Seeing through the Human Reporting Bias

Visual Classifiers from Noisy Human-Centric Labels

(CVPR'16)

Wonwoong Cho

What do you see?

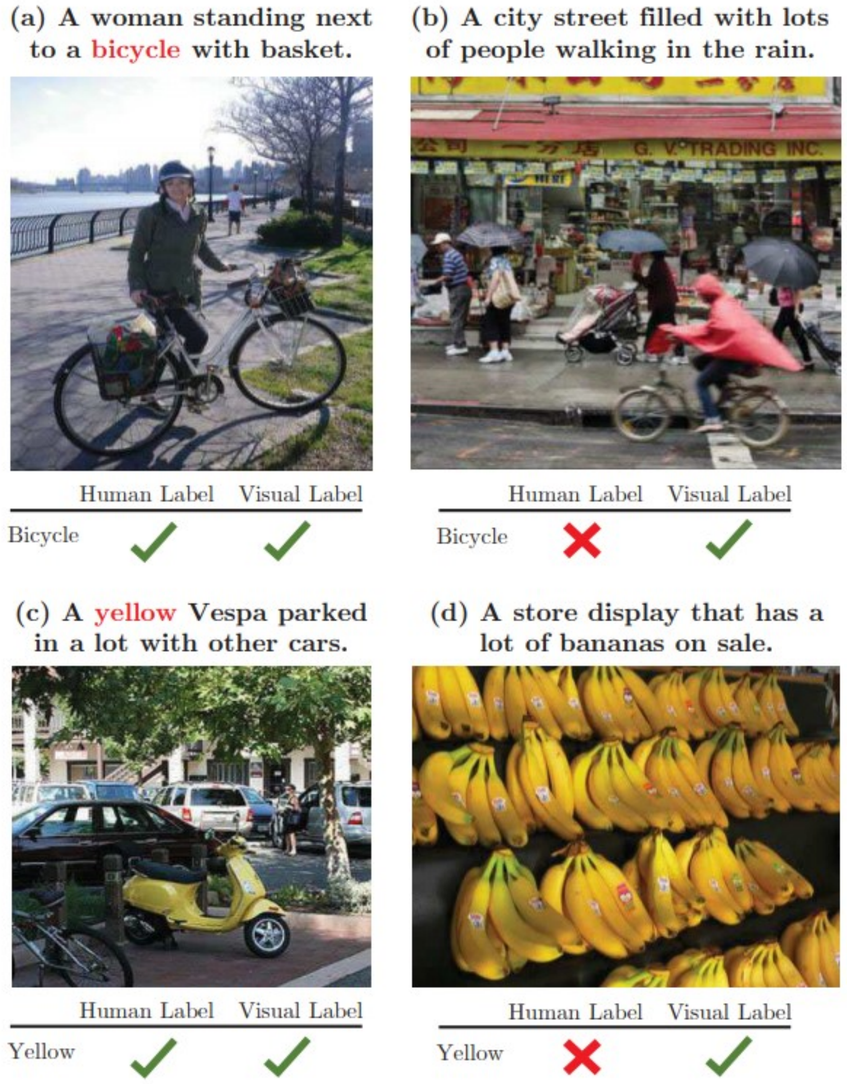
- Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

- Yellow Bananas

Motivation

- When human annotators are given a choice about what to label in an image, they apply their own subjective judgments on what to ignore and what to mention.
- We refer to these noisy “human-centric” annotations as exhibiting human reporting bias.
- Such annotations do not use consistent vocabulary, and miss a significant amount of the information present in an image
- we demonstrate that the noise in these annotations exhibits structure and can be modeled.

Motivation



Human descriptions capture only some of the visual concepts present in an image. For instance, the bicycle in (a) is described, while the bicycle in (b) is not mentioned.

How to tackle?

- This paper proposes to train a model that explicitly factors human-centric label prediction into a visual presence classifier and a relevance classifier.

- Visual presence classifier: 이미지에 해당 visual concept이 있냐?
- Relevance classifier: 사람이 {Banana, Yellow} 가 주어졌을 때 뭘 선택하는지를 학습

What can be expected from this approach?

- visual representation을 더 잘 뽑게 되서 다양한 테스크에서 성능 증진

Basic setting

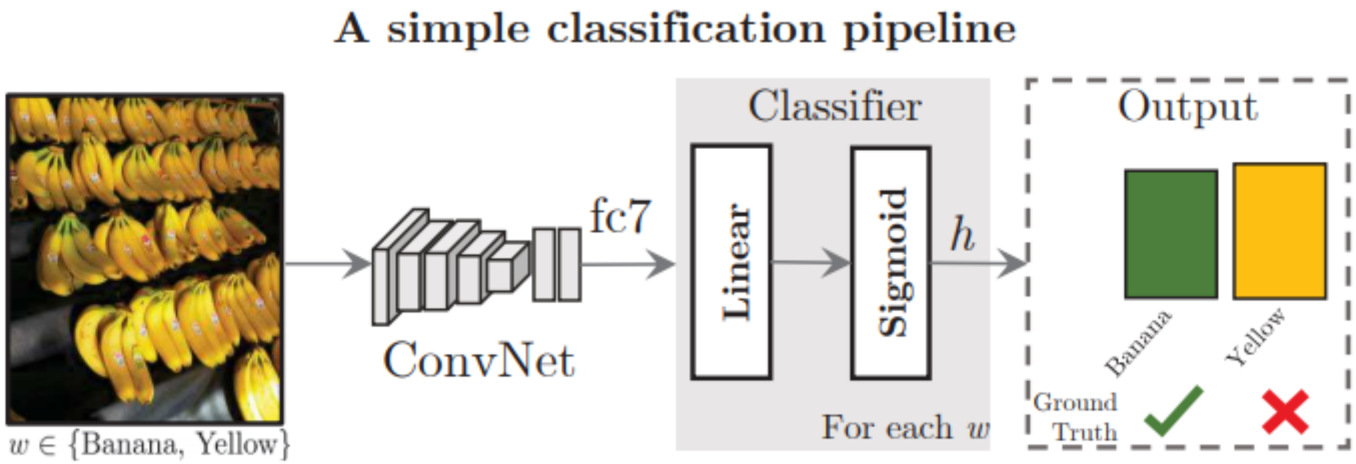


Figure 2: A simple classification model for learning from human-centric annotations. The noisy labels (`banana` is not annotated as `yellow`) impede the learning process.

Method

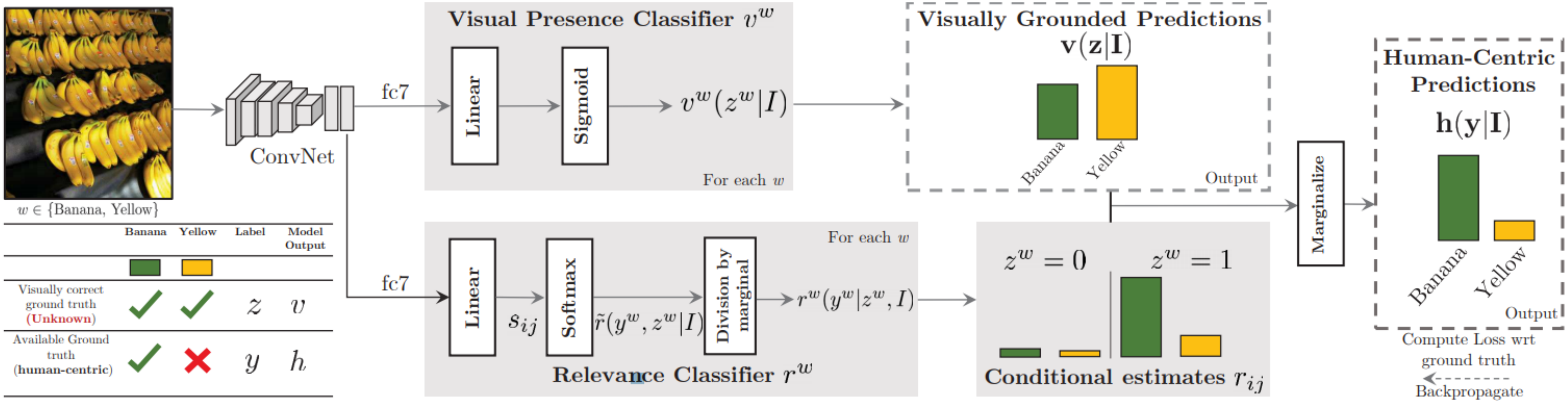
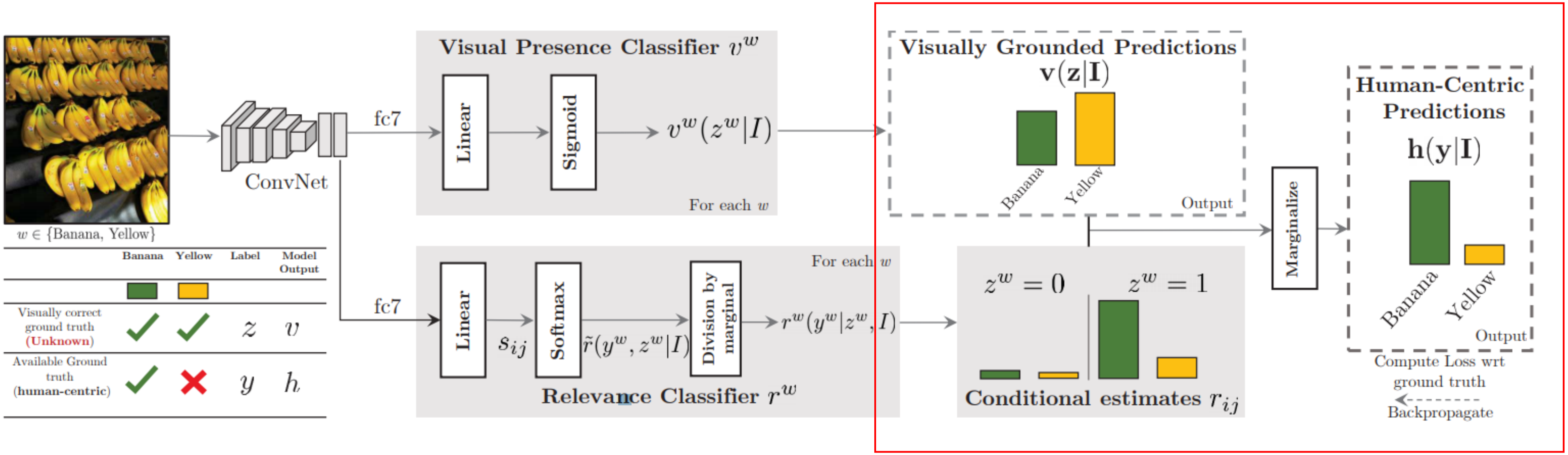


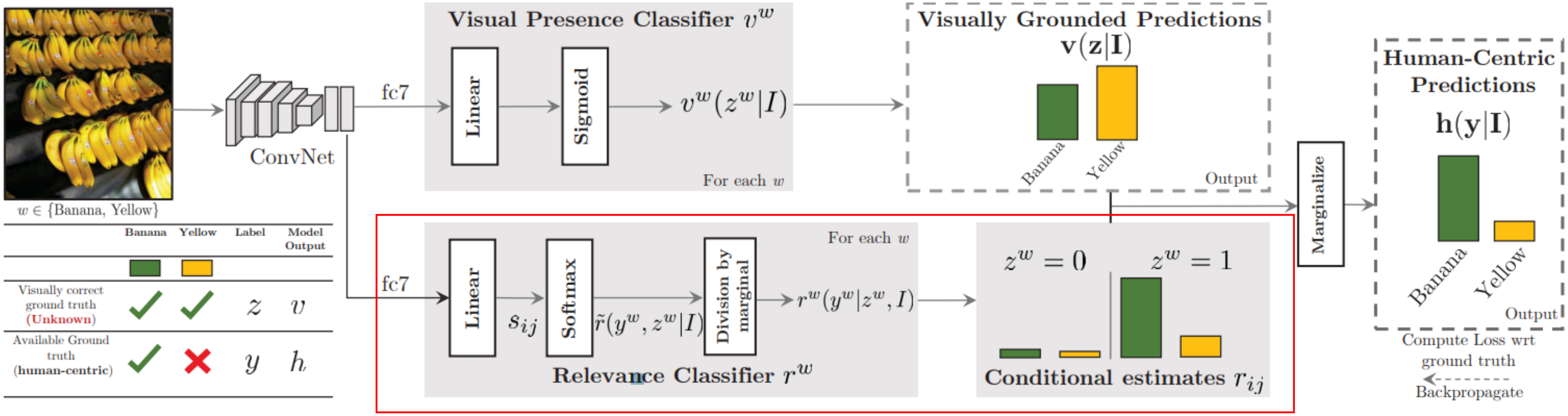
Figure 3: Our model uses noisy human-centric annotations y for learning visually grounded classifiers without access to the visually correct ground truth z. It uses two classifiers: a visual presence classifier v and a relevance classifier r. The visual presence classifier v predicts whether the visual concept w is visually present in an image. The relevance classifier r models the noise and predicts whether the concept should be mentioned or not. We combine these predictions to get the human-centric prediction h.

Method



$$h^w(y^w|\mathcal{I}) = \sum_{j \in \{0,1\}} r^w(y^w|z^w = j, \mathcal{I})v^w(z^w = j|\mathcal{I})$$

Method



$$s_{ij} = m_{ij}^T \phi(\mathcal{I}) + b_{ij},$$

$$\tilde{r}_{ij} = \exp(s_{ij}) / \sum_{i'j'} \exp(s_{i'j'}). \qquad r_{ij} = \tilde{r}_{ij} / \sum_{i'} \tilde{r}_{i'j}.$$

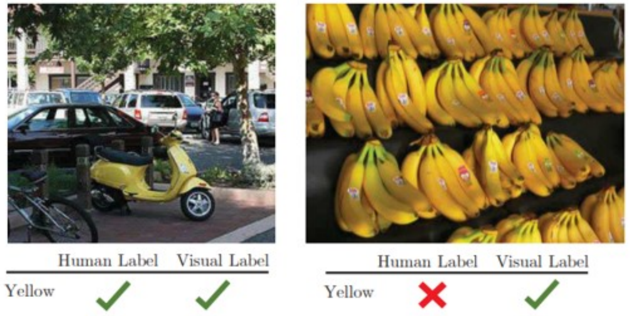
Experiments

- MS COCO dataset => visual concept 1000개 추출
- TrainSet:

  각 이미지 당 caption 5개를 훑어서 visual concept가 있으면 1, 없으면 0으로 총 1000dim label
- TestSet:

  unmentioned concept을 데이터에서 뽑아야 함

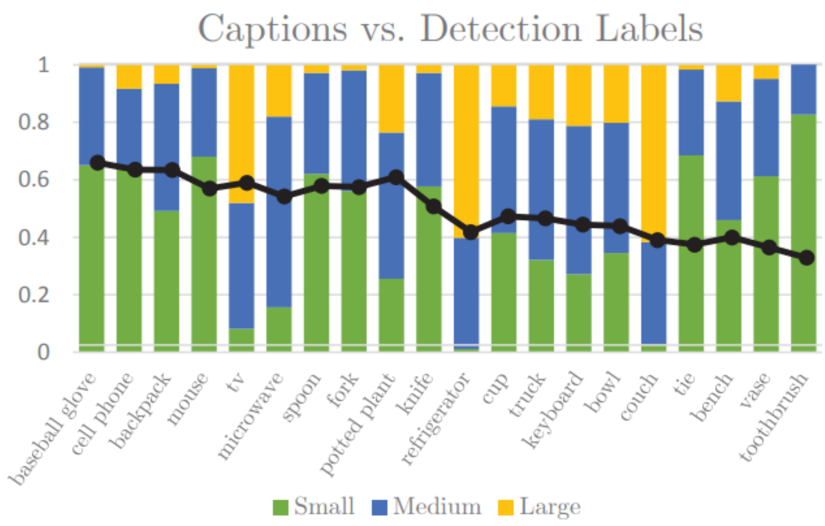
(c) A yellow Vespa parked in a lot with other cars. (d) A store display that has a lot of bananas on sale.



Experiments

- TestSet:

1. Caption 및 detection 데이터에서 동시에 많이 존재하는 73개 클래스 선별
2. Caption ground truth 에는 없고 (y=0),
3. Detection ground truth 에는 있음 (z=1)



Experiments

			Mean Average Precision								Precision at Human Recall							
			NN	VB	JJ	DT	PRP	IN	Others	All	NN	VB	JJ	DT	PRP	IN	Others	All
		Prob	616	176	119	10	11	38	30	1000	← Count							
VGG16	MILVC [12]	-	41.6	20.7	23.9	33.4	20.4	22.5	16.3	34.0	52.7	32.8	40.5	40.3	32.2	33.0	24.6	45.8
VGG16	MILVC + Multiple-`fc8`	-	41.1	20.9	23.7	33.6	21.1	22.8	16.8	33.8	51.2	32.6	40.8	41.1	31.7	33.5	27.3	45.0
VGG16	MILVC + Latent (Ours)	v	42.9	21.7	24.9	33.1	19.6	23.0	16.2	35.1	53.6	35.4	43.3	41.3	28.0	36.0	24.4	47.2
VGG16	MILVC + Latent (Ours)	h	44.3	22.3	25.8	34.4	21.8	23.6	17.3	**36.3**	55.5	36.3	44.7	42.9	32.1	37.3	26.4	**48.9**
AlexNet	MILVC [12]	-	33.2	16.2	20.1	30.9	16.4	19.9	14.6	27.4	40.0	26.4	36.0	38.2	24.2	27.5	21.9	35.9
AlexNet	MILVC + Latent (Ours)	v	35.6	17.7	21.9	32.4	16.9	20.7	15.2	29.4	43.9	28.3	37.5	41.2	29.2	29.9	23.3	39.0
AlexNet	MILVC + Latent (Ours)	h	36.5	18.0	22.4	32.9	17.8	21.4	15.6	**30.1**	45.1	28.7	38.0	41.2	32.2	31.0	24.0	**40.0**
VGG16	Classif.	-	34.9	18.1	20.5	32.8	19.2	21.8	16.3	29.0	42.5	30.4	33.9	40.5	30.4	30.7	23.8	38.2
VGG16	Classif. + Multiple-`fc8`	-	34.2	17.7	19.9	32.6	19.0	21.5	15.9	28.4	41.3	27.9	32.3	39.6	29.6	31.2	22.6	36.8
VGG16	Classif. + Latent (Ours)	v	37.7	19.6	22.0	32.6	20.2	22.0	16.3	31.2	46.3	32.9	36.8	38.9	32.3	33.1	27.0	41.5
VGG16	Classif. + Latent (Ours)	h	38.7	20.1	22.6	33.8	21.2	23.0	17.5	**32.0**	47.8	33.7	37.9	42.5	34.2	34.4	29.0	**42.9**

Experiments

Experiments