DEVIEW 2020

데이터 라벨링 너무 귀찮아요 : 컨센서스 라벨링 도입기

신윤식 AIMMO

이 세션에서 다루는 것

컨센서스 라벨링?
당시의 상황과 결정까지의 과정
시도했던 방법들과 시행착오 내용

이 세션에서 다루지 않는 것

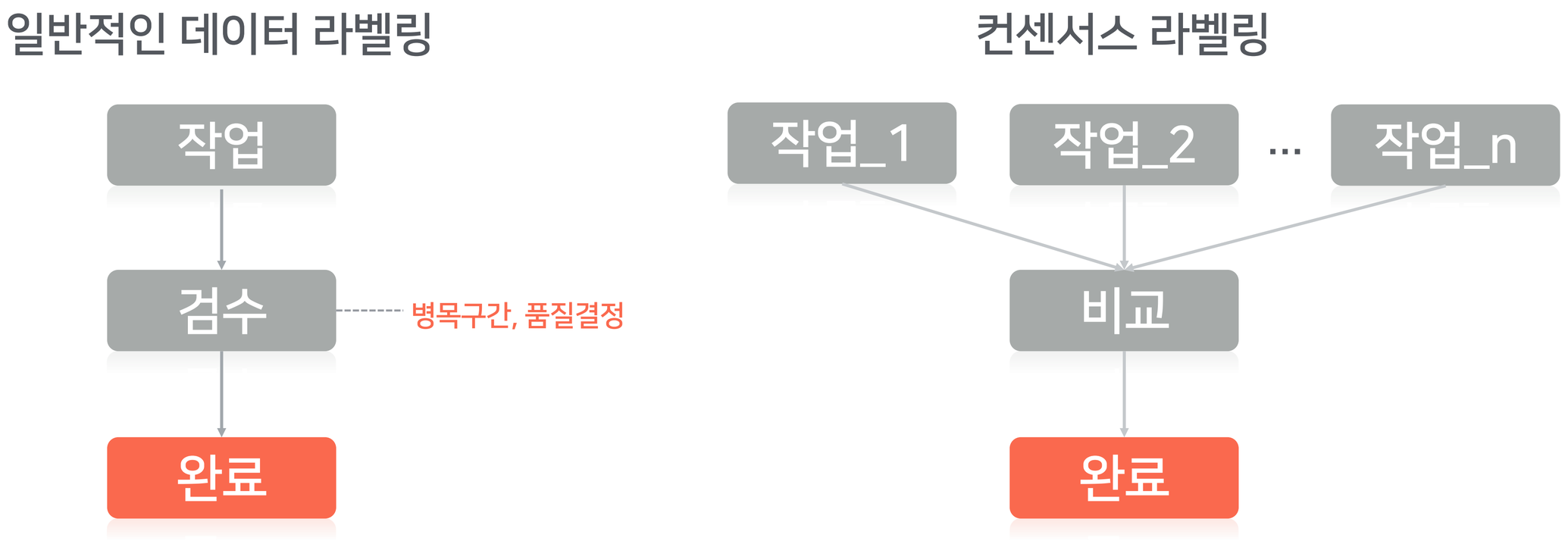
딥러닝
컨센서스 라벨링 외 다른 라벨링 방법

CONTENTS

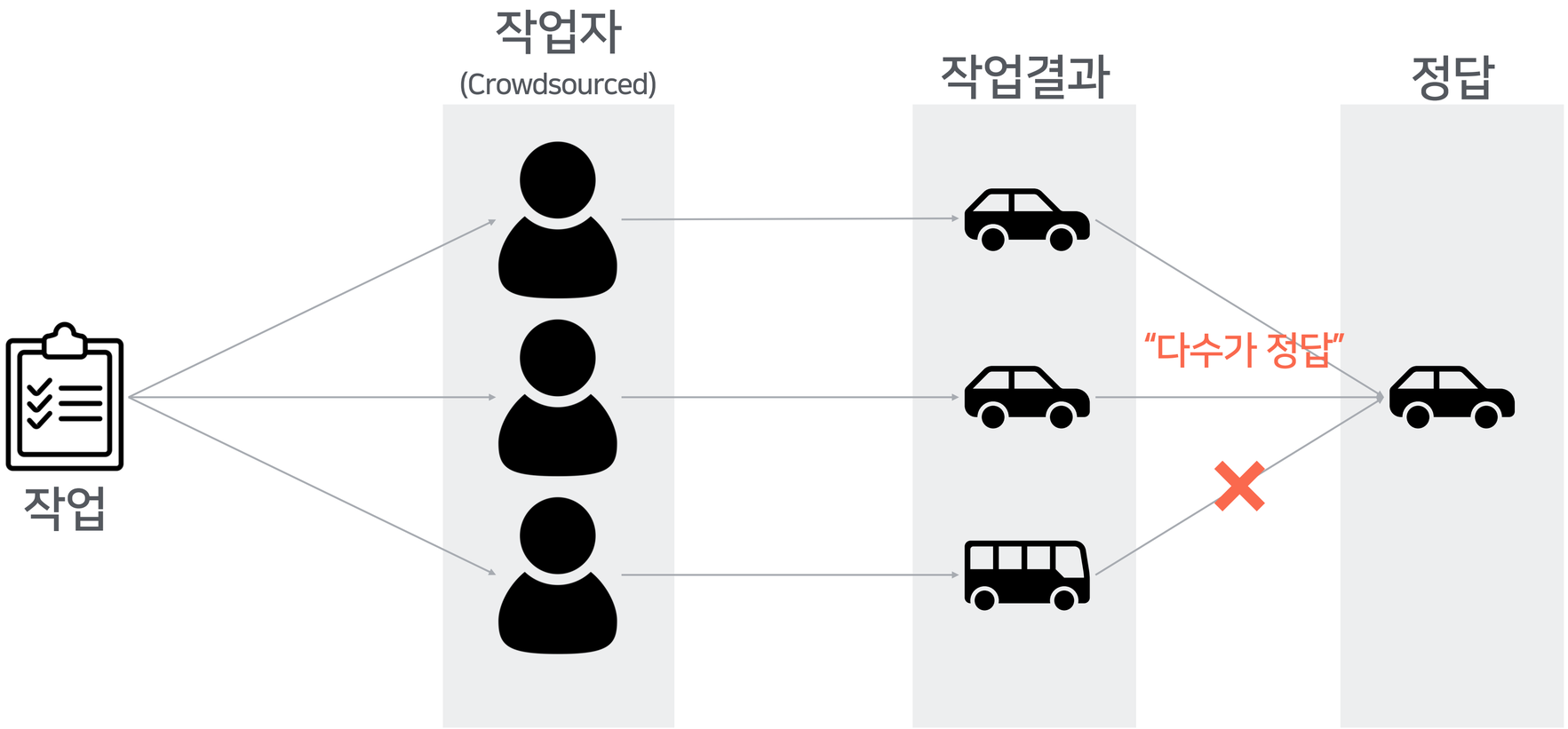
1. 컨센서스 라벨링

1.1 컨센서스 라벨링이란?

합의, 투표 기반의 정답(Ground Truth) 도출 방식



1.1 컨센서스 라벨링이란?



1.2 왜 컨센서스 라벨링을 사용하는가?

DEVIEW 2020

장점: 검수 병목현상 해소

데이터 라벨링 프로젝트에서의 병목현상

1. 작업 수 만큼 확인작업 발생
2. 작업품질에 따라 수정, 재확인 작업 추가발생
3. 작업자 수와 검수자 수의 불균형으로 병목발생

단점: (작업)비용

2. 라벨링 프로젝트 주요 이슈

2.1 컨센서스 라벨링을 선택한 이유

10명 문서 100만건

운영, 관리비용 최소화 필요!

작업자 관리는 누가? → 크라우드 소싱

검수(Quality Control)는 누가? → 크라우드 소싱?

2.1 컨센서스 라벨링을 선택한 이유

1. 간단하게 끝날 양이 아니었습니다.
2. 현실적으로 운영, 관리에 쏟을 자원이 없었습니다.
3. 학습데이터는 정확하길 바랐습니다.

2.2 라벨링 프로젝트 주요 이슈

DEVIEW 2020

작업

- 작업 운영부담
- 작업자 모집 및 교육
- 우수 작업자 선별

검수

- 라벨링 품질관리 부담
- 작업자에서 전환하는 문제
- 작업 일관성을 보장하는 문제

2.2 라벨링 프로젝트 주요 이슈

DEVIEW 2020

그렇게 고민할 일일까?

직접 해야하는 상황...

누가 필요할까? 연구개발팀 (10명)

얼마나 필요할까? 엄청나게 많이 (COCO 데이터셋)

전체의 1%만 살펴봐도 2,000장 이상

2.2 라벨링 프로젝트 주요 이슈

연구개발팀

연구, 개발

데이터 라벨링

운영, 관리

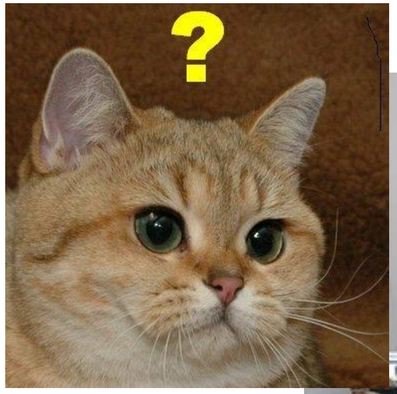
추가로 채용하기엔 애매하고…

자, 누가 해볼래?

2.3 주요 이슈 1: 작업 운영부담

DEVIEW 2020

작업은 크라우드소싱(Crowdsourcing)으로!



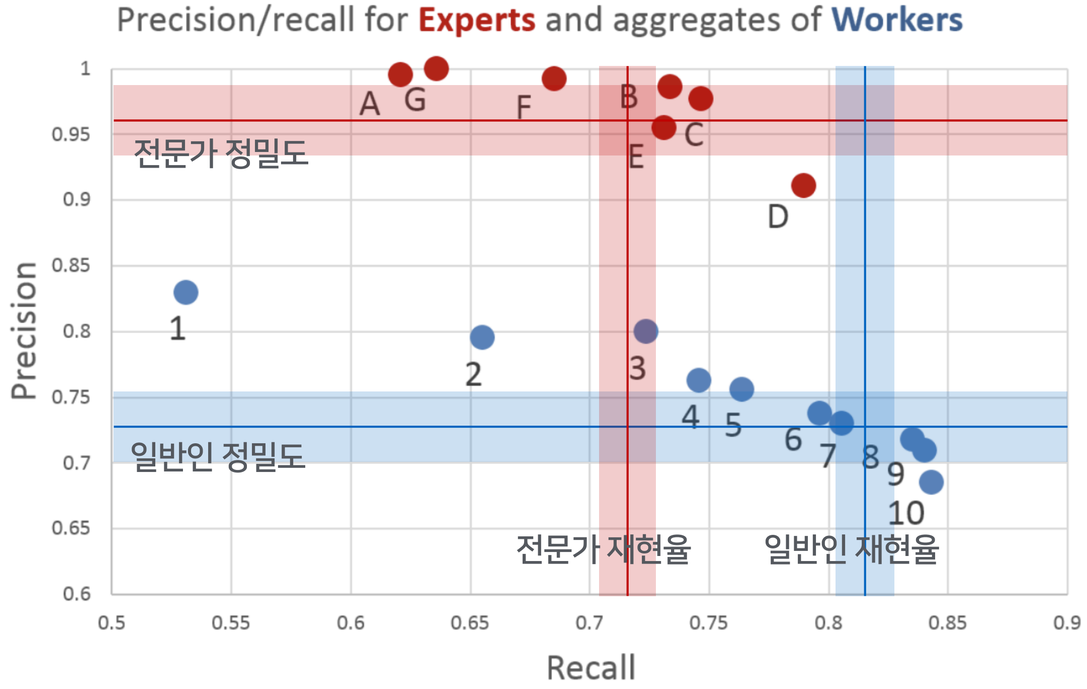
Ref.: NYT, “How Cheap Labor Drives China’s A.I. Ambitions”(2018.11.25)

2.3 주요 이슈 1: 작업 운영부담

전문가, 일반인 그룹의 작업특성 차이(MS Coco Dataset)

Ref.: Chen, Xinlei, et al. "Microsoft COCO captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325(2015).

이미지(328,000장)내 객체 카테고리 표시 작업



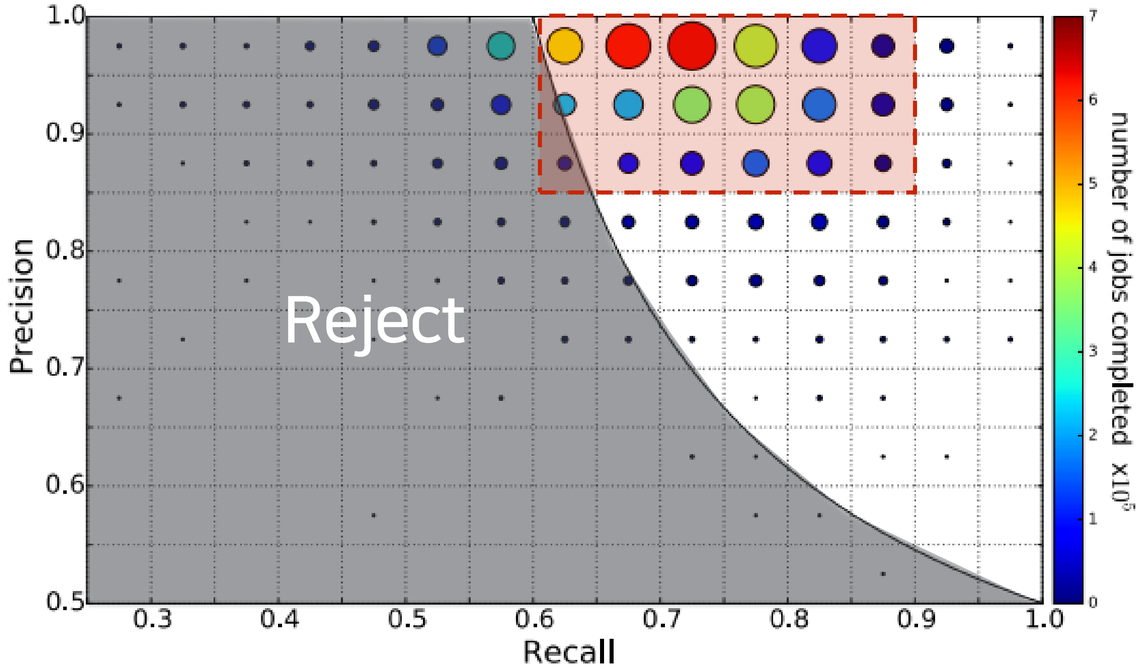
* 일반인 그룹: Amazon Mechanical Turk에서 모집된 Crowd

2.3 주요 이슈 1: 작업 운영부담

DEVIEW 2020

전문가, 일반인 그룹의 작업특성 차이(MS Coco Dataset)

Ref.: Chen, Xinlei, et al. "Microsoft COCO captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325(2015).

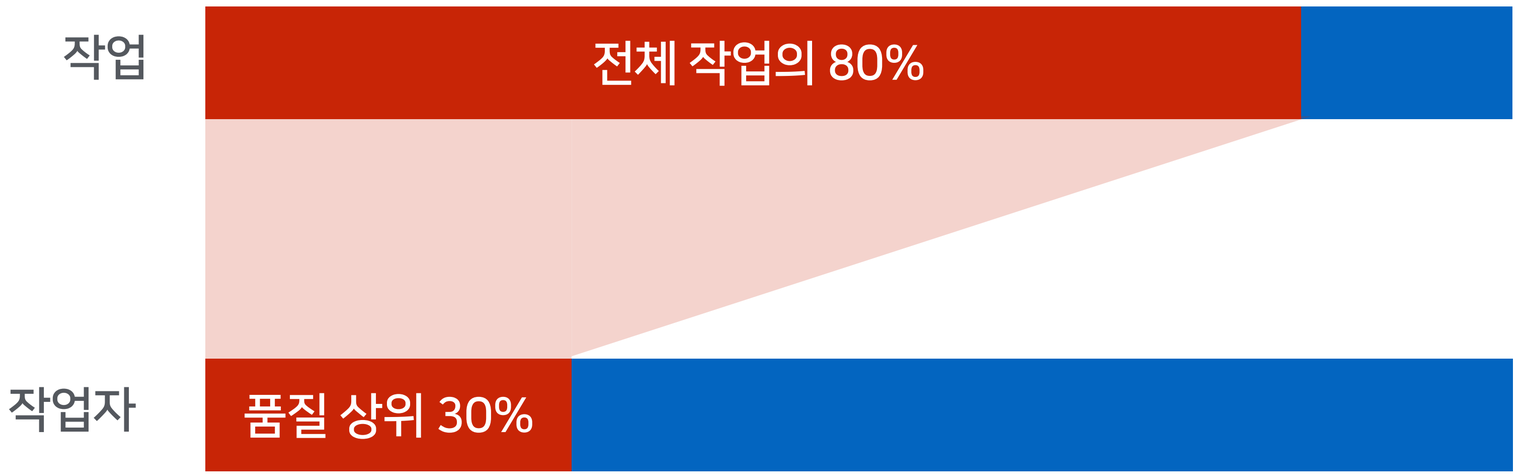


1. Crowdsourcing으로도 High F1-score 도달 가능
2. 생각보다 많은 작업자가 High Precision에 도달
3. 많은 작업이 고수준 작업자에 의해 완료

2.3 주요 이슈 1: 작업 운영부담

잘하는 소수가 대부분의 작업 완료

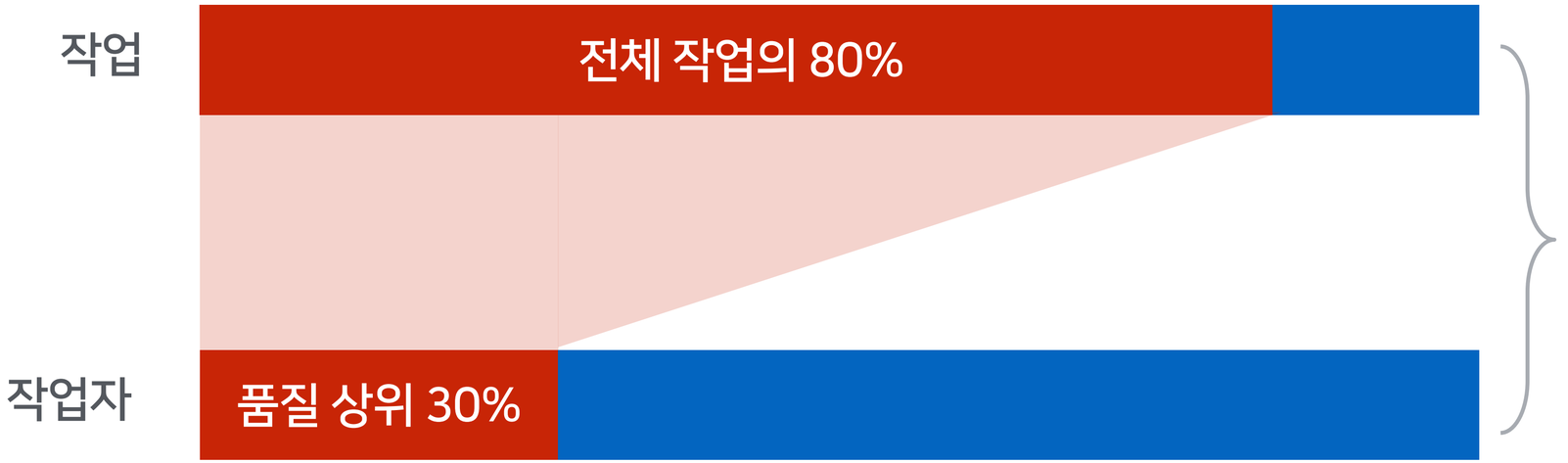
실제 서비스에서의 작업 완료 비중



2.3 주요 이슈 2: 검수 품질관리

DEVIEW 2020

방법: 품질 상위 작업자를 검수자로 전환



상위 10% 전환 시,

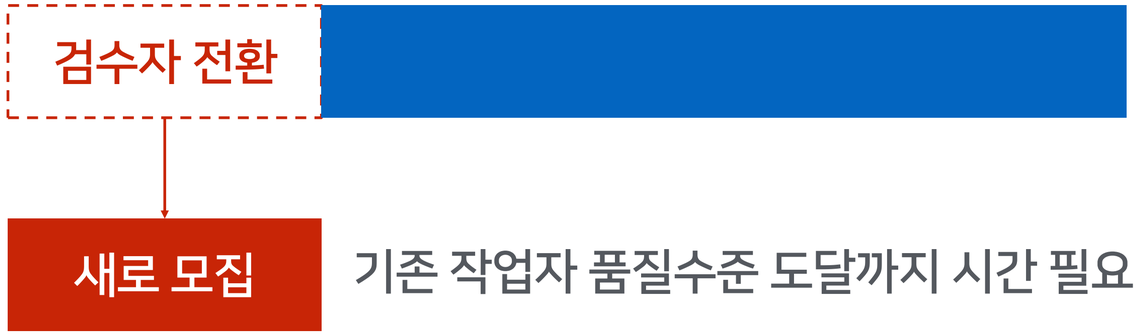
전체 작업량 24% 감소

+ 고수준 작업자 감소로 **전체적 품질하락** 초래

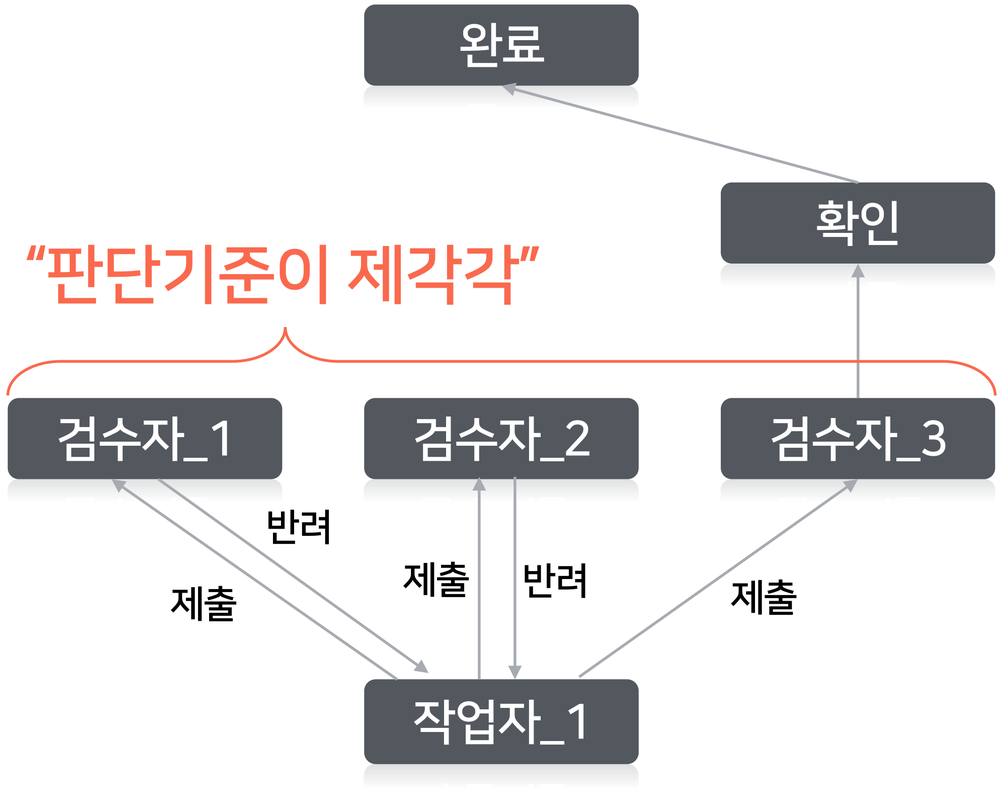
+ 많은 수의 검수자로 인한 **반복검수 발생**

2.3 주요 이슈 2: 검수 품질관리

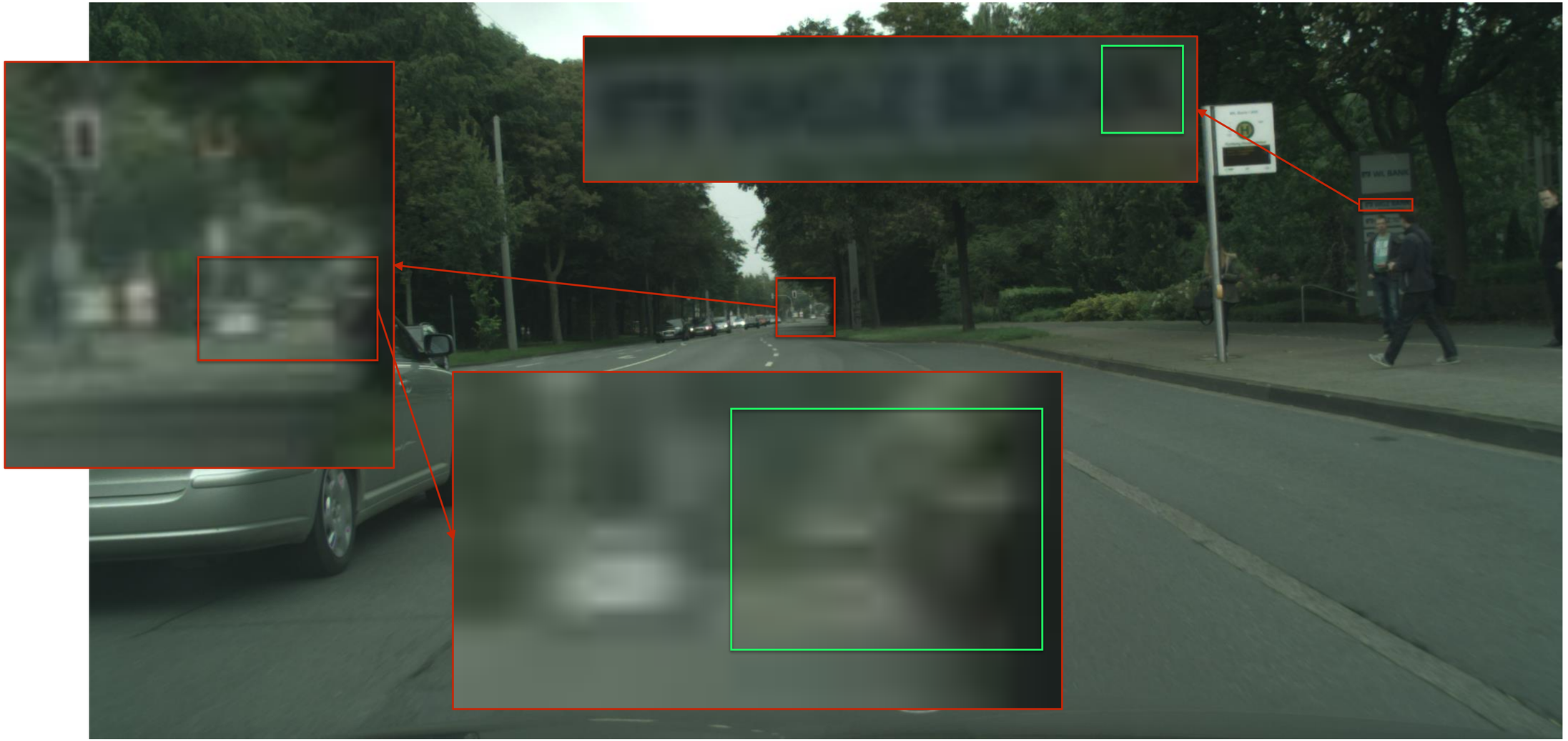
고수준 작업자 감소



반복검수 발생



2.3 주요 이슈 2: 검수 품질관리

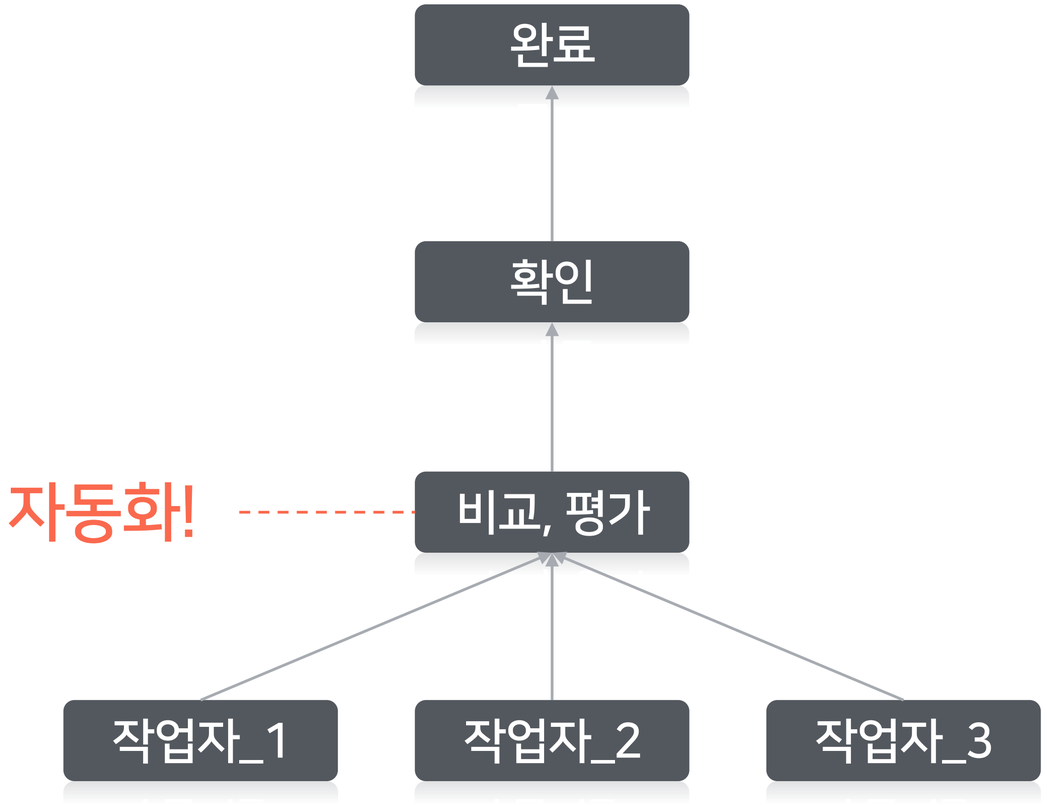


Ref.: M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele,
"The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

우리의 상태…


"아무것도 안하고 싶다"
"이미 아무것도 안하고 있지만 더 격렬
하고 적극적으로 아무것도 안하고 싶다"

선택한 방법


완료
확인
자동화!
비교, 평가
작업자_1
작업자_2
작업자_3

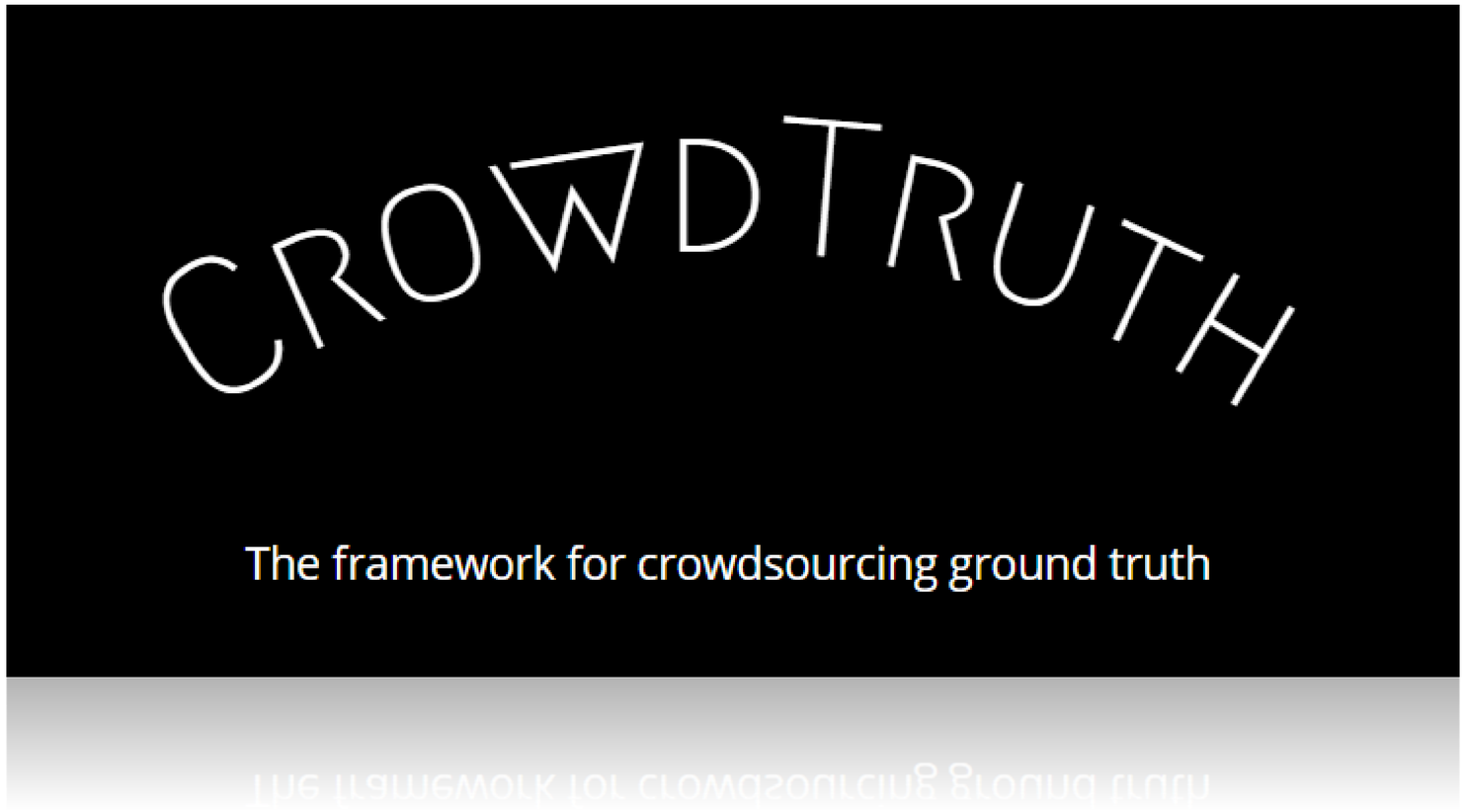
3. 컨센서스 라벨링 자동화: SQIP

3.1 참고사례: Crowd Truth

DEVIEW 2020

Measuring Crowd Truth: Disagreement Metrics Combined with Worker Behavior Filters

Ref.: Soberón, Guillermo, et al. "Measuring crowd truth: Disagreement metrics combined with worker behavior filters." CrowdSem 2013 Workshop. Vol. 2. 2013.



Ref.: http://crowdtruth.org/

3.1 참고사례: Crowd Truth

DEVIEW 2020

주로 자연어 관련 라벨링 프로젝트에 활용

Data processed with CrowdTruth 2.0

- **NYT – Topical Relevance** – Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, Lora Aroyo (2018).
- **Causal Relations between Events** – Tommaso Caselli, Oana Inel (2018).
- **Frame Disambiguation** – Anca Dumitrache, Lora Aroyo, Chris Welty (2018).
- **Open Domain Relation Extraction** – Anca Dumitrache, Lora Aroyo, Chris Welty (2017).

Data processed with CrowdTruth 1.0

- **Medical Relation Extraction** – Anca Dumitrache, Lora Aroyo, Chris Welty (2015). *DOI: 10.5281/zenodo.31890*
- **VU Sound Annotation Corpus** – Emiel van Miltenburg, Benjamin Timmermans, Lora Aroyo (2015).
- **Salience in News and Tweets** – Oana Inel, Tommaso Caselli, Lora Aroyo (2015). *DOI: 10.5281/zenodo.46477*
- **Crowdsourcing Events in Videos** – Robert Iepsma, Theo Gevers, Zoltan Szlavik and Lora Aroyo (2016).
- **Crowdsourcing Named Entities Gold Standards** – Oana Inel, Lora Aroyo (2017). *DOI: 10.5281/zenodo.235452*

Ref.: http://crowdtruth.org/

3.1 참고사례: Crowd Truth

DEVIEW 2020

작업 예시

작업대상 문장

In the sentence: "We studied mononuclear cell (MNC)-mediated natural killing (NK) of [VARICELLA]-zoster [VIRUS] (VZV)-infected fibroblasts in normal children, children with VZV infections, and children with Hodgkin's disease."

Is [VARICELLA] ----*related-to*---- [VIRUS]?

질문

STEP 1: Select the valid RELATION(s)

- [ ] [TREATS]
- [ ] [PREVENTS]
- [ ] [DIAGNOSED_BY_TEST_OR_DRUG]
- [ ] [PART_OF]
- [ ] [OTHER]
- [ ] [NONE]
- [ ] [CAUSES]
- [ ] [LOCATION]
- [ ] [SYMPTOM]
- [ ] [MANIFESTATION]
- [ ] [CONTRAINDICATES]
- [ ] [ASSOCIATED_WITH]
- [ ] [SIDE_EFFECT]
- [ ] [IS_A]

답변 1

답변 2

STEP 2a: Copy & Paste ONLY the words from the SENTENCE that express the RELATION you selected in STEP1

Answer N/A if you selected [NONE] in

Copy & Paste from the sentence ONLY the words that express the RELATION you have selected in STEP1. DO NOT copy the whole sentence.

STEP 2b: If you selected [NONE] in STEP 1, explain why

Answer N/A if you have selected a

If you think there is a relation between those two words, but it is different than any of the relations in STEP 1, then type the relation here. If you think there is no relation between those terms, explain why do you think it is.

Ref.: Soberón, Guillermo, et al. "Measuring crowd truth: Disagreement metrics combined with worker behavior filters." CrowdSem 2013 Workshop. Vol. 2. 2013.

3.1 참고사례: Crowd Truth

	Worker	Sentence	Relation
Representation	각 단위를 벡터로 표현		
Metrics	각 단위 간 유사도를 측정		

* 각 작업자의 결과차이를 기반으로 모호한 정도를 측정

Representation

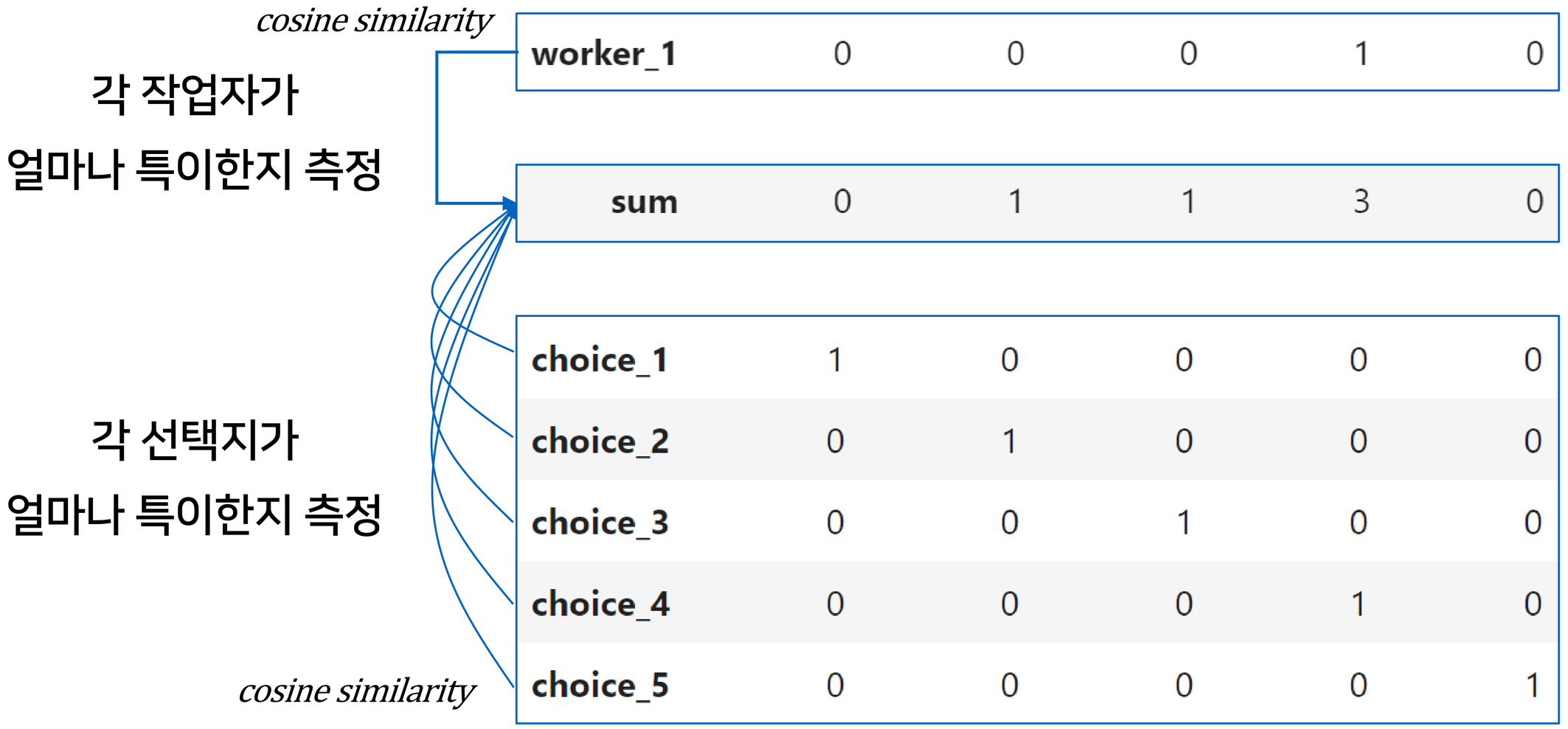
Relation_1 Vector

Worker_1 Vector

Sentence Vector

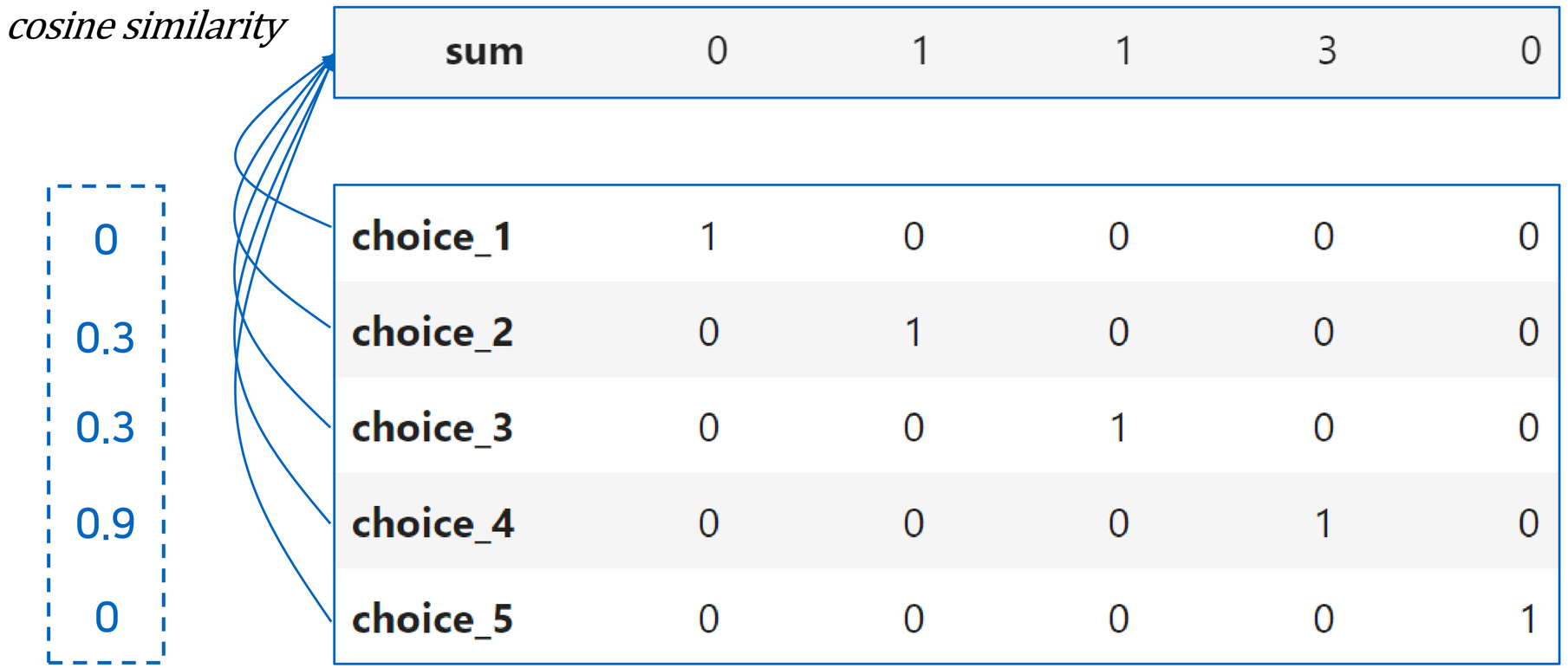
	choice_1	choice_2	choice_3	choice_4	choice_5
worker_1	0	0	0	1	0
worker_2	0	0	0	1	0
worker_3	0	1	0	0	0
worker_4	0	0	1	0	0
worker_5	0	0	0	1	0
sum	0	1	1	3	0

Metrics



worker_1	0	0	0	1	0
sum	0	1	1	3	0
choice_1	1	0	0	0	0
choice_2	0	1	0	0	0
choice_3	0	0	1	0	0
choice_4	0	0	0	1	0
choice_5	0	0	0	0	1

Metrics



max_value = 0.9

"작업자 간 이견이 클수록 감소"

이미지에 맞춰서 써보자!

3.2 컨센서스 라벨링 자동화 개요

DEVIEW 2020

SQIP

발명자: 신윤식, 홍기섭

(Statistical Quality Inference Protocol)

라벨링 품질관리 작업 최소화, 반자동화

3.2 컨센서스 라벨링 자동화 개요

요소별 품질 저해요인과 특징

품질 저해요인

- 모호한 가이드
- 모호한 선택지,
- 인지하기 어려운 이미지 등

품질 저해요인

- 의도적: Spammer, Bot 등
- 비의도적: 잘못된 가이드 이해

종속적 특징

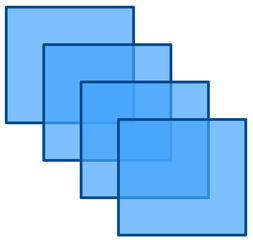
- 정답 없음.
- 다수(Majority)의 선택이 정답
- 작업결과로부터 추론 필요

3.2 컨센서스 라벨링 자동화 개요

DEVIEW 2020

품질평가 개요(Segmentation 예시)

Mask



작업결과물 n개 획득

Alpha map

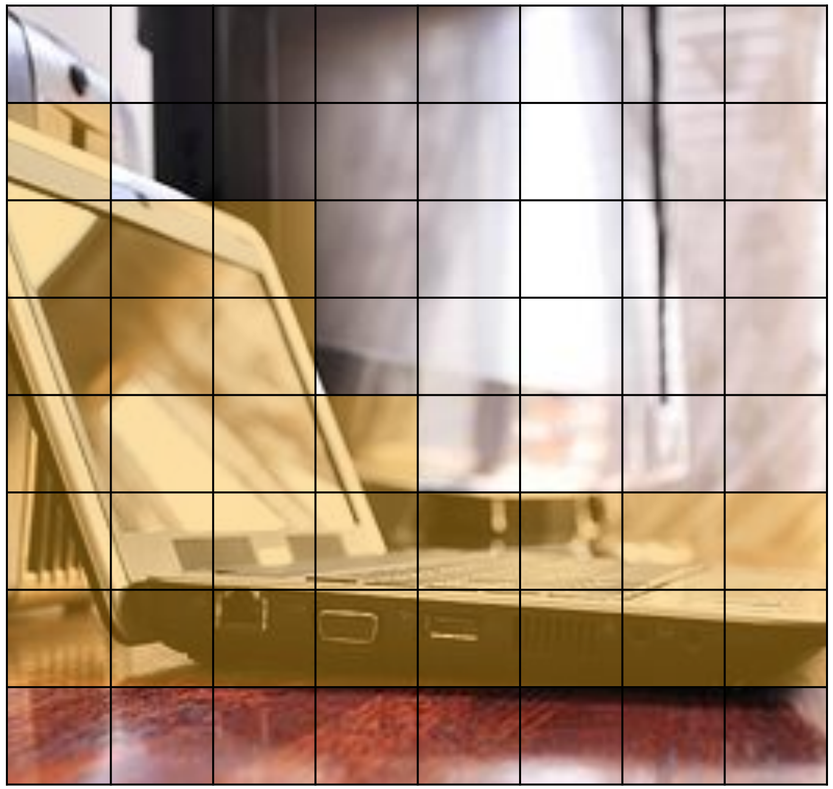
1개로 병합

Evaluation

병합결과 및 개별 결과물들을 비교평가

3.3 데이터의 표현

Binary Mask

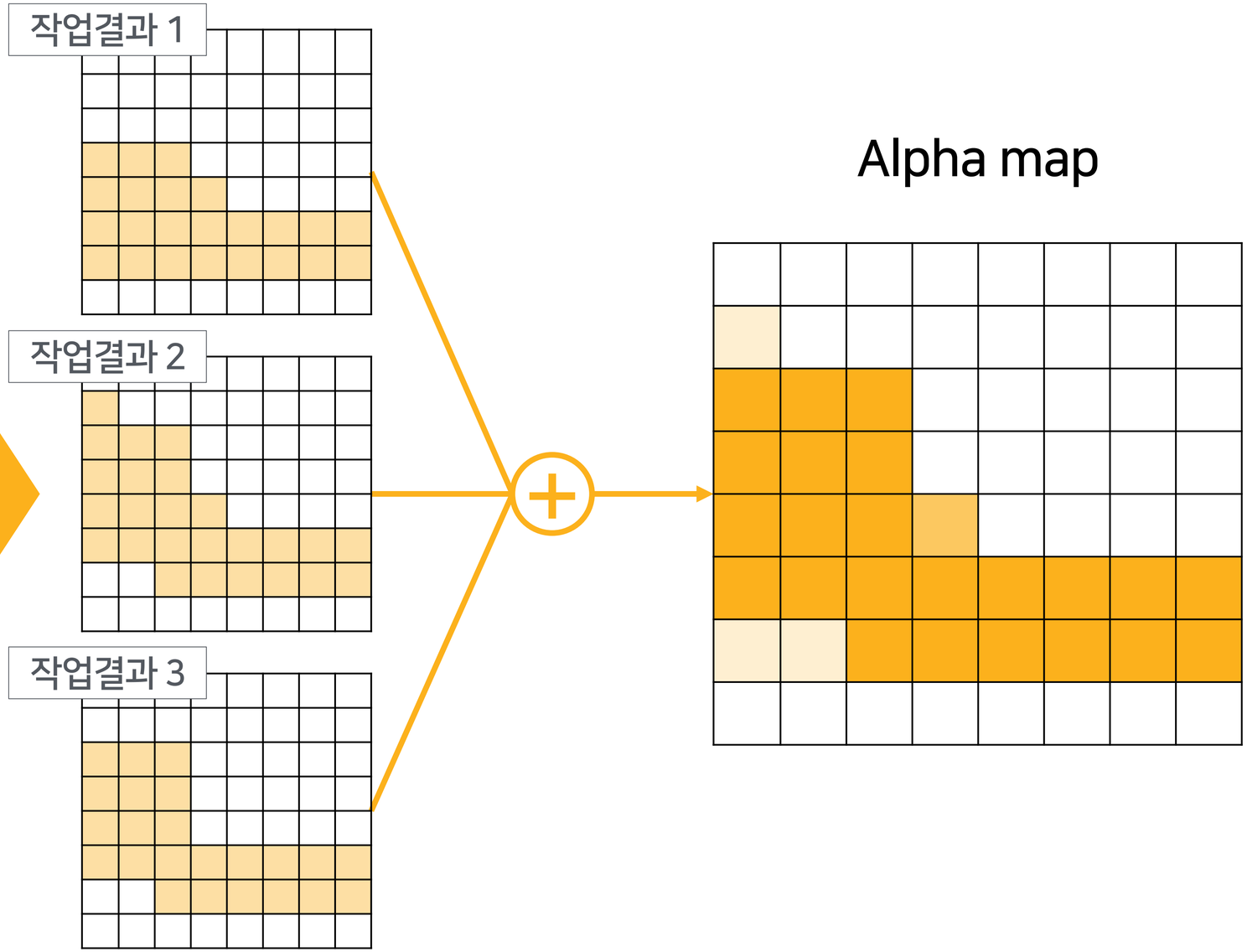


Ref.: MS COCO, https://cocodataset.org/

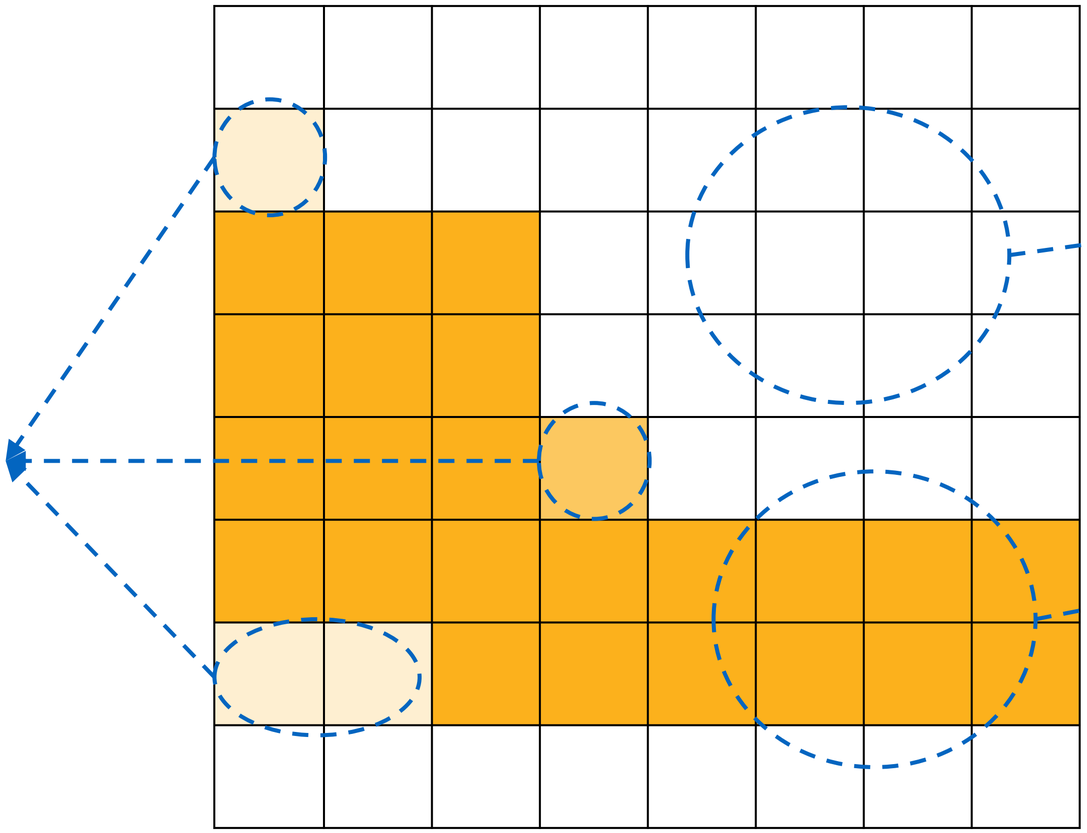
2D Matrix

0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0
1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0

Ref.: MS COCO, https://cocodataset.org/



Alpha map



Uncertain area

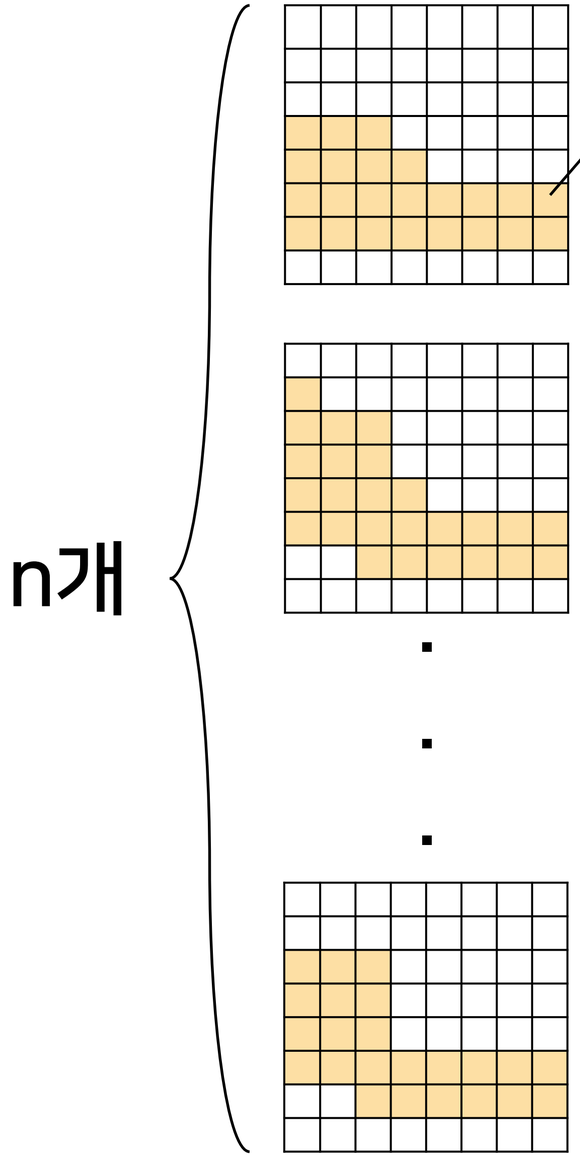
: 작업자별 이견 발생

Background

: 모든 작업자가 미선택

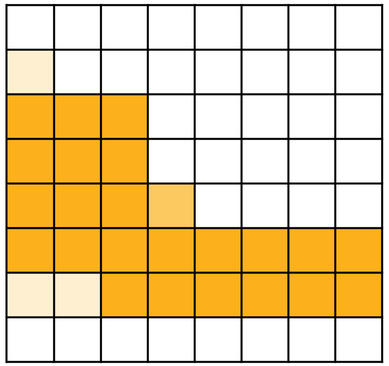
Foreground

: 모든 작업자가 선택



$$\mathbf{Alpha(pixel\ value)} = \frac{\mathbf{1}}{\boldsymbol{n}}$$

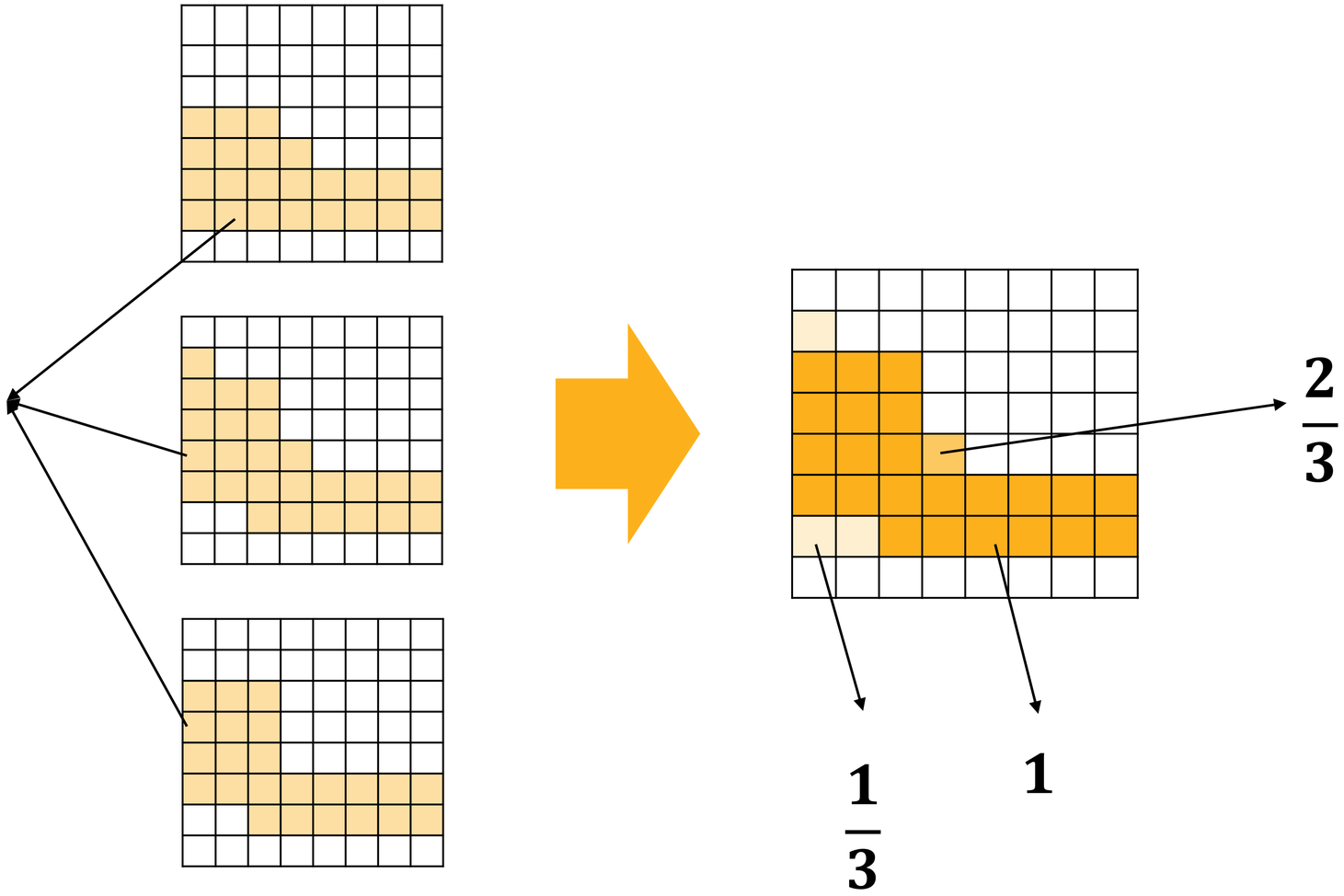
$$\boldsymbol{AlphaMap} = \sum \boldsymbol{AlphaMask}$$



- $\boldsymbol{Foreground} = \mathbf{1}$
- $\boldsymbol{Background} = \mathbf{0}$
- $0 < \boldsymbol{UncertainArea} < \mathbf{1}$

$$n = 3$$

$$\text{Alpha(pixel value)} = \frac{1}{3}$$



3.4 지표 정의

Image Task Metrics

- 모든 Metrics는 작업 단위로 비교
- AS(Alpha Score): 특정 작업결과에 대한 나머지 작업결과들의 일치도
- TWS(Task-Worker Similarity): 특정 작업결과와 AlphaMap의 유사도
- AMS(Alpha Map Score): 특정 작업결과와 AlphaMap의 일치도
- ICS(Image Clarity Score): 작업결과 간 일치도

명칭	비교단위	평가대상	의미
AS	작업	작업결과, 작업자	작업결과, 작업자 성향파악
TWS			높을수록 나머지 작업과 유사
AMS			
ICS		작업	높을수록 작업이 작업결과 이견 없이 명확

3.4 지표 정의

AS(Alpha Score)

- 의미: 특정 작업결과에 대한 나머지 작업결과들의 일치도
- 목적: 작업결과 및 작업자의 성향파악(과소표시, 과대표시)

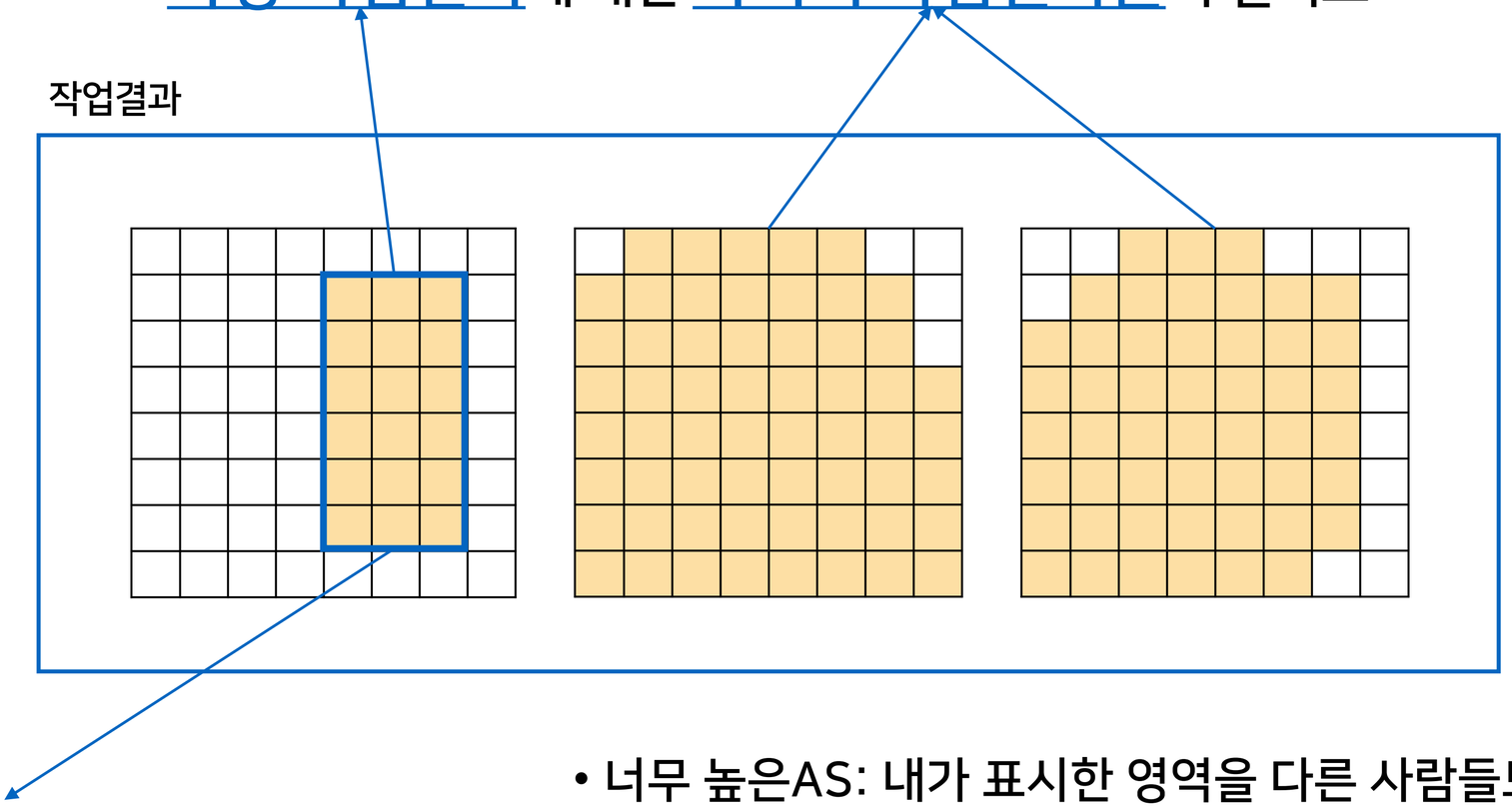
AlphaMask에서 선택된 작업자 표시영역의 값 총합

$$AS(t,i) = \frac{AlphaMap(t) \circ BM(t,i)}{\Sigma BM(t,i)}$$

선택된 작업자의 표시영역 면적

- $t = task\ number$
- $i = worker\ number$

특정 작업결과에 대한 나머지 작업결과들의 일치도

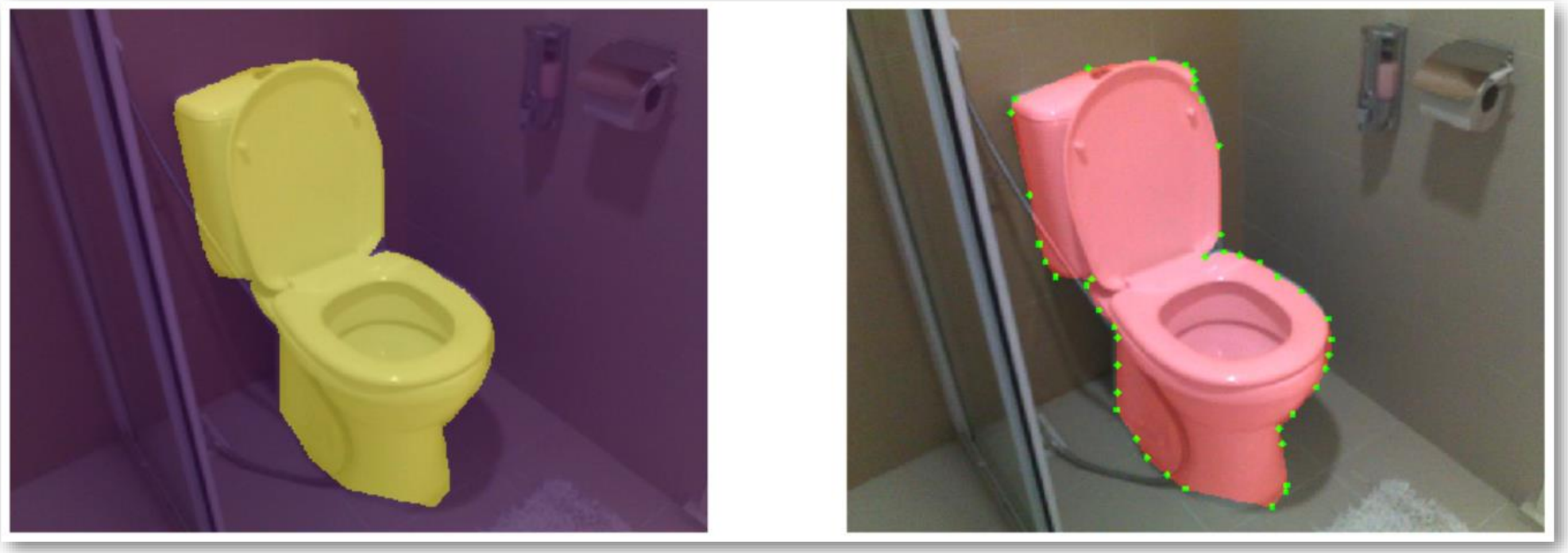


내가 표시한 부분을

다른 사람들도 표시했는지

- 너무 높은AS: 내가 표시한 영역을 다른 사람들도 대부분 표시
  → 과소표시 가능성(High Precision, Low Recall)
- 너무 낮은 AS: 내가 표시한 영역을 다른 사람들은 대부분 미표시
  → 과대표시 가능성(Low Precision, High Recall)

Ground Truth



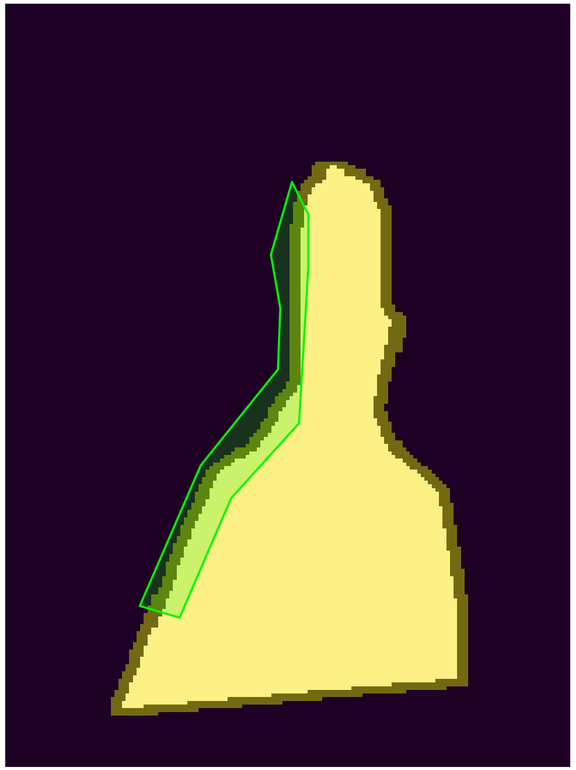
Ref.: MS COCO, https://cocodataset.org/

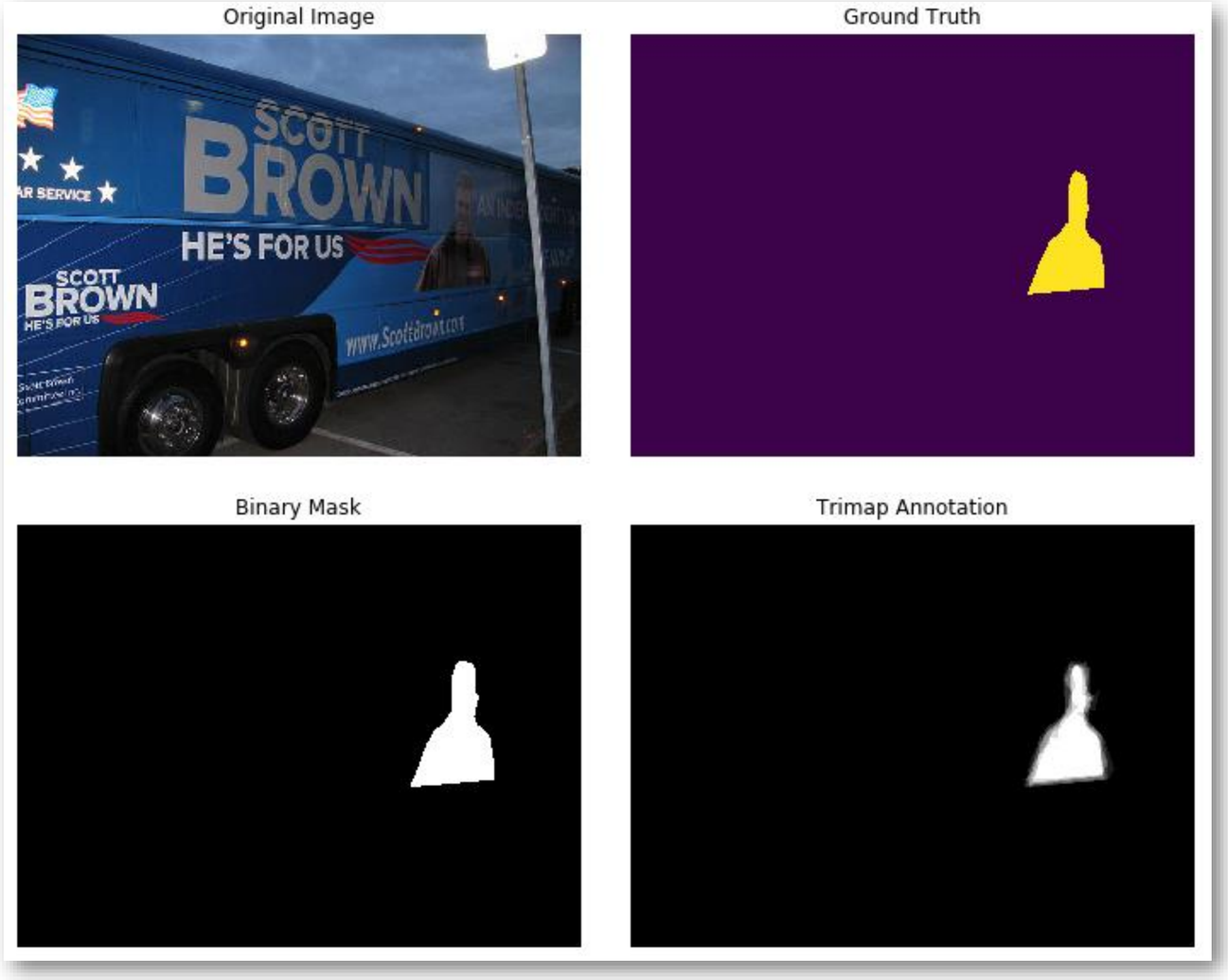
작업유형을 가정하여 데이터 생성

Ref.: MS COCO, https://cocodataset.org/

과대표시(Overbounded) 예시

AS: 0.830

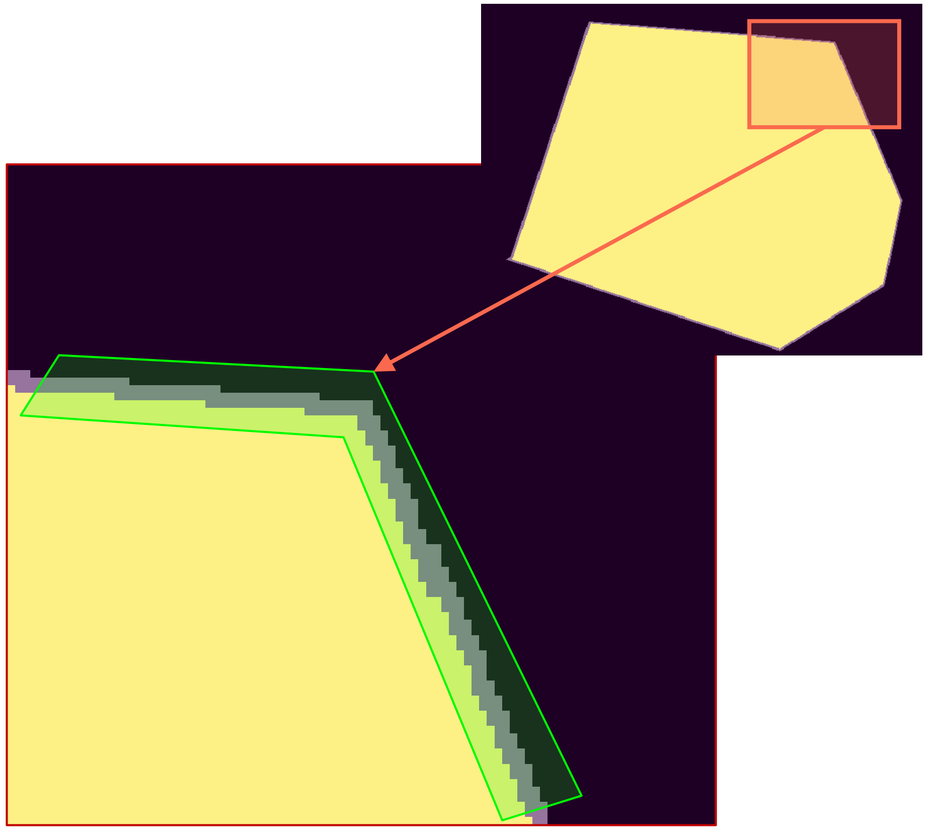


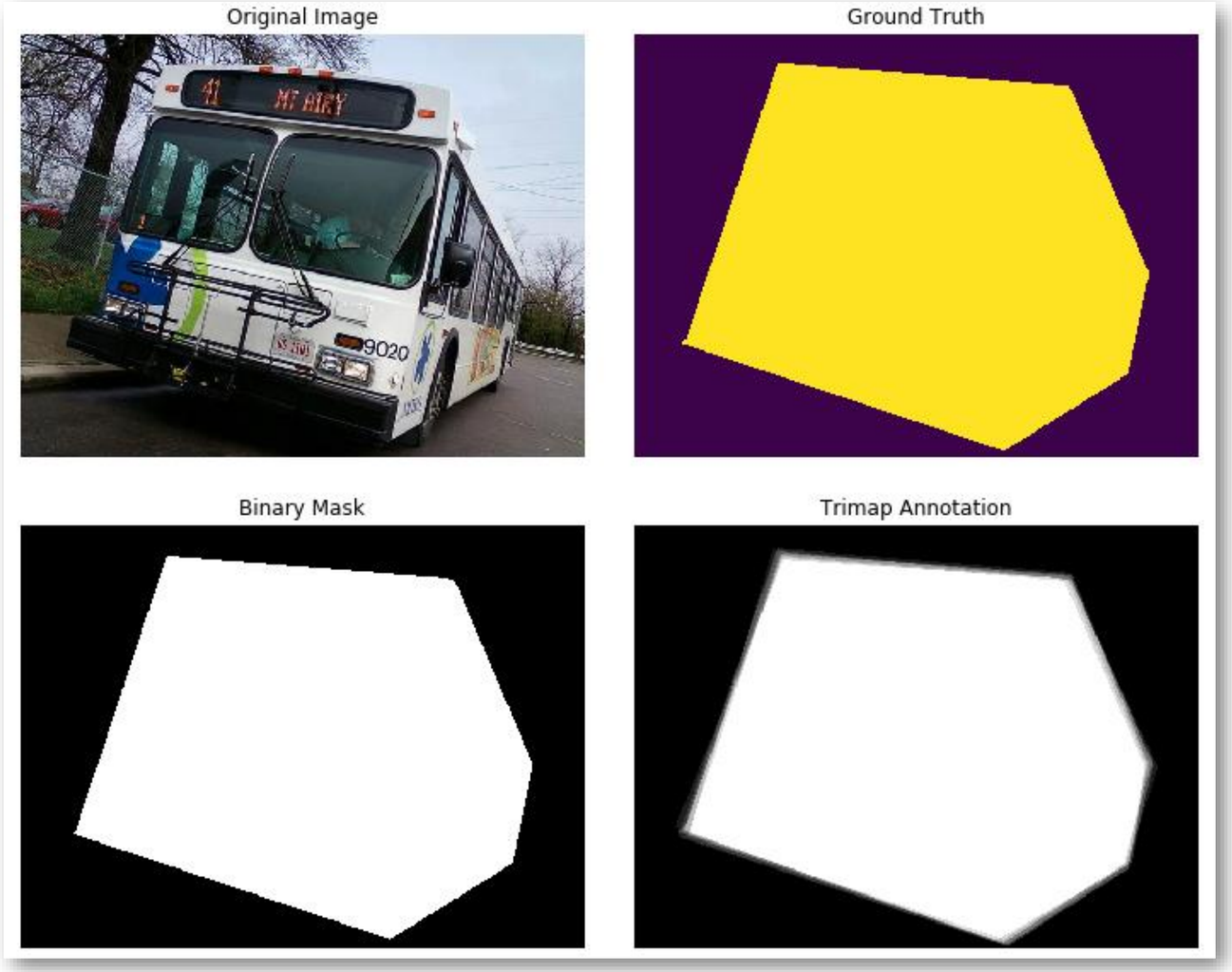


Ref.: MS COCO, https://cocodataset.org/

과소표시(Underbounded) 예시

AS: 0.997





Ref.: MS COCO, https://cocodataset.org/

작업유형에 따른 지표 차이(AS)

	Large_Jitter	Normal_Dilate	Normal_Erode	Normal_Jitter1	Normal_Jitter2	Normal_Jitter3	Normal_Jitter4	Superior
0	0.959	0.927	0.994	0.975	0.971	0.970	0.966	0.971
1	0.984	0.963	0.993	0.982	0.983	0.977	0.982	0.983
2	0.914	0.892	0.991	0.949	0.939	0.949	0.946	0.940
3	0.951	0.926	0.991	0.946	0.941	0.951	0.947	0.953
4	0.881	0.874	0.989	0.929	0.937	0.915	0.923	0.939
5	0.956	0.955	0.993	0.978	0.969	0.981	0.979	0.977
6	0.938	0.943	0.990	0.961	0.964	0.963	0.969	0.970
7	0.911	0.932	0.990	0.951	0.949	0.960	0.965	0.958
8	0.924	0.934	0.992	0.981	0.962	0.944	0.969	0.968
9	0.986	0.971	0.998	0.996	0.975	0.976	0.988	0.987
10	0.989	0.977	0.995	0.985	0.995	0.992	0.985	0.991
11	0.873	0.831	0.978	0.852	0.886	0.889	0.933	0.930

3.4 지표 정의

TWS(Task-Worker Similarty)

- 의미: 선택된 작업자가 다른 작업자들의 결과와 유사한 정도
- 목적: 일반적인 작업경향에서 얼마나 벗어났는지 판별

$$TWS(t,i) = \cos\left(V_t - V_{t,i}, V_{t,i}\right)$$

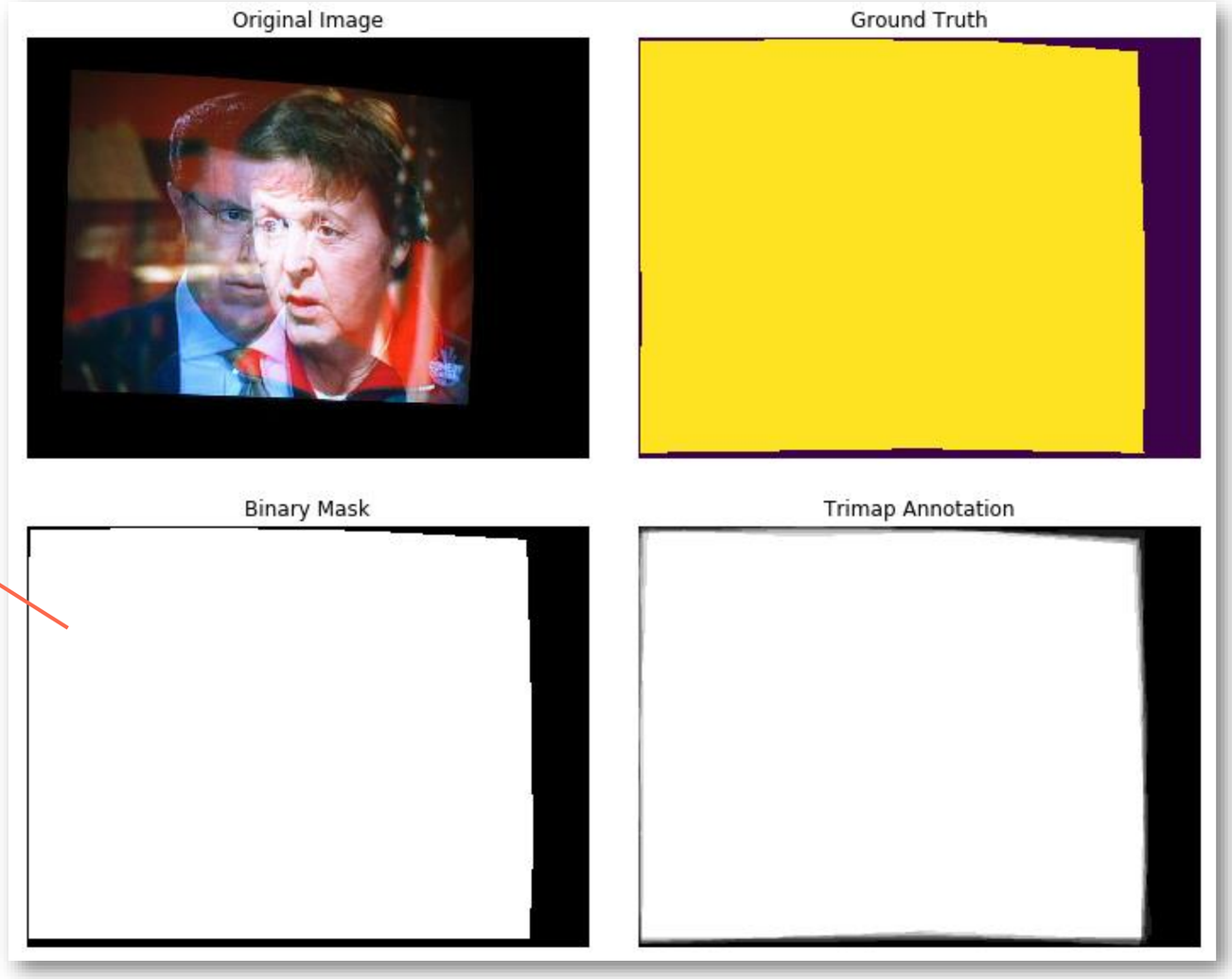
- $V_t = flatten\{AlphaMap(t)\}$
- $V_{t,i} = flatten\{AlphaMask(t,i)\}$

높은 TWS 예시

TWS: 0.996

다른 작업결과와

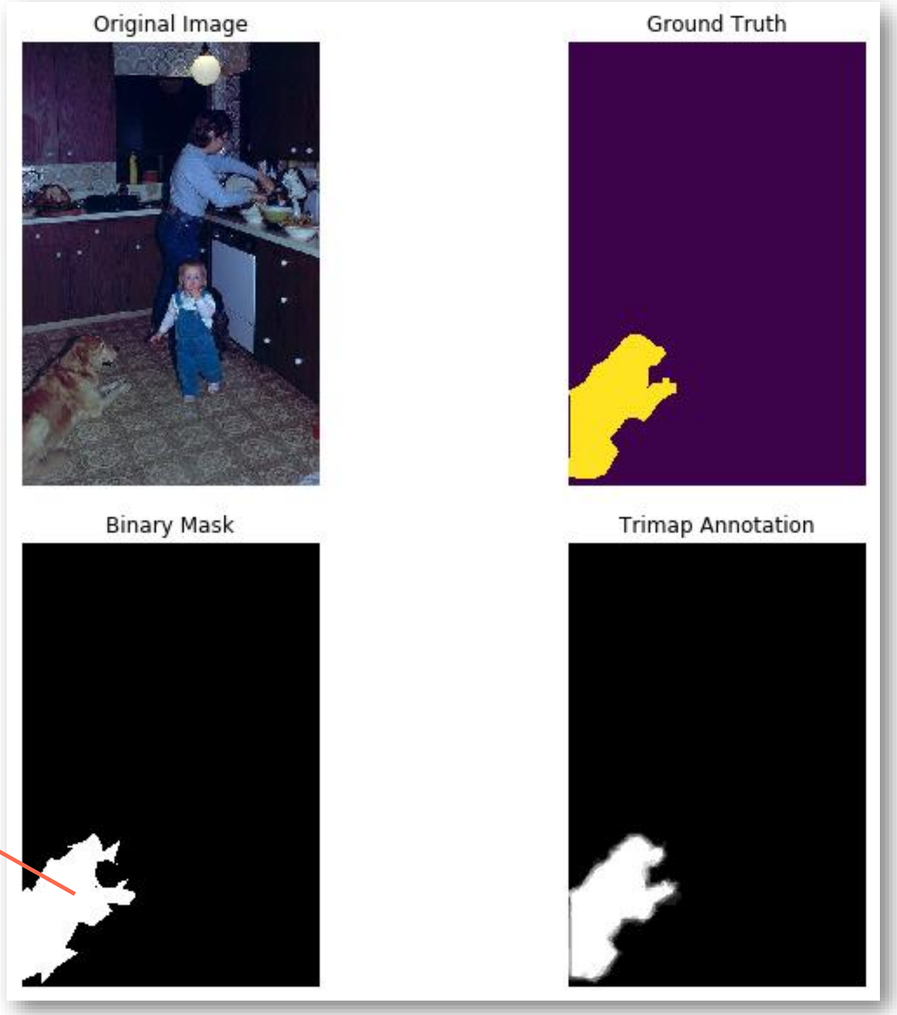
차이가 적을수록 상승



Ref.: MS COCO, https://cocodataset.org/

낮은 TWS 예시

TWS: 0.936



Ref.: MS COCO, https://cocodataset.org/

작업유형에 따른 지표 차이(TWS)

	Large_Jitter	Normal_Dilate	Normal_Erode	Normal_Jitter1	Normal_Jitter2	Normal_Jitter3	Normal_Jitter4	Superior
0	0.964	0.972	0.974	0.982	0.983	0.981	0.981	0.987
1	0.978	0.987	0.986	0.988	0.991	0.986	0.991	0.992
2	0.936	0.960	0.944	0.972	0.965	0.966	0.958	0.975
3	0.943	0.974	0.960	0.971	0.972	0.967	0.969	0.979
4	0.901	0.954	0.930	0.955	0.959	0.953	0.960	0.971
5	0.963	0.983	0.983	0.986	0.985	0.985	0.985	0.989
6	0.953	0.979	0.973	0.978	0.976	0.978	0.977	0.985
7	0.940	0.976	0.965	0.972	0.969	0.977	0.966	0.979
8	0.955	0.977	0.971	0.962	0.977	0.977	0.979	0.985
9	0.987	0.989	0.988	0.987	0.989	0.990	0.994	0.994
10	0.988	0.992	0.993	0.994	0.991	0.994	0.993	0.996
11	0.897	0.940	0.919	0.924	0.936	0.923	0.931	0.958

실제로 사용해보니…

정확도에 따른 차이가 적어 직관적인 판단이 어렵다.

정말 잘된 것	심각한 문제 발생
0.996	0.936

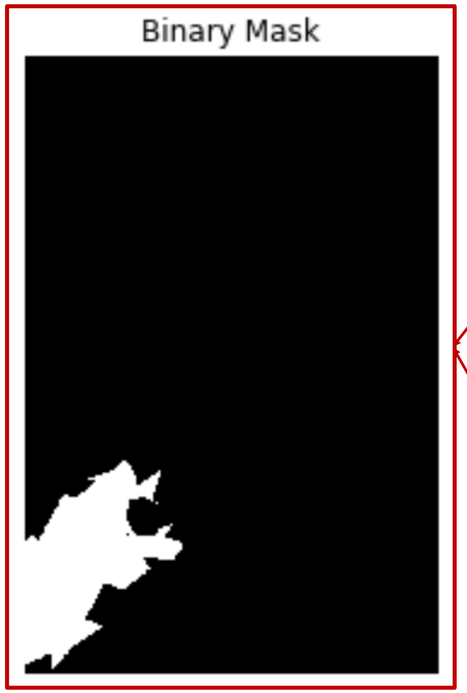
3.4 지표 정의

AMS(Alpha Map Score)

- 의미: 선택된 작업결과와 AlphaMap의 일치도
- 목적: TWS보다 더 직관적인 수치 제시

$$TMS(t,i) = 1 - \frac{\Sigma\{|TM(t) - BM(t,i)|\}}{n(TM(t) > 0)}$$

- $t = task$
- $i = worker$



AMS

	Large_Jitter	Normal_Dilate	Normal_Erode	Normal_Jitter1	Normal_Jitter2	Normal_Jitter3	Normal_Jitter4	Superior
0	0.919	0.926	0.934	0.945	0.945	0.942	0.942	0.950
1	0.950	0.961	0.962	0.965	0.968	0.961	0.968	0.971
2	0.864	0.889	0.879	0.911	0.902	0.904	0.894	0.914
3	0.883	0.919	0.905	0.917	0.918	0.913	0.915	0.928
4	0.816	0.871	0.857	0.882	0.887	0.877	0.887	0.900

TWS

	Large_Jitter	Normal_Dilate	Normal_Erode	Normal_Jitter1	Normal_Jitter2	Normal_Jitter3	Normal_Jitter4	Superior
0	0.964	0.972	0.974	0.982	0.983	0.981	0.981	0.987
1	0.978	0.987	0.986	0.988	0.991	0.986	0.991	0.992
2	0.936	0.960	0.944	0.972	0.965	0.966	0.958	0.975
3	0.943	0.974	0.960	0.971	0.972	0.967	0.969	0.979
4	0.901	0.954	0.930	0.955	0.959	0.953	0.960	0.971

3.4 지표 정의

ICS(Image Clarity Score) 어렵고 모호할수록 작업자 간 이견이 커질 것이다.

- 의미: 작업결과들의 차이 정도
- 목적: 작업의 모호함이나 어려운 정도를 판별

AlphaMap 표시영역 내 값의 합

$$ICS(t) = \frac{\Sigma AlphaMap(t)}{\Sigma\{AlphaMap(t) > 0\}}$$

AlphaMap 표시영역의 면적

- $t = task$

0.765

0.810

0.971

0.958

Ref.: MS COCO, https://cocodataset.org/

식별이 어렵거나,

형태가 복잡할 수록 차이 발생

추가로 고려해야할 내용

작업

품질 저해요인

- 모호한 가이드
- 모호한 선택지,
- 인지하기 어려운 이미지 등

작업결과 일관성을
기준으로 판단

작업자

품질 저해요인

- 의도적: Spammer, Bot 등
- 비의도적: 잘못된 가이드 이해

작업결과

종속적 특징

- 다수(Majority)의 선택이 정답

가장 일반적인 작업을
정답으로 간주

추가로 고려해야할 내용

작업

품질 저해요인

- 모호한 가이드
- 모호한 선택지,
- 인지하기 어려운 이미지 등

작업자

품질 저해요인

- 의도적: Spammer, Bot 등
- 비의도적: 잘못된 가이드 이해

작업결과

종속적 특징

- 다수(Majority)의 선택이 정답

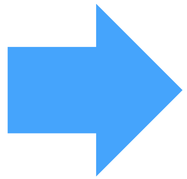
문제성 작업자에 Penalty,

고품질 작업자에 Advantage

3.5 정확도 향상 시도 1: 영향력 차등화

영향력 차등화가 필요한 이유

		객관식 선택지
전문가	Worker1	1번
매크로	Worker2	3번
스팸	Worker3	3번

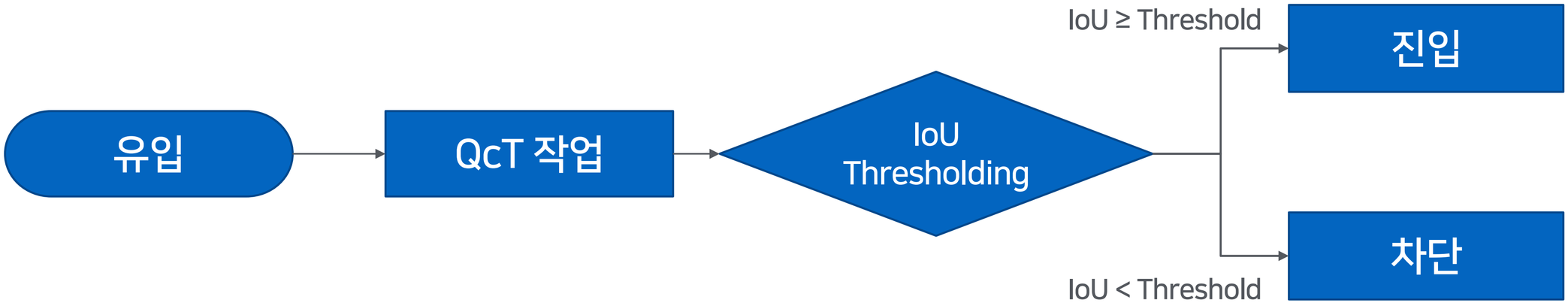


3번이 정답?

3.5 정확도 향상 시도 1: 영향력 차등화

Spammer 등 사전차단을 위한 작업(QcT) 준비

- QcT(Quality Control Task): 성과 측정을 위해 정답을 정해놓은 작업



3.5 정확도 향상 시도 1: 영향력 차등화

고려해야할 요소

- QcT의 난이도: 어려울수록 가점 부여
- 작업결과의 최신성: 최신 작업결과 위주로 고려

변수정의

- Impact Point: 영향력 점수
- Voting Power: 영향력 점수를 정규화한 값
- QcT Range: 영향력 산정에 반영할 가장 최근 QcT의 갯수
- Decay Weight: 오래된 작업결과에 대한 패널티 가중치
- QScore: IoU에 난이도(ICS)에 따른 가점이 부여된 QcT 정확도(점수)

변수상세

$$\boldsymbol{IP(Impact\ Point)} = \frac{\boldsymbol{DecayWeight}\ \cdot \boldsymbol{QScore}}{\boldsymbol{QcT\ Range}}$$

$$\boldsymbol{VP(Voting\ Power)} = \frac{\boldsymbol{IP_{task,i}}}{\sum \boldsymbol{IP_{task}}}$$

- $task = task\ number$
- $i = worker\ number$

$$\boldsymbol{DecayWeight(weight, t)} = \boldsymbol{weight^{i}}$$

- $t = QcT\ Range$
- $i = 1,2, \ldots, t-1$

$$\boldsymbol{QScore\left(QcT_{w,i}, t\right)} = \frac{\boldsymbol{IoU(QcT_{gt,i}, QcT_{w,i})}}{\boldsymbol{ICS(QcT_{i})}}$$

- $w = worker$
- $t = QcT\ Range$
- $i = 1,2, \ldots, t-1$
- $QcT_{w,i} = i_{th}\ recent\ result\ of\ worker_{w}`s\ QcT$
- $QcT_{gt,i} = i_{th}\ recent\ result\ of\ worker_{w}`s\ QcT\ Ground\ Truth$

예시 1: IP 산출

	IoU	ICS
QcT_1	0.9	0.8
QcT_2	0.8	0.9
QcT_3	0.8	0.7
QcT_4	0.9	0.9
QcT_5	0.7	0.8

$t = 5$

$weight = 0.9$

$QScore(QcT_{w,i}, 5)$

$= [1.12, 0.88, 1.14, 1.00, 0.87]$

$DecayWeight(0.9, 5)$

$= [1.0, 0.9, 0.81, 0.73, 0.65]$

$DecayWeight$

1.0	0.9	0.81	0.73	0.65

$\cdot$

$QScore$

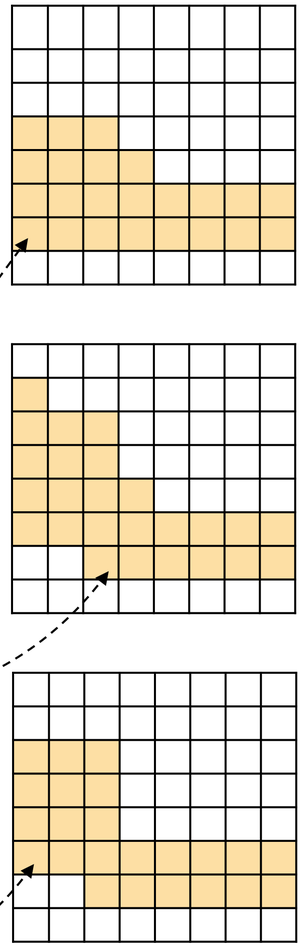
1.12
0.88
1.14
1.00
0.87

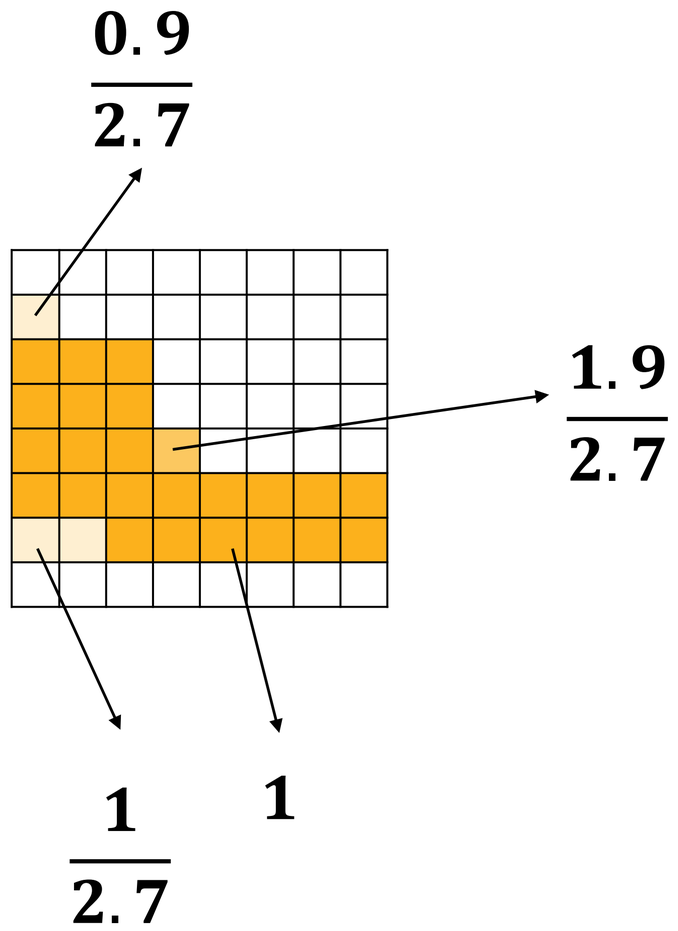
$= 4.13$

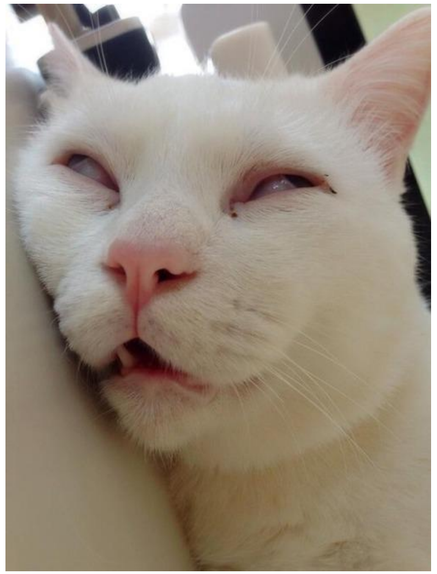
$$IP = \frac{4.13}{5} = 0.82$$

예시 2: VP 산출 및 활용

Worker	w1	w2	w3
IP	1	0.9	0.8
VP	$\frac{1}{2.7}$	$\frac{0.9}{2.7}$	$\frac{0.8}{2.7}$







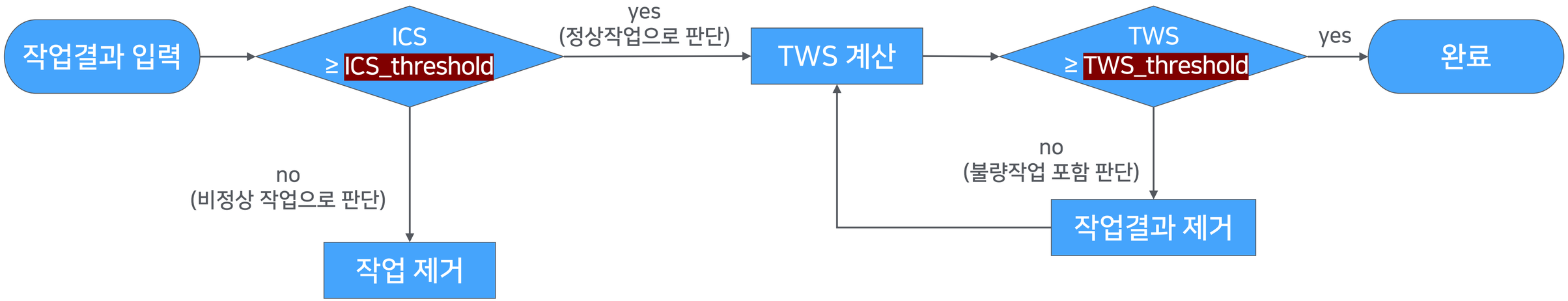
그럼에도 불구하고…

불량작업은 계속 발생

3.6 정확도 향상 시도 2: 불량작업 제거 로직

DEVIEW 2020

EM 알고리즘과 유사하게 진행(Iterative method)



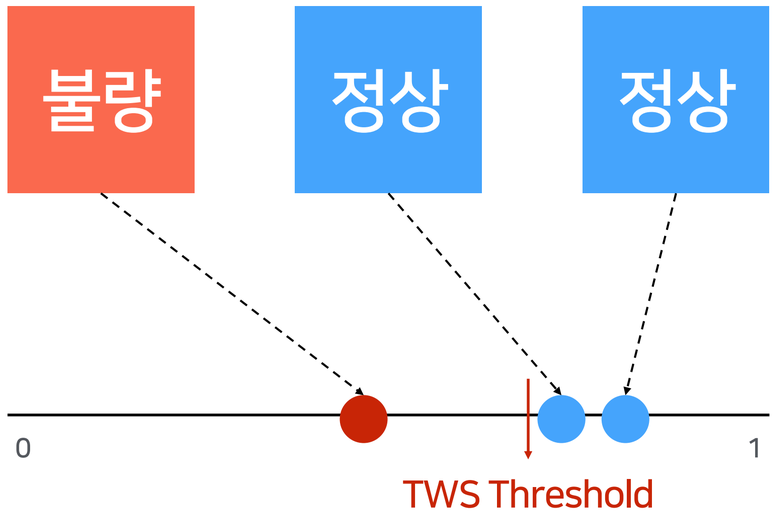
3.6 정확도 향상 시도 2: 불량작업 제거 로직

DEVIEW 2020

Expectation

불량작업 포함 시,

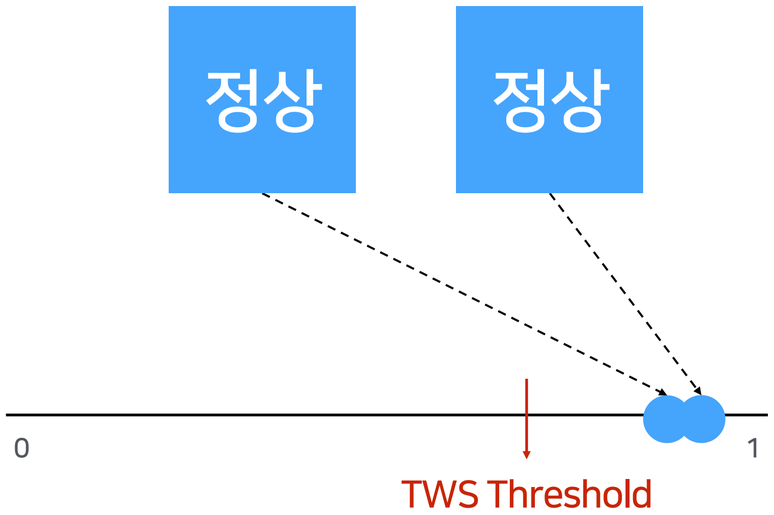
정상작업 점수하락 발생



Maximization

불량작업 제거로

불이익 없이 정상적인 점수 기록



3.6 정확도 향상 시도 2: 불량작업 제거 로직

사용 예시

특이값 제거를 위한 허용범위 설정

$$\mu(TWS) \pm threshold$$

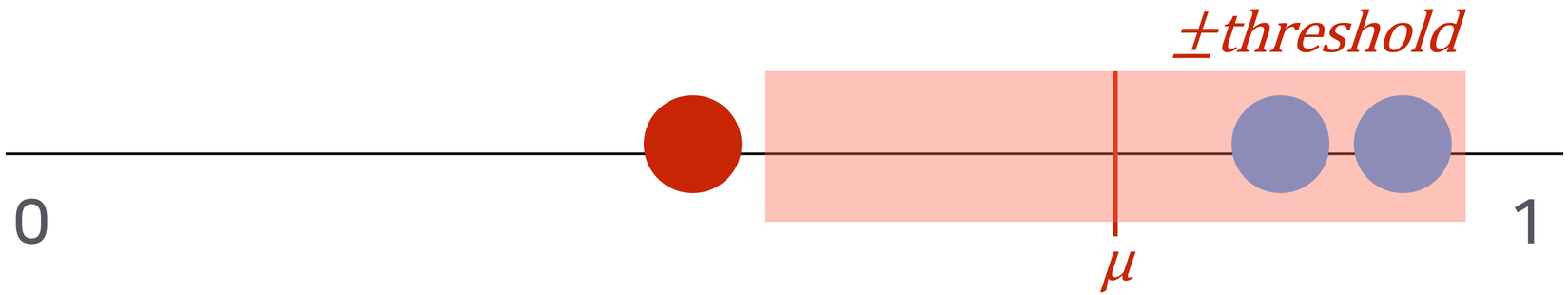
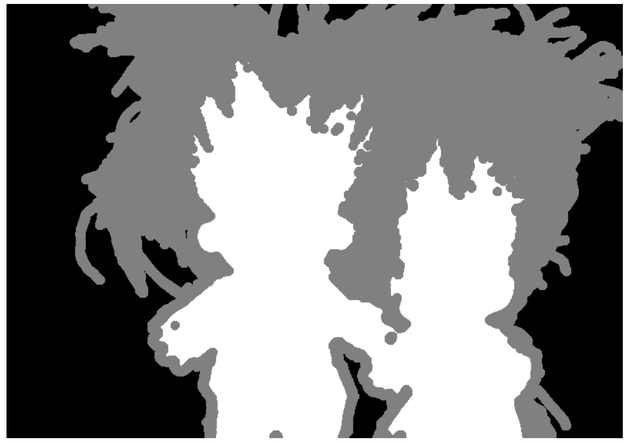
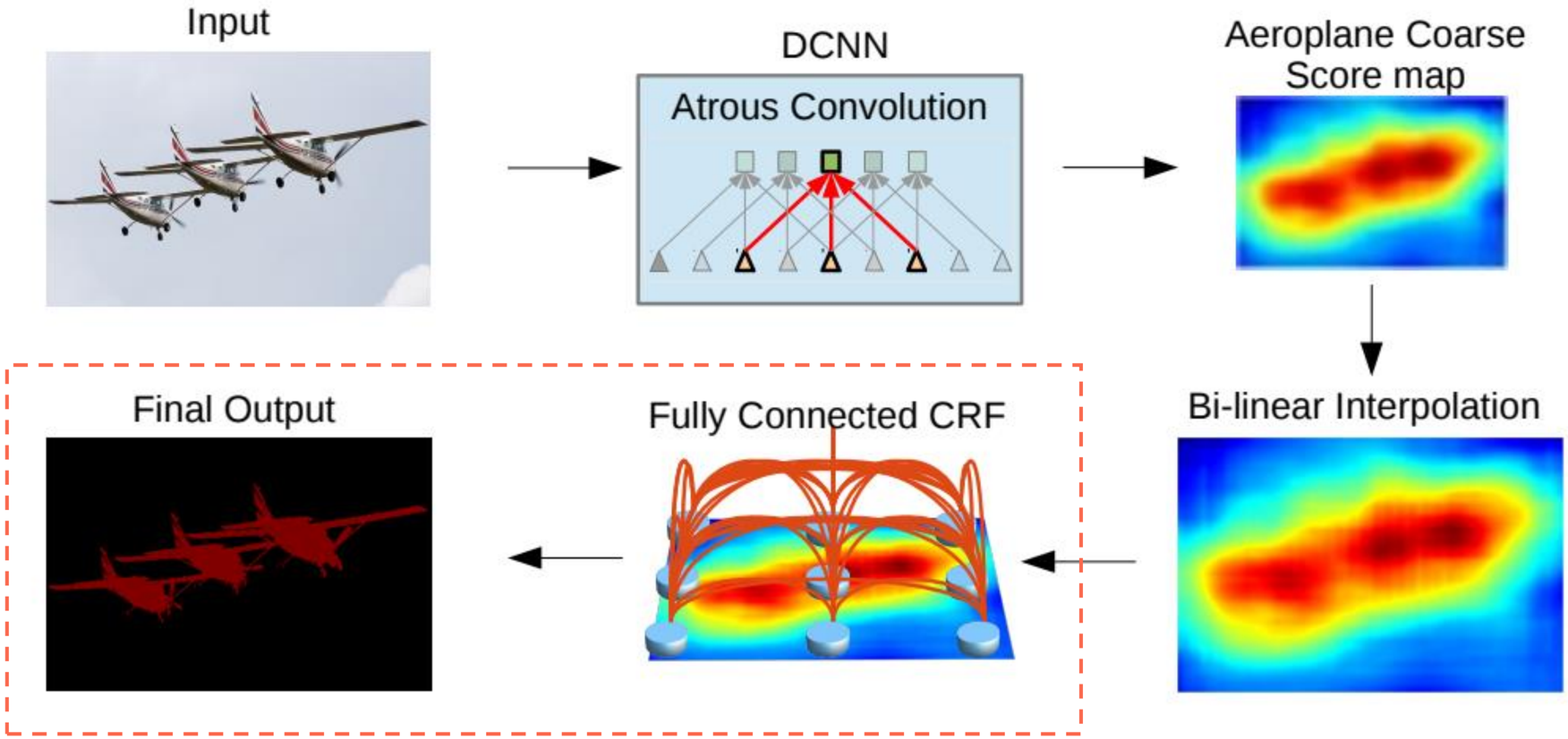


Image Matting을 시도해보면 어떨까?



Ref: http://alphamatting.com/

DeepLab V2



Ref: Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017): 834-848.

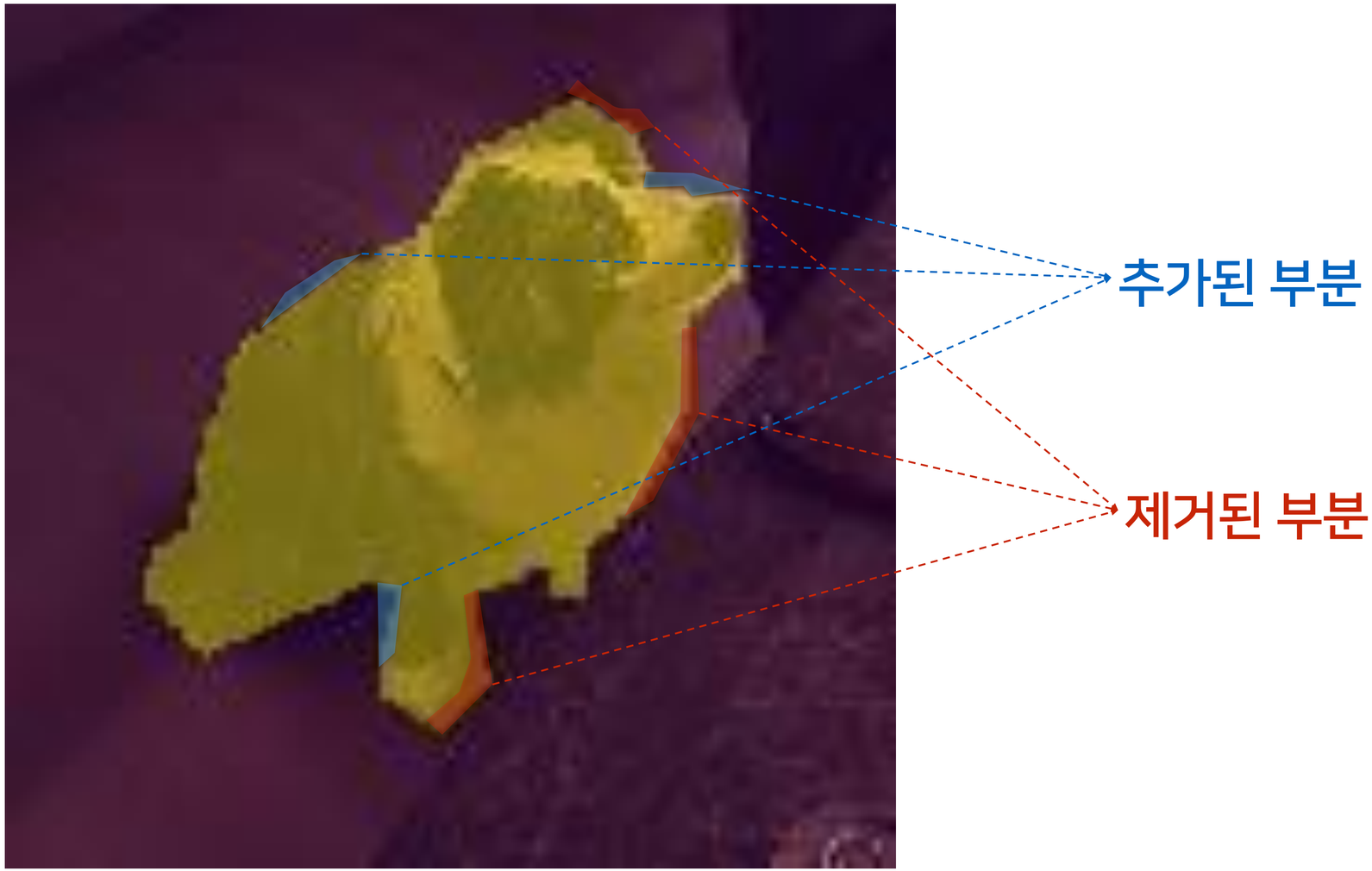
3.7 정확도 향상 시도 3: Image Matting

DEVIEW 2020

Dense-CRF를 활용한 Ground Truth 추론

Ref.: MS COCO, https://cocodataset.org/

전반적으로 납득할만한 결과 도출

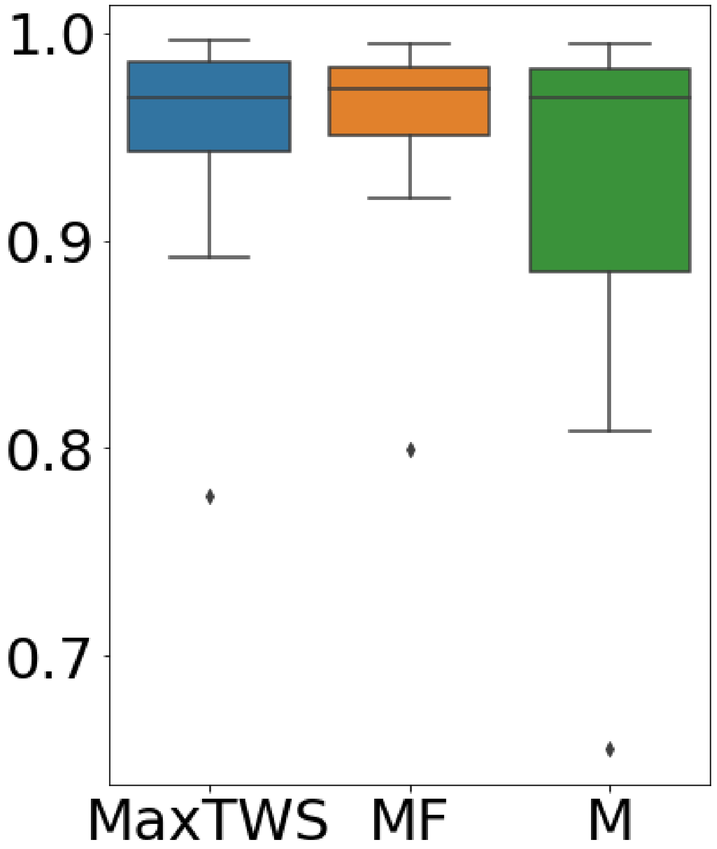


Ref.: MS COCO, https://cocodataset.org/

MaxTWS: TWS 최대값을 정답으로 선택

MF: 작업결과 필터링 후 Matting 적용

M: 작업결과 필터링 없이 Matting 적용



전체적인 정확도 및 일관성 향상

저품질 작업 위주로 정확도 향상

	MaxTWS	MF	M
mean	0.958658	0.962483	0.931437
std	0.043394	0.036833	0.074676
min	0.776778	0.799829	0.655274
25%	0.942956	0.950411	0.884710
50%	0.968784	0.972690	0.968601
75%	0.985726	0.982987	0.982654
max	0.996611	0.994935	0.994781

낮은 정확도, 낮은 일관성

당연히 불량 작업 필요

3.7 정확도 향상 시도 3: Image Matting

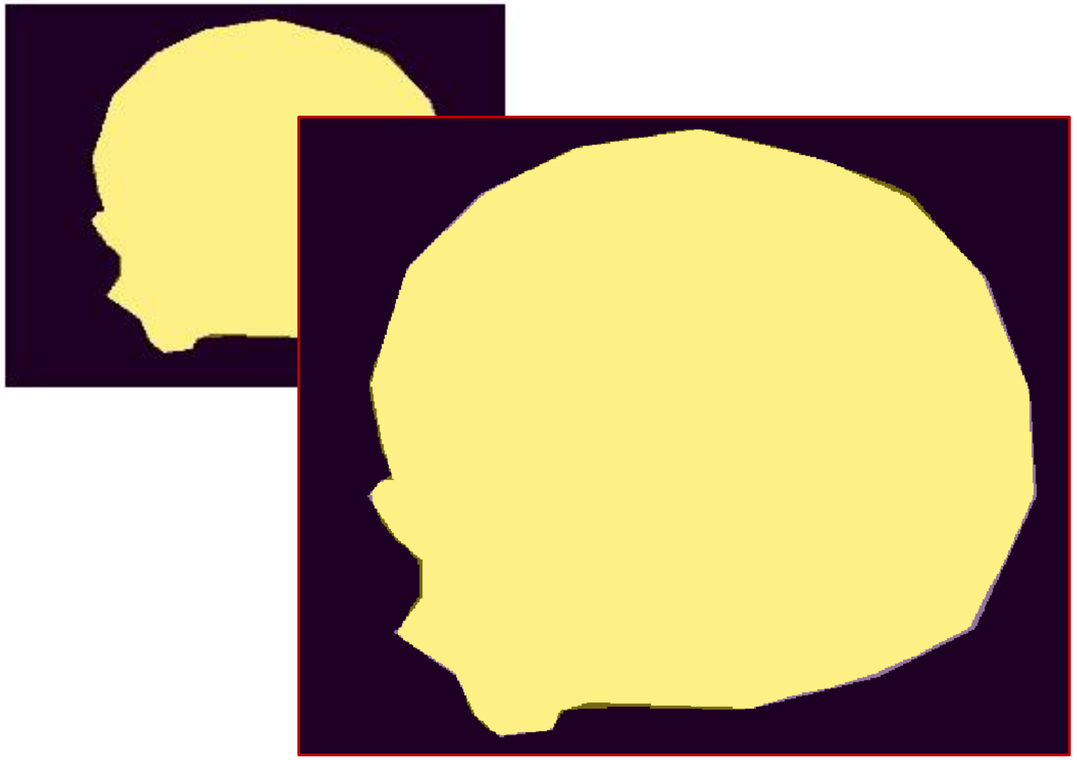
DEVIEW 2020

별다른 효과가 없는 경우 1: 다 같이 잘한 경우

Ground Truth

Trimap Annotation

TA Dense-CRF



Ref.: MS COCO, https://cocodataset.org/

3.7 정확도 향상 시도 3: Image Matting

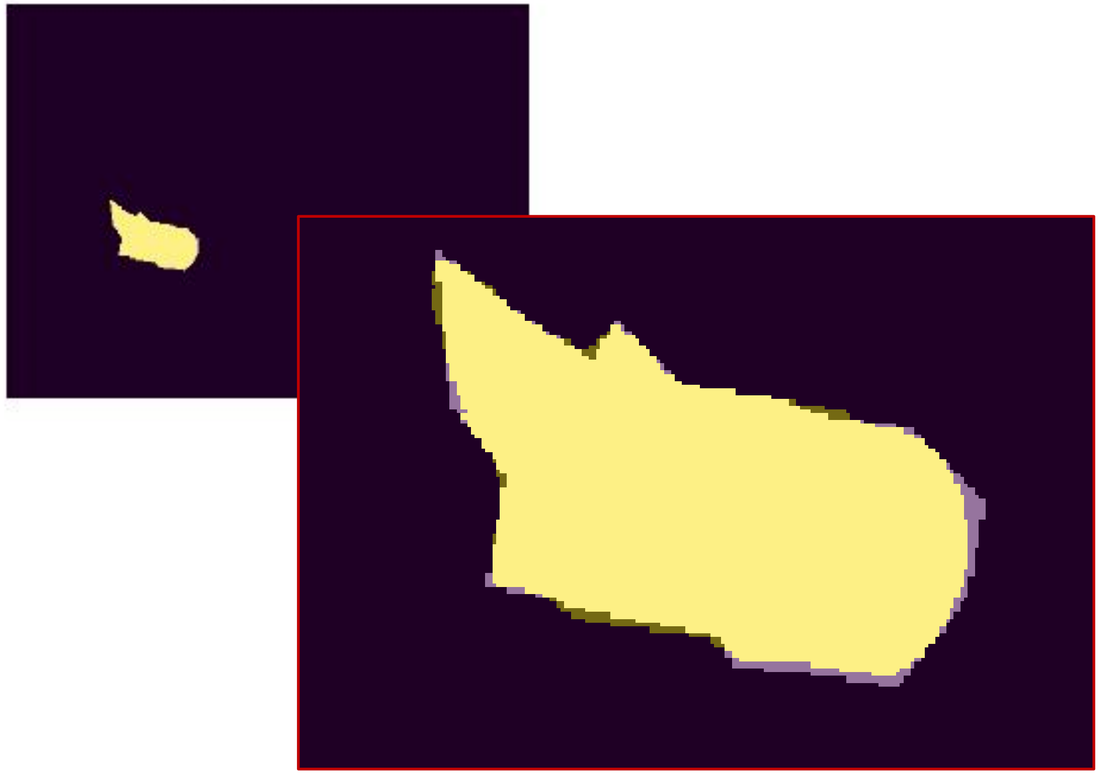
DEVIEW 2020

별다른 효과가 없는 경우 2: 다 같이 못한 경우

Ground Truth

Trimap Annotation

TA Dense-CRF



Ref.: MS COCO, https://cocodataset.org/

3.7 정확도 향상 시도 3: Image Matting

DEVIEW 2020

적용 후 결론

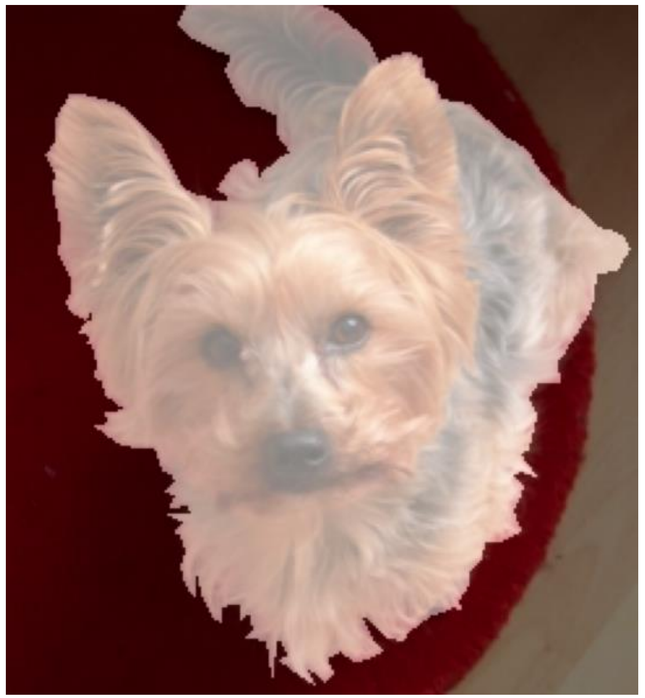
- 매번 Hyperparameter 적정값을 탐색해야하는 번거로움 발생
- 알고리즘 적용결과가 어떤 영향을 미쳤는지 추적이 어려움.
- 전수 검수를 진행한다면 검수자의 수정소요 경감 가능

4.적용사례

4.1 Instance Segmentation

- 1 이미지 / 1 인스턴스
- 작업 1개 당 7 ~ 14명 할당

Ground Truth example

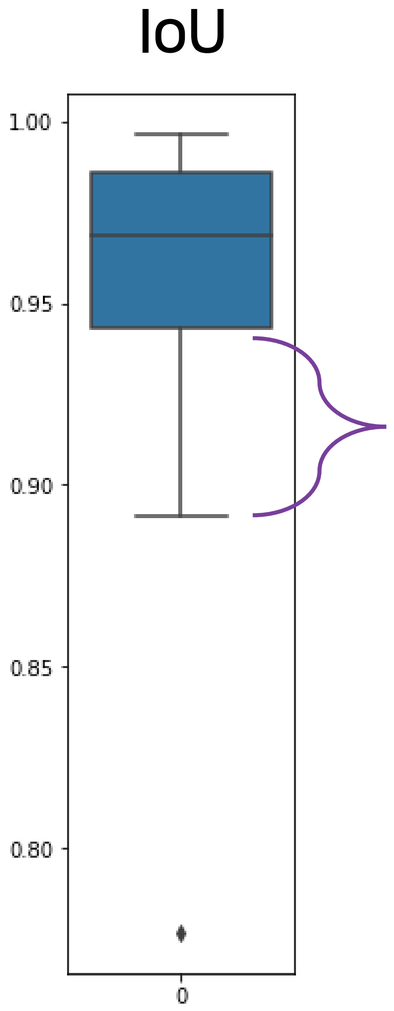


Ref.: flickr, https://www.flickr.com/

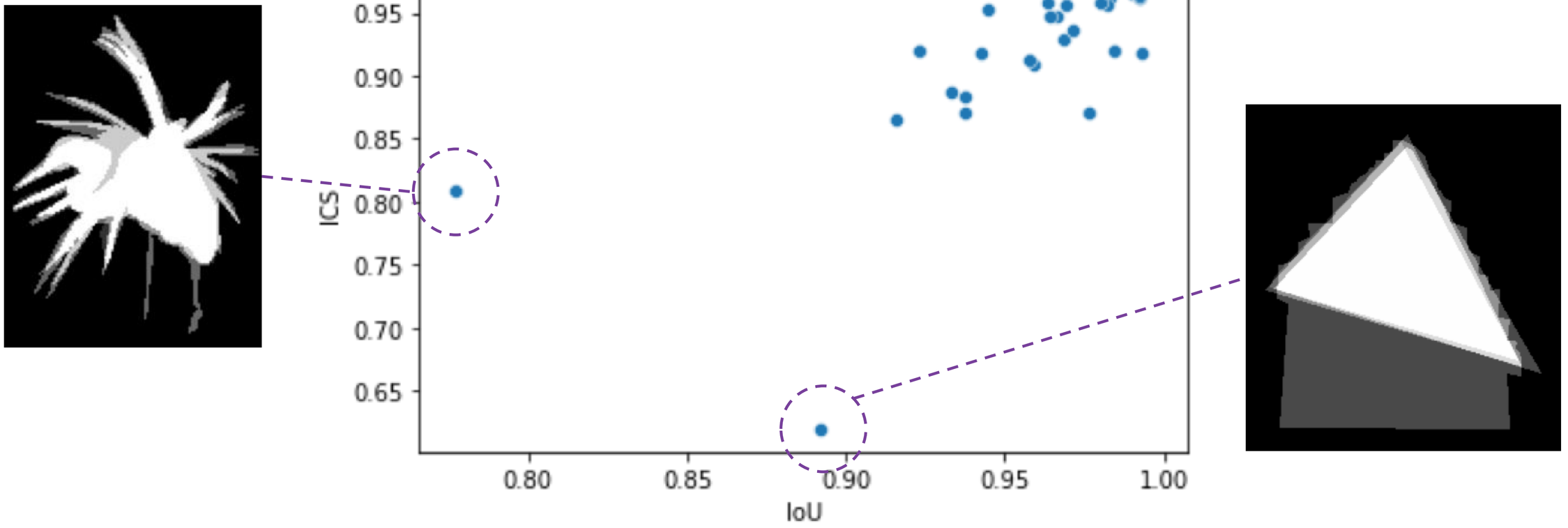
4.1 Instance Segmentation

DEVIEW 2020

mean	0.958658
std	0.043394
min	0.776778
25%	0.942956
50%	0.968784
75%	0.985726
max	0.996611



4.1 Instance Segmentation



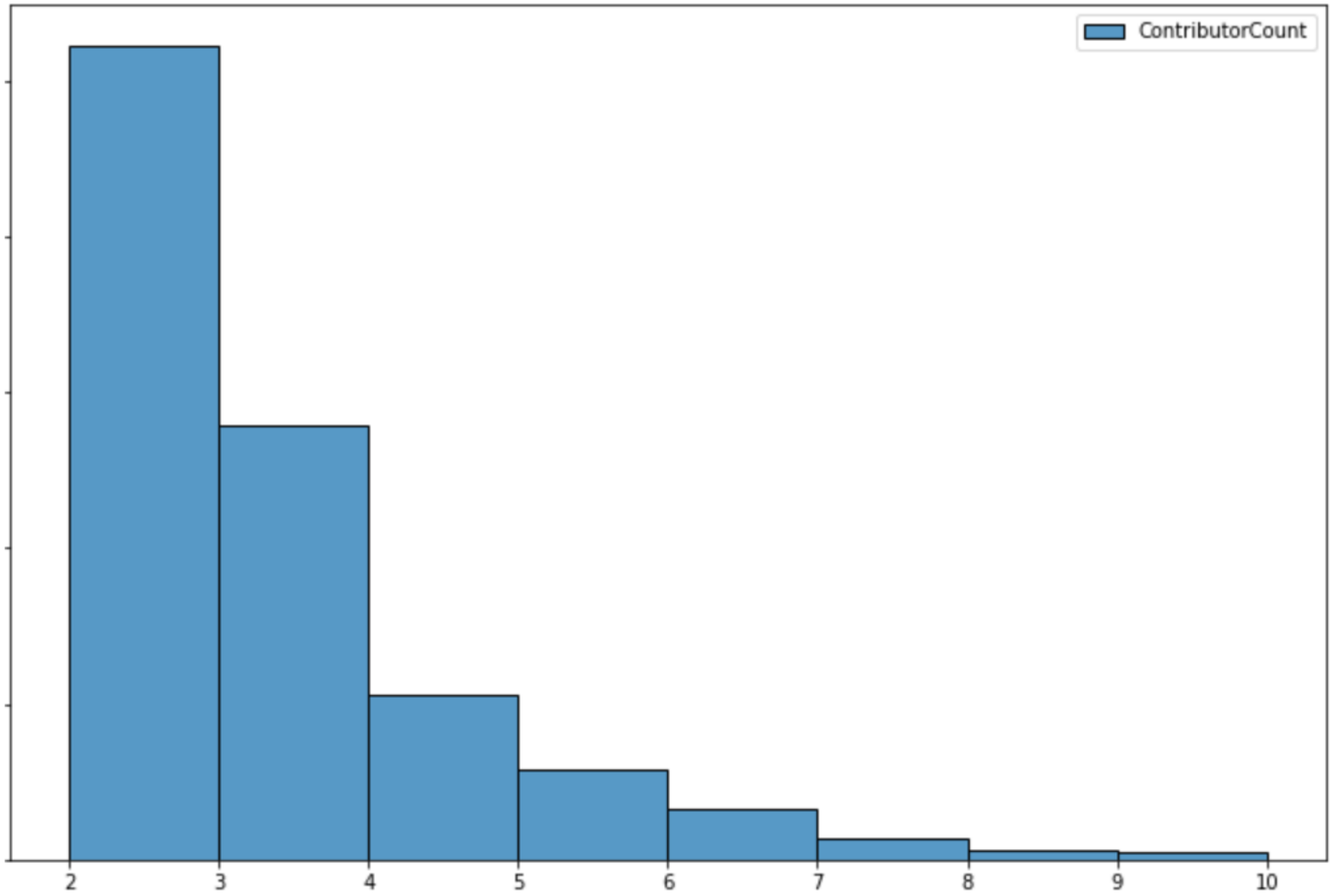
4.2 Image Classification

- 이진 분류 문제
- 2명에게 할당 후 과반일치 발생까지 할당 반복(Pass 선택가능)

작업예시) 다음 중 “재규어”를 선택해주세요.



4.2 Image Classification



- Precision = 0.95
- Recall = 0.89
- F1-score = 0.92
- Accuracy = 0.97

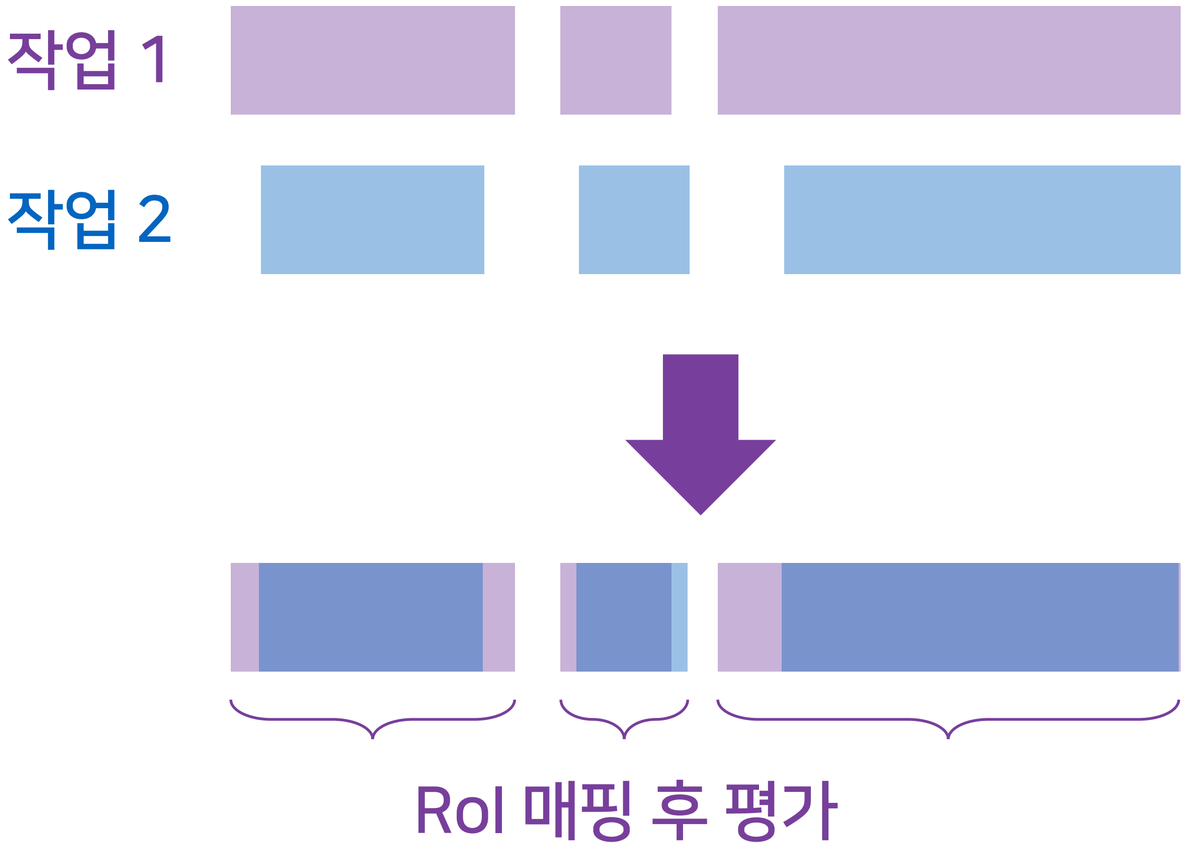
4.3 Scene Text Localization

- Rotated Bounding Box
- 가로쓰기, 세로쓰기 포함

가로쓰기
안녕하세요

세로쓰기
반갑습니다

4.3 Scene Text Localization



ICS 일정 수치 미만

RoI 선별검수

4.3 Scene Text Localization

- Precision = 1.0
- Recall = 0.97
- F1-score = 0.98

4.4% 검수 대부분 판단이 쉽지않은 모호한 작업

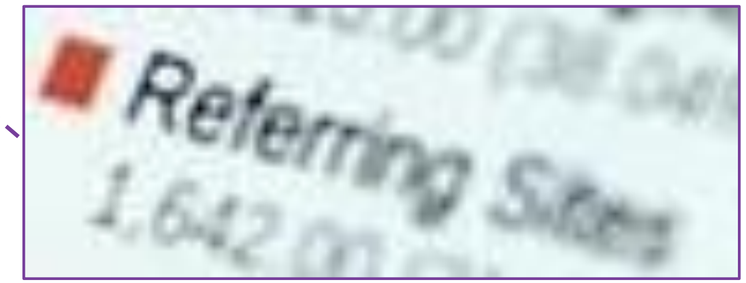
4.3 Scene Text Localization

DEVIEW 2020

모호한 작업의 예시

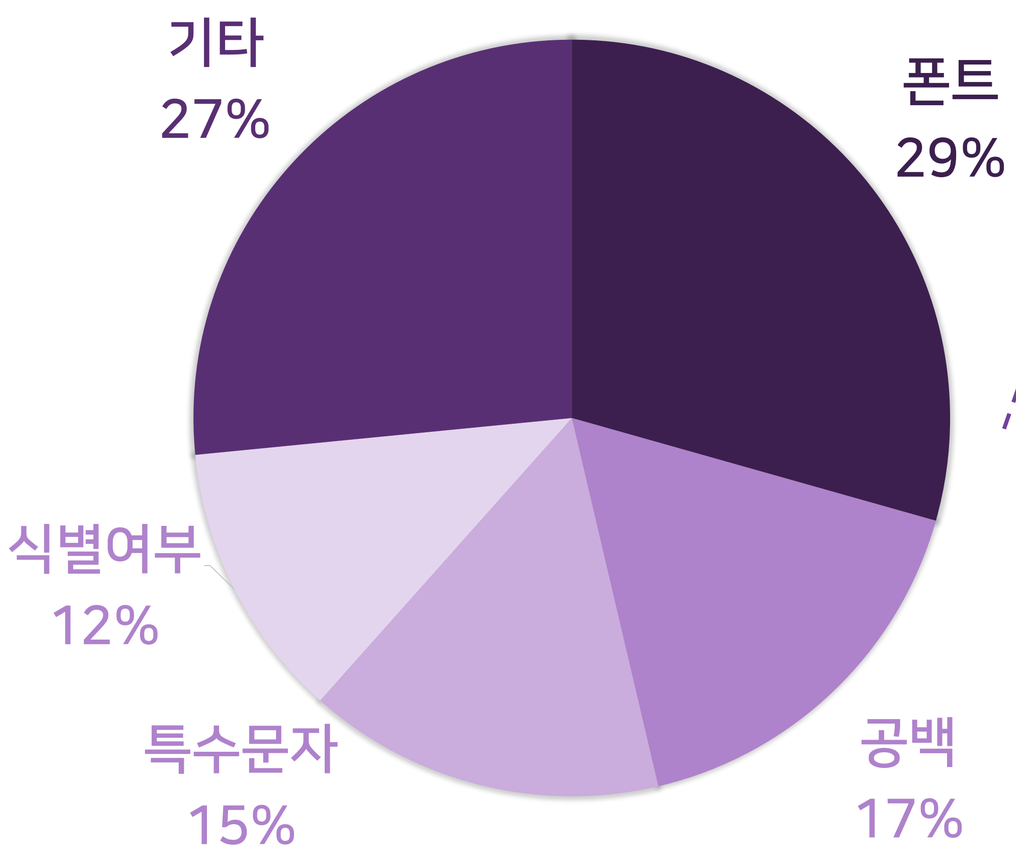
"Map"인 것 같긴 한데…

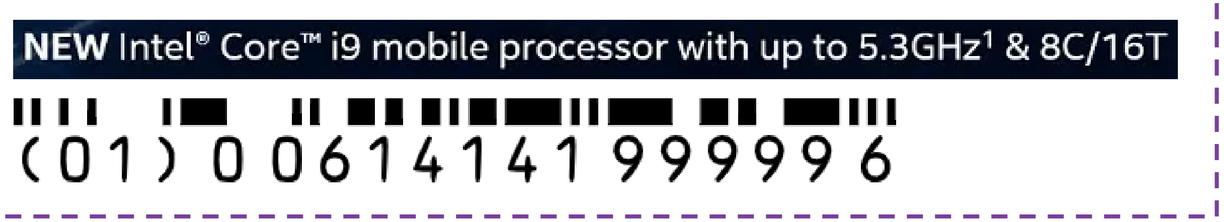
식별이 가능하다고 해야할까?



어디까지 박스를 지정하지?

4.3 Scene Text Localization





기술적인 접근보다는…

명료한 작업기준

작업기준의 일관된 이해(교육)

5.정리

운영부담을 덜 수 있다.

대응해야 할 상황이 너무 많다.

5. 정리

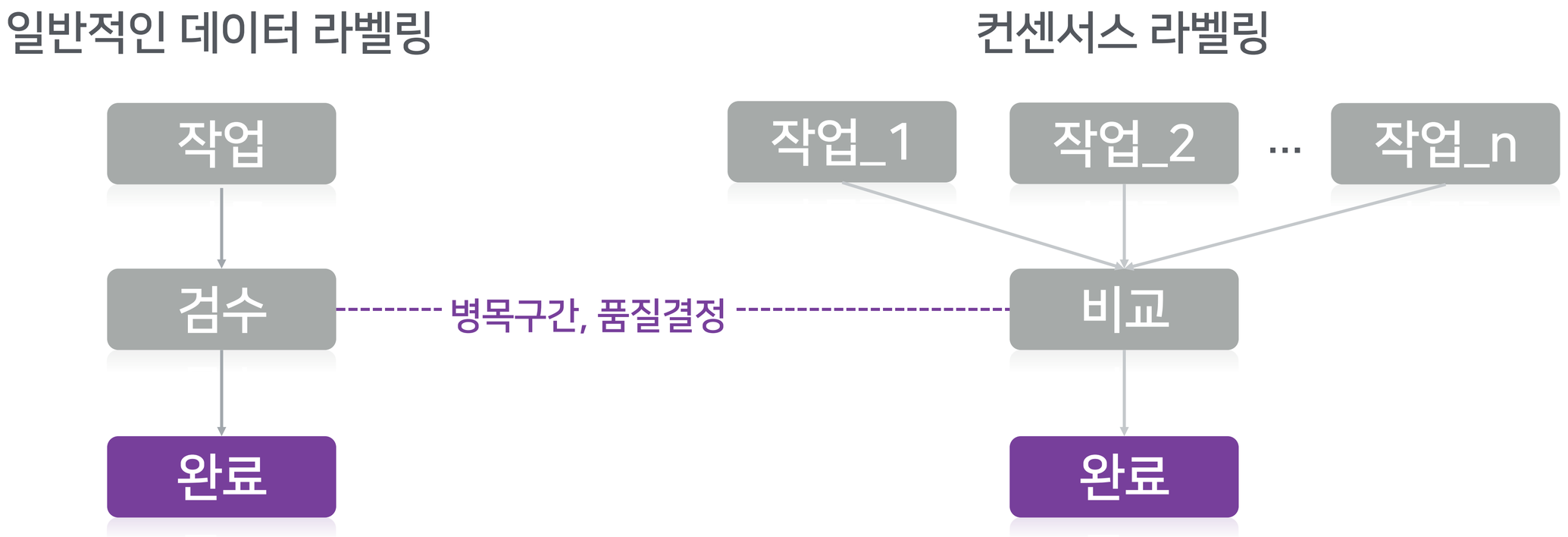
좋았던 점

- 예상보다 적은 비용: 튜토리얼 도입 후 1 작업 당 3명 할당까지 가능
- 운영소요 경감: 검수작업 90% 이상 감소

힘들었던 점

- 각 라벨링 프로젝트별로 대응해야 할 다양한 도메인 특성 존재
- 평가 로직 개발에 생각보다 많은 시간 소요

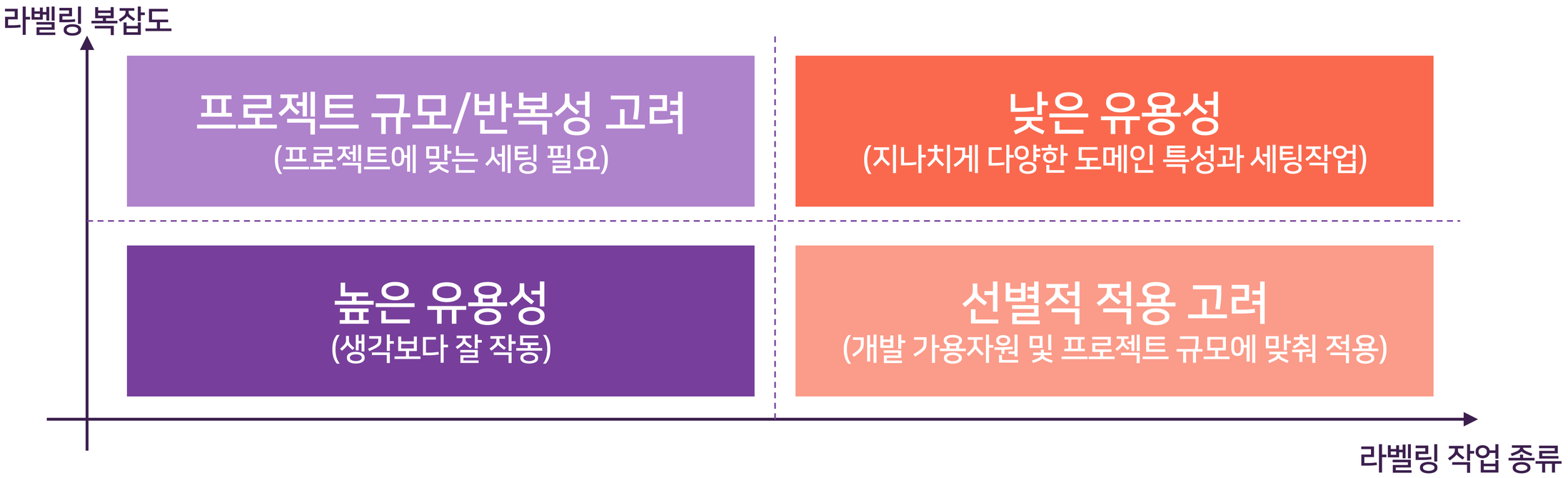
5. 정리



5. 정리

그래서 어떻게하면 돼요?

- 라벨링 작업의 종류가 얼마나 많은지
- 라벨링 작업이 얼마나 복잡한지



Q&A

Thank You