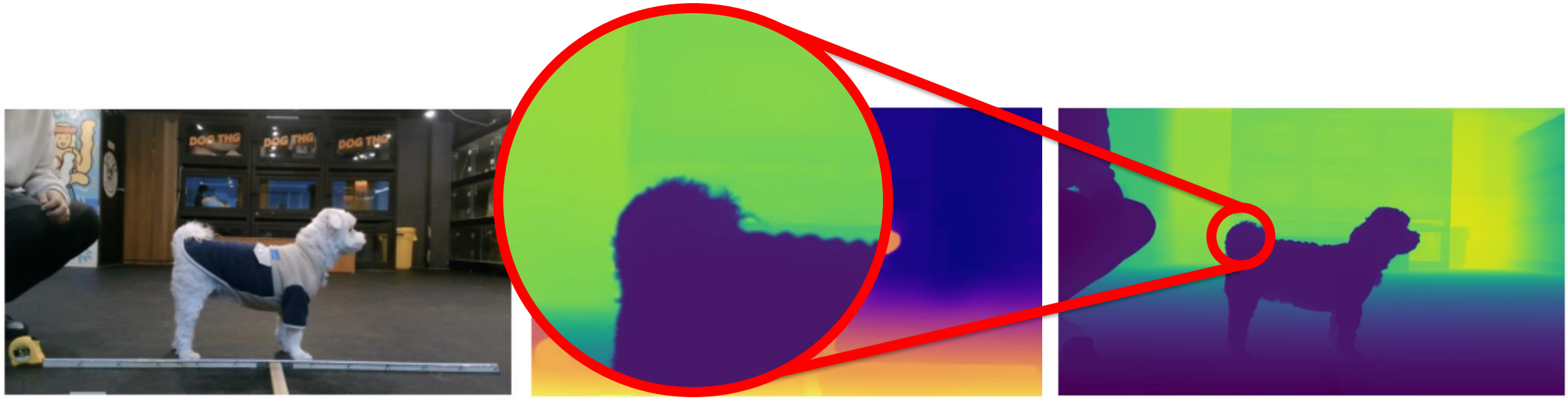


Figure 8. SAM-2 Masking을 통한 MMPose의 성능 개선. Masking 방식에 대한 예시와 (Top) SAM-2 Masking 이전과 이후의 결과 차이 (Bottom).

Depth Model Comparison



(a) Original Image

(b) Depth Anything v2

(c) Depth Pro

Figure 9. Depth Estimation 결과에 대한 비교. Depth Anything v2와 달리, Depth Pro의 경우 상당히 coarse한 부분까지 depth를 구분해내는 것을 확인할 수 있음.

Depth Model Comparison

(단위: cm)

Method	Depth of Neck	Depth of Rear	Length w/ model f_x	Length w/ given f_x
Ground Truth	115	115	-	-
Baseline (Depth Anything v2, NIPS 24)	+112.79 (± 47.79)	+104.78 (± 41.91)	-	+98.15 (± 67.15)
Metric 3D [5] (ICCV 23)	-25.16 (± 30.56)	-21.69 (± 31.58)	-	8.56 (± 6.60)
UniDepth [6] (CVPR 24)	+124.08 (± 92.77)	+155.63 (± 122.28)	+45.98 (± 52.09)	+155.88 (± 168.92)
Depth Pro (Under review)	+18.15 (± 30.05)	+19.52 (± 32.61)	-4.57 (± 11.69)	+34.39 (± 27.71)

Table 1. Depth Estimation Model에 따른 예측 성능 비교 (on 기업 데이터). 왼쪽에서부터 Neck Point의 Depth와 Rear Point의 Depth, 최종적으로 취득한 최종 Length 두 종류에 대한 L1 Loss 통계.

Depth Model Comparison

(단위: m)

Method	Length (cm)
Baseline (Depth Anything v2, NIPS 24)	+ 0.39 (±0.25)
Metric 3D (ICCV 23)	- 0.13 (±0.16)
UniDepth (CVPR 24)	+ 0.17 (±0.18)
Depth Pro (Under review)	+ 0.01 (±0.16)

Table 2. Depth Estimation Model에 따른 예측 성능 비교 (on AI-Hub 데이터). 최종적으로 취득한 최종 Length에 대한 L1 Loss 통계.

Patch-wise Depth Estimation

Intrinsic Noise of Depth Estimation

- Depth Estimation Model 자체는 pixel-wise estimation을 진행하기에 Noisy한 특성을 가짐.

→ Q. Depth 값을 단순 pixel이 아닌 Patch를 이용해서 연산을 진행하면 성능 향상을 기대할 수 있을까?

1. AVG in Patch
2. Mode Patch
3. 2D Spline Smoothing

→ Fig 10. 에서 나타나듯이 단일 Patch 내에서 유의미한 Depth value의 차이를 살피기 힘들어 기각.

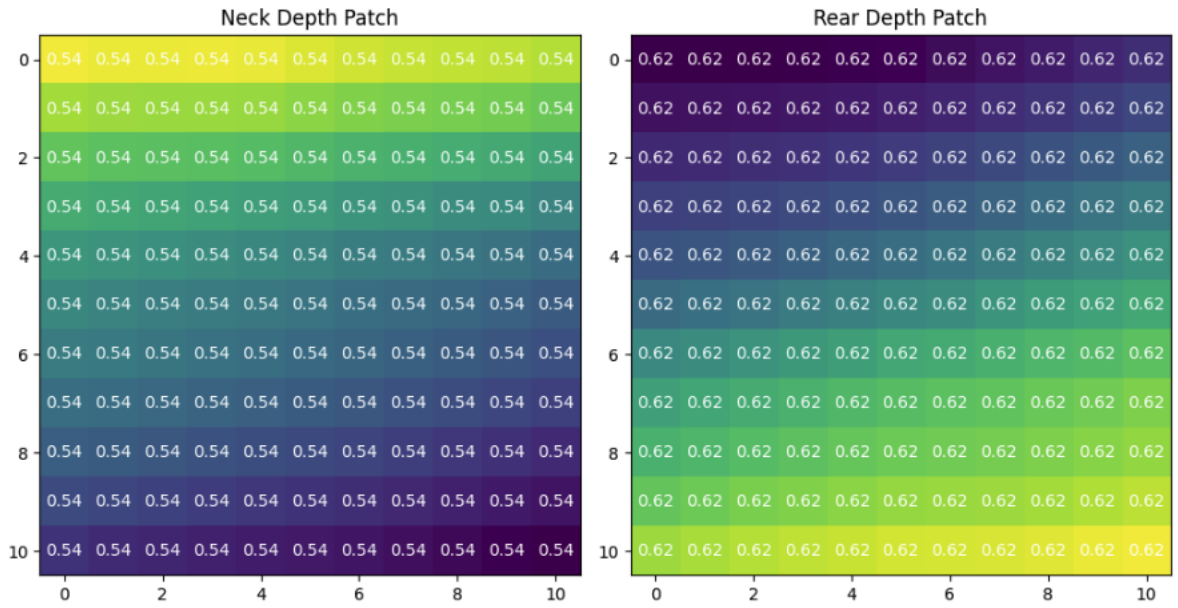
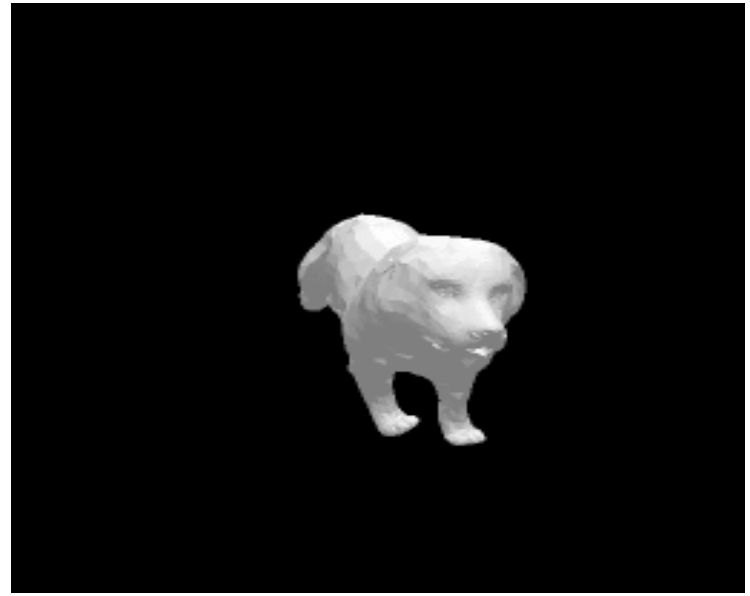


Figure 10. Patch 단위로 살펴본 Depth Map의 수치. Cm 단위에서 유의미한 차이를 살펴보기 힘듦.

3D Reconstruction: BARC

(a) Original Image

(b) 3D Mesh on Image



(c) Visualization of 3D Mesh

Figure 11. 3D Reconstruction 모델 BARC [7] 을 활용하여 단일 이미지로부터 복원한 개의 3D Mesh. Ground Truth가 존재하지 않기에 수치적으로 확인하긴 어려우나, 육안으로 평가 시 상당히 유사하게 Mesh를 생성해냄을 확인할 수 있음.

3D Reconstruction: BARC

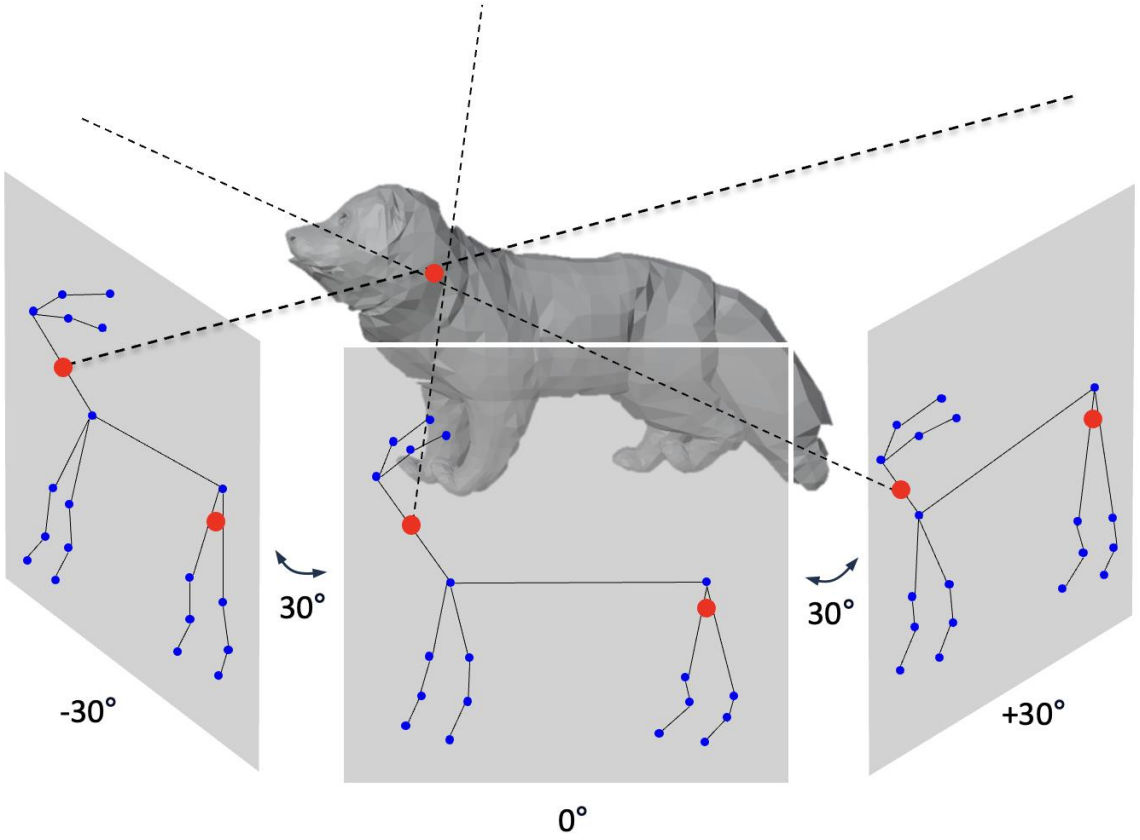


Figure 12. 3D Mesh에서의 Length Estimation 방식. 3개의 서로 다른 각도에서 측정한 MMPose Keypoint들을 기반으로 3D space에서의 keypoint 위치를 근사.

(On-gong Work)

Method	Length (cm)
Baseline (Depth Anything v2)	+0.981 (± 0.672)
BARC	-0.034 (± 0.131)

Table 3. BARC를 사용한 Length Estimation 결과 (on 기업 데이터).

Key Contributions

1. 기존에 사용하던 Depth Anything v2 모델을 개선함으로써 PetNow 데이터 기준 4cm 내외, AI-Hub 데이터 기준 1cm 내외의 오차 범위로 성능 개선.

2. MMPose가 keypoint extraction에 실패하는 문제를 해결하기 위해 SAM-2 모델을 이용해 보완.

3. 기존 기업의 pipeline과는 완전히 다른 3D Reconstruction-based 방법론인 BARC를 이용해 시도.

Future Works

1. BITE [8], MagicPony [9] 등의 추가적인 3D Reconstruction 모델을 이용해 성능 비교 및 개선.

2. 2D-to-3D Transformation 수식에서 focal length와 principal point 외에 다양한 camera parameter들의 적용 및 개선 가능성에 대한 확인.

3. Validation Set 외의 추가적인 데이터를 이용한 성능 비교 및 개선점 탐색.

4. (On-going) 3D Mesh에서 MMPose로 추출한 keypoint를 3D space에서의 점으로 근사.

References

[1] MMPose Github Repo, https://github.com/open-mmlab/mmpose

[2] Yang et al. Depth Anything V2 (NeurIPS 2024)

[3] Nikhila et al. SAM-2: Segment Anything in Images and Videos (arXiv Preprint)

[4] Aleksei et al. Depth ProL Sharp Monocular Metric Depth in Less Than a Second (arXiv Preprint)

[5] Wei et al. Towards Zero-shot Metric 3D Prediction from a Single Image (ICCV 2023)

[6] Piccinelli et al. UniDepth: Universal Monocular Metric Depth Estimation (CVPR 2024)

[7] Nadine et al. BARC: Learning to Regress 3D Dog Shape from Images by Exploiting Breed Information (CVPR 2022)

[8] Nadine et al. BITE: Beyond Priors for Improved Three-D Dog Pose Estimation (CVPR 2023)

[9] Shangzhe et al. MagicPony: Learning Articulated 3D Animals in the Wild (CVPR 2023)

Thank you!

Yonsei DataScienceLab
dslab.yonsei@gmail.com