ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO — Supplementary Materials —

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang$^\diamondsuit$, Seong Joon Oh$^\clubsuit$

NAVER AI Lab
$\diamondsuit$ Now at Google Research $\clubsuit$ Now at University of Tübingen

We include additional materials in this document. We first describe the details of our machine annotators (Appendix A), including the explanation of each model (Appendix A.1) and diversity between each model (Appendix A.2). We provide the details of Human Intelligence Tasks (HITs) for ECCV Caption construction (Appendix B), such as detailed questionnaire (Appendix B.1), MTurk worker statistics (Appendix B.2) and the results (Appendix B.3). Appendix C describes the post-processing details, including the full list of invalid items (Appendix C.1) and the examples of ECCV Caption (Appendix C.2). We include more evaluation results in Appendix D, such as user study details for comparing mAP@R and Recall@k (Appendix D.1), training details of re-implemented methods (Appendix D.2), the full results with various evaluation metrics (Appendix D.3). Finally, we provide the full bias analysis in Appendix E and the discussions of noisy crowdsource annotations in Appendix F.

A ECCV Caption Machine Annotators Details

A.1 Machine annotators

To cover both diversity and practical relevance, we have choose five state-of-the-art cross-modal retrieval models with diverse properties.

- VSRN [24] builds up connections between image regions, and perform reasoning with Graph Convolutional Networks to generates features with semantic relationships. VSRN uses the Faster R-CNN detector [33] as the visual encoder following [1]. VSRN employs the triplet loss [36] with hardest negative mining (HNM) [13].
- PVSE [38] learns a one-to-many function to solve ambiguous matching by one-to-one function. PVSE is a multi-headed model, focusing on diverse matching between two diverse concepts. PVSE also employs the triplet loss with HNM as VSRN.
- PCME [7] is a stochastic model for learning many-to-many correspondences in multi-modal matching tasks. PCME is trained by a probabilistic matching objective function based on the pair-wise matching loss.

- ViLT [20] is a vision-language pre-training method with massive paired data (4.1M images and 9.9M captions). While other methods have separated text and visual backbones, ViLT has a unified shared Transformer [39] backbone for text and visual modalities.
- CLIP [32] is a contrastive approach for massive but noisy associations and shows powerful zero-shot classification performances. CLIP is trained with 4M image and caption pairs. We use the ViT-B/32 CLIP, the largest one when we start the annotation process.

PVSE, VSRN and PCME use pre-trained visual backbones (ImageNet-trained ResNet, Visual Genome [21]-trained Faster R-CNN) and only use COCO Caption dataset as the training dataset. We use the official weights provided by the authors, except for PCME. We re-train PCME with CutMix [40] pre-trained ResNet-152. This slightly boosts the original performances. We illustrate the example retrieved images by each model in Figure A.1.

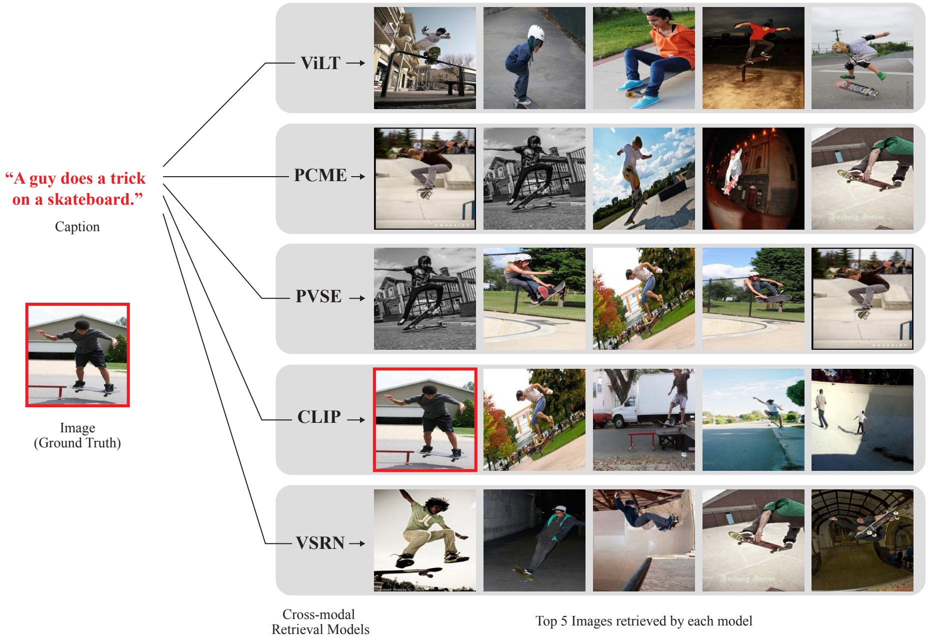


Fig. A.1: **Example retrieved images by the machine annotators.** For the given caption ("A guy does a trick on a skateboard."), we show the top-5 images retrieved by models. The matched pair in the dataset is denoted by red boxes.

A.2 Diversity between machine annotators

We illustrate the quantify of the diversity between machine annotators by their retrieved items in Table A.1. We retrieve 25 images by each model from the

	PVSE	VSRN	PCME	ViLT	CLIP
PVSE	-	0.266	0.430	0.089	0.088
VSRN	0.266	-	0.260	0.060	0.049
PCME	0.430	0.260	-	0.110	0.120
ViLT	0.089	0.060	0.110	-	-0.013
CLIP	0.088	0.049	0.120	-0.013	-

(a) **Model similarity analysis by Kendall's τ.** A higher score means that two models are more correlated.

	PVSE	VSRN	PCME	ViLT	CLIP
PVSE	-	4.52	3.87	5.95	6.24
VSRN	4.50	-	4.47	5.45	5.97
PCME	3.87	4.50	-	5.79	6.20
ViLT	5.79	5.19	5.78	-	6.18
CLIP	6.16	5.81	6.12	6.03	-

(b) **Model similarity analysis by the average ranking.** A smaller rank means that two models are more similar.

Table A.1: **Model similarity analyses.** We measure similarties between the machine annotators in two different ways. (a) We measure the ranking correlations using Kendall's τ; 1.0 means two lists are identical, -1.0 indicates two lists are strongly disagreed each other. (b) We measure the average rankings of the image retrieved by a model for other models. Each row indicates the average ranking of the top-1 retrieved image of the row model for other column models.

COCO validation captions, and measure (1) Kendall's τ (Table A.1a), and (2) the average ranking of the top-1 retrieved items by a model of another model.

Table A.1a shows the Kendall's rank correlation coefficients (Kendall's τ) between the models. Kendall's τ is computed on two ranked lists $[x_1, x_2, \ldots, x_n]$ and $[y_1, y_2, \ldots, y_n]$. We say that two pairs ,(x_i, x_j) and (y_i, y_j), *agree* if either $(x_i > x_j$ and $y_i > y_j)$ or $(x_i < x_j$ and $y_i < y_j)$. Kendall's τ is computed by $\tau = \frac{\text{\#agreed pairs}-\text{\#non-agreed pairs}}{\text{\#all pairs}}$. We use the tau-B variant for the tie-breaking.

B Human Intelligence Tasks (HITs) for ECCV Caption Construction

B.1 HIT details

The example HIT for crowd workers is shown in Figure B.1. Each of the 20 questions in the HIT ask the workers to select the degree of belief that the given image-description pair is a positive match. We have designed the HITs in such a way that not only the positivity of the match is recorded, but also the degrees and rationales for the workers' judgments are collected. Workers can choose among "100% YES", "Partially YES, but", "Mostly NO, because", and "100% NO". Here, we use four choices instead three level ("YES", "Not Sure", and "NO") to avoid encouraging the workers to select "Not Sure" for all questions. If a worker chooses "Partially YES, but" or "Mostly NO, because", then they are asked further questions on the rationale behind their uncertainty. Four possible shortcomings for the image-description match are presented as choices: "the description describes concepts that *do not appear* in the image", "the description *does not describe* the main concepts in the image", "the description describes the main concepts in *a wrong way*", and "the description is grammatically incorrect". Finally, if a worker thinks the description describes the image in a wrong way, we ask *how* the description is wrong. The possible choices here (*e.g.*, quantity, color, . . .) have been crystallized from an internal, preliminary study.

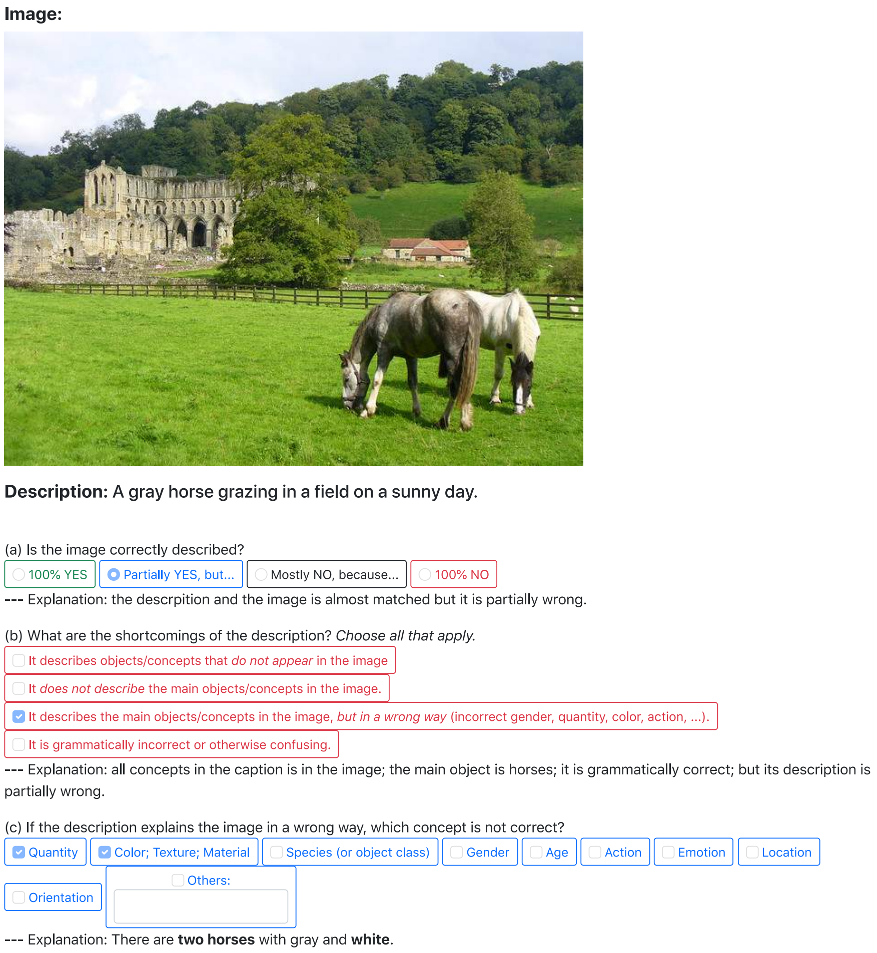


Image:

Description: A gray horse grazing in a field on a sunny day.

(a) Is the image correctly described?

○ 100% YES ● Partially YES, but... ○ Mostly NO, because... ○ 100% NO

--- Explanation: the descrpition and the image is almost matched but it is partially wrong.

(b) What are the shortcomings of the description? *Choose all that apply.*

□ It describes objects/concepts that *do not appear* in the image

□ It *does not describe* the main objects/concepts in the image.

☑ It describes the main objects/concepts in the image, *but in a wrong way* (incorrect gender, quantity, color, action, ...).

□ It is grammatically incorrect or otherwise confusing.

--- Explanation: all concepts in the caption is in the image; the main object is horses; it is grammatically correct; but its description is partially wrong.

(c) If the description explains the image in a wrong way, which concept is not correct?

☑ Quantity ☑ Color; Texture; Material □ Species (or object class) □ Gender □ Age □ Action □ Emotion □ Location □ Orientation □ Others:

--- Explanation: There are **two horses** with gray and **white**.

Fig. B.1: **Example question in a MTurk HIT.** The question asks whether the image is correctly described. If unsure ("Partially YES, but ..." or "Mostly NO, because ..."), the question prompts the worker to provide the rationale. There are 20 of such questions in each HIT.

We make two separated HITs for the annotation process. In the first stage, we verify the results of the image-to-caption retrieval results of five models. We also ask the crowd workers to justify their answer if they choose "Partially YES" or "Mostly NO". In the second stage, we verify the results of the caption-to-

# HITs / worker	# Workers	# Submit. HITs	# Approv. HITs	Approve ratio
1	599	599	459	76.6%
1 < and ≤ 5	254	681	535	78.6%
5 < and ≤ 10	61	475	374	78.7%
10 < and ≤ 15	23	297	216	72.7%
15 < and ≤ 25	21	411	285	69.3%
25 <	12	506	293	57.9%

Table B.1: **MTurk worker statistics.** The number of the unique workers, the submitted HITs, the approved HITs and the average approve ratio by the number of completed HITs are shown.

image retrieval results of the models. After we analyzed the first HITs, we have concluded that the justification stage is not highly useful as our expectation. We omit the justification questions in the second stage for reducing annotation costs. The first annotation round was between 23rd Aug 2021 to 7th Sep 2021. During the first stage, 1,000 HITs are verified by human annotators. The second stage was between 24th Jun 2022 to 10th Feb 2022, and 1,160 HITs are verified during this stage.

B.2 MTurk workers

Before launching the crowdsourcing on AMT, we have conducted an in-lab study involving 70 HITs and 27 workers for 4 days. We have observed that if workers continuously complete HITs, the average elapsed time per HIT is about 4 to 8 minutes. Based on this estimate, we have set the compensation level for each HIT to $1.4 so that a worker can earn $15 per hour in the first stage. For the second stage, we have set the the compensation level for each HIT to $0.65, based on the similar in-lab study without justification questions. The final costs including platform fees for the first and the second stages are $1.65 and $0.78, respectively.

In the main crowdsourcing phases, crowd workers are recruited through AMT. The detailed statistics for workers are shown in Table B.1. Overall, 970 unique workers have completed 2,969 unique Human Intelligence Tasks (HITs), while 807 HITs of them (37.3%) have been rejected by our qualification process. The average elapsed time for each HIT of the first and second annotation phase are 9.5 and 13.8 minutes, respectively. The average HITs per worker is 3.06.

B.3 MTurk results

We summarize the results of the crowdsourced annotations, corresponding to 2,160 approved HITs on MTurk, in Table B.2 and Table B.3. In Table B.2, we show the ratios of “Yes”, “Weak yes”, “Weak no” and “No” for different models and rankings. Here, we observe that weaker results (*i.e.*, worse ranked pairs)

	PVSE				VSRN			
Rank	Yes	Weak yes	Weak no	No	Yes	Weak yes	Weak no	No
1	52.3	31.7	11.2	4.7	54.4	31.6	10.0	4.0
2	39.0	36.7	16.7	7.5	39.9	37.6	15.8	6.7
3	34.5	36.9	18.0	10.6	36.6	36.8	17.6	9.0
4	30.9	39.2	19.6	10.4	33.2	38.1	17.6	11.1
5	28.0	39.2	20.5	12.3	30.4	38.6	19.0	12.0
Avg.	36.8	36.8	17.3	9.1	38.8	36.6	16.0	8.6

	PCME				ViLT			
Rank	Yes	Weak yes	Weak no	No	Yes	Weak yes	Weak no	No
1	52.3	33.2	10.4	4.2	47.6	33.7	12.0	6.7
2	39.8	37.2	15.1	7.9	36.6	33.5	19.7	10.2
3	35.6	39.0	17.5	7.9	30.3	35.6	20.2	13.9
4	31.9	37.8	21.1	9.2	26.7	35.6	22.8	14.9
5	27.8	40.0	21.5	10.6	25.8	35.6	22.5	16.1
Avg.	37.4	37.5	17.2	8.0	32.9	34.8	19.7	12.5

	CLIP			
Rank	Yes	Weak yes	Weak no	No
1	50.5	33.5	12.0	3.9
2	35.3	38.5	18.0	8.2
3	33.8	39.9	16.1	10.1
4	31.0	38.6	18.0	12.5
5	28.4	39.8	20.2	11.6
Avg.	36.0	38.0	16.8	9.2

Table B.2: **Model-wise annotation overview.** The percentages of "100% YES", "Partially YES", "Mostly NO" and "100% NO" for each model and each ranking are shown. For example, the first row indicates the annotation results for the top-1 retrieved image and description pairs.

have lower "Yes" and "Weak yes" ratios. For example, the annotation results for PCME show that the "Yes" ratios monotonically decrease as the rank goes down: 52.3% for the most similar pairs, but 27.8% for the least similar pairs. Interestingly, we observe that the average ratio of "Yes" + "Weak yes" for the top-1 retrieved items exceed 80% for all five models (*e.g.*, 86.0% for VSRN), while the R@1 score of each model is known to less than 60% (See Table 4 in the main paper).

From the table, we observe that by letting annotators verify more similar pairs by machines, the annotation process becomes more efficient, *i.e.*, we can acquire the same amount of positive annotations with less number of human ver-

	Not in image	Not in caption	Incorrect object description	grammar error
PVSE	30.4	17.5	48.0	4.1
VSRN	30.1	18.4	46.5	5.0
PCME	31.6	17.9	46.1	4.4
ViLT	31.2	18.9	45.7	4.2
CLIP	28.8	18.5	47.1	5.6
Avg.	30.4	18.2	46.7	4.7

(a) **Shortcomings by models.**

	Quantity	Color	Species	Gender	Age	Action	Emotion	Location	Orientation	Others
PVSE	25.7	20.1	11.9	5.4	2.2	14.9	0.9	3.4	12.1	3.2
VSRN	23.8	20.1	12.7	4.0	1.7	15.5	1.2	4.3	13.6	3.0
PCME	24.8	20.9	12.1	5.6	2.1	14.3	0.8	4.2	12.1	3.1
ViLT	24.9	20.8	11.2	6.3	2.1	14.9	0.9	3.7	12.8	2.4
CLIP	27.2	19.7	13.5	3.5	1.7	14.5	1.1	3.9	11.3	3.5
Avg.	25.3	20.3	12.3	5.0	2.0	14.8	1.0	3.9	12.4	3.0

(b) **Detailed errors by models.**

Table B.3: **Error types.** The percentage of the error types by models. There is no statistical significant difference by models.

ification. However, as we will discuss in depth later, we emphasize that the model power is not only factor to consider: model biases emerge in MITL-produced datasets regardless of the strength of the model.

We additionally show the rationales for the uncertain matches and the specification of the errors in Table B.3. We observe that the models result in similar patterns in the annotations' rationale and specification of errors. Finally, by our annotation process, the average number of "100% YES" and "Partially YES" images for each caption is 8.3 and 7.1, respectively. It is remarkable since original COCO annotations allow only one image to be positively paired to a caption, revealing the massive amount of missed positive matches.

C ECCV Caption Post-processing Details

C.1 The full list of invalid captions and images

In this subsection, we list up the invalid captions and images in the original COCO test split. We filter the invalid captions by the following process: (1) We first list up the "true positive" (*i.e.*, the positive pairs in the original COCO test set) annotated by "100% No" items by Turkers or CxC [30]. (2) We manually validate the items into two categories: totally wrong captions (*e.g.*, "I don't know" captions) and semantically incorrect captions (*e.g.*, "A group of birds flying above the beach" for an image with kites), The full list of invalid captions with their COCO caption ids are as follows:

- 607516 The first picture is blank all the time on purpose.
- 607486 Why is my first one a blank every time.
- 433639 There is no image here to provide a caption for.
- 248212 I am unable to see an image above.
- 469834 There is no image here to provide a caption for.
- 462530 I really cant see this image very well.
- 469102 There is no image to be reviewed on this hit.
- 743575 There is no image showing on this page to describe.
- 246706 I am unable to see an image above.
- 61717 There is no picture here to describe with a caption.
- 500797 I am unable to see the image above.
- 19273 There is no image for me to write about.
- 630298 There is no image to provide a caption for.
- 576409 I am unable to see the image above.
- 390637 I am unable to see an image above.
- 296557 There is no image here to provide a caption for.
- 450553 I am unable to see an image above.
- 44809 blank image with no pictures available to write about

We also show the list of semantically incorrect captions as the follows:

- 610564 An individual is in the open view in the image.
- 359139 I cant tell if the bears may be fighting or kissing.
- 218995 A baseball player hugging another player as lovers do.
- 609235 Individuals are up and doing something fun today.
- 143250 The bar of the small bathroom has many remotes on it.
- 375316 A photo duplicated a few times and put together.
- 712683 Talk about a bad hair day, his is frightful.
- 75083 A group of birds flying above the beach.
- 625605 It is always wise to have bottles of water on hand in case of an emergency.
- 620511 If the motorcycle brakes down, the bicycle will be good transportation.
- 613949 A full view of an outdoor space with many things to see.
- 129825 A picture of a comment that is open.
- 605566 There is a room with various items in the picture.
- 634829 That thing is really red and slow lol

Finally, we omit `COCO_val2014_000000578492.jpg` from our test set, where the image is duplicated to training images: `COCO_train2014_000000388662.jpg` and `COCO_train2014_000000397819.jpg`.

C.2 More examples of ECCV Caption

We illustrate the samples from ECCV Caption in Figure C.2.

D More Evaluation Results on ECCV Caption

D.1 User study for evaluation metrics

In this subsection, we describe the details of the user study to compare mAP@R and Recall@k in terms of human judgement. We first randomly sample 40 captions from the captions whose the number of corresponding images is between 5

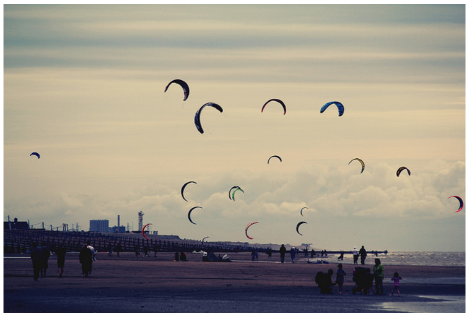


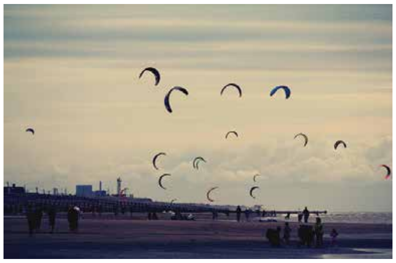
Fig. C.1: **An example image of semantically wrong captions.** We annotate "A group of birds flying above the beach" as a wrong caption of the figure, while the other captions are available in Figure C.2.

to 8. Then, we construct five rankings for each caption: (A) only top-1 is wrong (B) only top-1 is correct (C) top1 to 5 are wrong (D) only top-5 is correct, and (E) all items are wrong. When the number of corresponding images is 5, then we treat (C) as (E). Each ranking system shows different mAP@R and Recall@k; if we assume the number of the positives is 8, then (A) shows 0 R@1, 100 R@5 and 66.0 mAP@R, (B) shows 100 R@k and 12.5 mAP@R, (C) shows 0 R@k and 10.3 mAP@R, and (D) shows 0 R@1, 100 R@5 and 2.5 mAP@R. The examples of each ranking system is illustrated in Figure D.1.

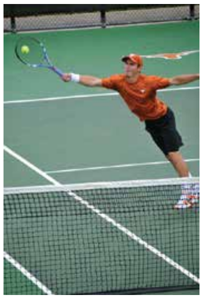
We collect binary preferences for the all possible combinations of (A) to (E), namely 10 binary pairs. We use MTurk for collecting participants, and we collect 8 participants for each question (the example question is in Figure D.2). As a result, we collect $40 \times 10 \times 8 = 3,200$ binary preferences of the five different rankings. We list the full binary preference in Table D.1. After collecting binary preferences, we restore the preference rankings using Bradley–Terry (BT) model [4]. The BT model assumes that for the given pair i and j, the probability to the pairwise comparison $i > j$ is linear to the true ranking score, *i.e.*, $P(i > j) = \frac{p_i}{p_i + p_j}$. Our goal is to estimate p_i, the true ranking preference for each method. Using BT model, we got the following results: A (70.85), B (13.15), C (10.66), D (4.89). This confirms that mAP@R is more aligned to humans than R@1; (A) shows 0 R@1 and B shows 100 R@1, however humans prefer (A) to (B) where (A) has higher mAP@R than (B) (66.0 and 12.5 if the number of positives is 8).

D.2 Training details

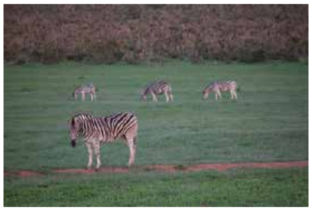
We follow the implementation details by Chun *et al.* [7]. We use the AdamP [16] optimizer and cosine annealing learning rate scheduling [26]. For re-implemented VSE, PVSE and PCME, we use the pre-trained ResNet-152 backbone and pre-trained Glove vectors following previous studies [7, 13, 38]. We use two-stage training scheme that includes "pre-training" (freezing pre-trained backbones, but only updating the additional modules) and "fine-tuning" (updating the whole



Kites being flown along the coast line in the morning
People watch multi colored items fly above the beach.
A beach where people are flying kites at sunset.
A crowd of people flying kites on the beach.
Dozens of kite skiers out in the ocean
People in the water and parachutes overhead.
Many different sails flying over a large body of water.
There are many large kites flying above the beach.
Kites are flying over people on a beach.
A bunch of kites flying in the sky on the beach.
Several gliders floating on the ocean next to an island.
A couple flying a kite at dusk on the seaside.
People on a sunny beach flying various kites



The tennis player is extending his reach to hit the racket.
A man swings his acket to hit a tennis ball
A tennis player swinging the rackets towards the ball.
A man that is standing up and has a tennis racquet.
A man lunging to hit a tennis ball in a match
A male tennis player walking on the tennis court.
A man on a court swinging a racket at a ball.
The man is playing tennis with a racket.
A man standing on a tennis court holding a racquet.
A man on a tennis court trying to hit the ball
A man taking a swing at a tennis ball
A man taking a swing at a tennis ball
A man throwing a tennis ball in the air for him to hit it with his racket.
A man hitting a tennis ball with a racquet.
A man with a tennis racket is running on a court
A man swinging his racket to hit the ball.
The tennis player is hitting the ball with his racket
A tennis player caught jumping up to hit the ball
A man is holding a tennis racquet prepared to hit the incoming ball.
A man holding a tennis racquet as a ball clears the net.
A man with a racket prepares to hit a tennis ball
A man in shorts and a long sleeve shirt playing tennis.
A man stands on a tennis court hitting a ball with a racket.
A man plays a game of tennis during the day.
A man with a tennis racket swings at a tennis ball
A man with a tennis racket on a court.
A man playing tennis on the tennis court
A person hitting a tennis ball with a tennis racket
A man playing tennis and holding back his racket to hit the ball.
A male tennis player swinging his tennis racket.
A man swinging at a tennis ball with a tennis racket.
A person hitting a tennis ball with a tennis racket.
A man on a tennis court that has a racquet.
A man in a head band hits a tennis ball
A man standing on top of a tennis court holding a racquet.
A male in a blue shirt playing tennis on a tennis court
A man holding a tennis racket on a tennis court.
A tennis player swinging a racket at a ball
A man holding a racquet on top of a tennis court.
A boy hitting a tennis ball on the tennis court.
A man on a court swinging a tennis racket.
A man in white shirt and shorts playing tennis.
A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball.
A person hitting a tennis ball with a tennis racket on a tennis court.
The man is playing tennis on the court.



Some zebras are seen grazing in the field.
Four adult zebra are grazing on a field of grass.
Four zebras are grazing on grass in a pasture.
Four zebras eating grass on a field.
A herd of zebra standing on top of a lush green field.
These four zebra are walking in a field.
There is a herd of zebras standing around.
A herd of zebras walking through the grass.
Clouds with a rainbow in the sky of an open field with zebras grazing on the grass.
Several zebras in an open area during a not so sunny day.
A group of zebras that are standing in the grass.
A group of zebras in a grassy and forested area
Four zebras are grazing at a nature reserve.
Zebras graze in a grass and bush fenced enclosure.
There are some zebras standing in a grassy field
The zebra is standing in the field with the other animals in the background.
Four zebras standing in the grass on a cloudy day.
Four zebras walking in a grassy area.
A herd of zebras stand on a pathway near brown grass.
Zebras graze on the plains with trees in the background.
Three zebras and two other animals grazing.
Several zebras walking the terrain of hills and mountains
Two zebras in some brown and green grass and some bushes
A herd of zebras grazing with a rainbow behind.
A herd of zebra grazing on a dry grass field.
Herd of five zebras grazing in a field
Two black and white zebras and some green grass and trees
A group of zebra standing on top of a dry grass field.
A zebra is standing outside on the grass by itself.
Three zebra in the middle of a field with a body of water in the distance.
Zebras grazing on sparse grass in an enclosure at an animal park.
A herd of zebras grazing in a field and a rainbow.
A group of zebras are on some grass with trees and bushes behind them.
A few deer and a zebra on a grass field
Several zebras eat the green grass in the pasture.

Query: Boats are traveling in the large open water.

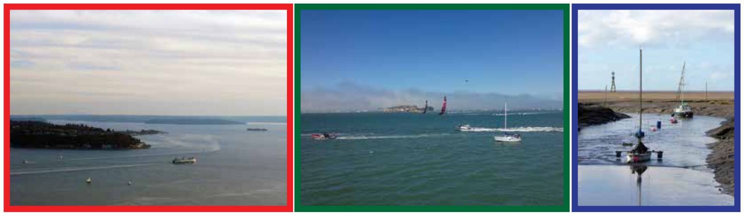
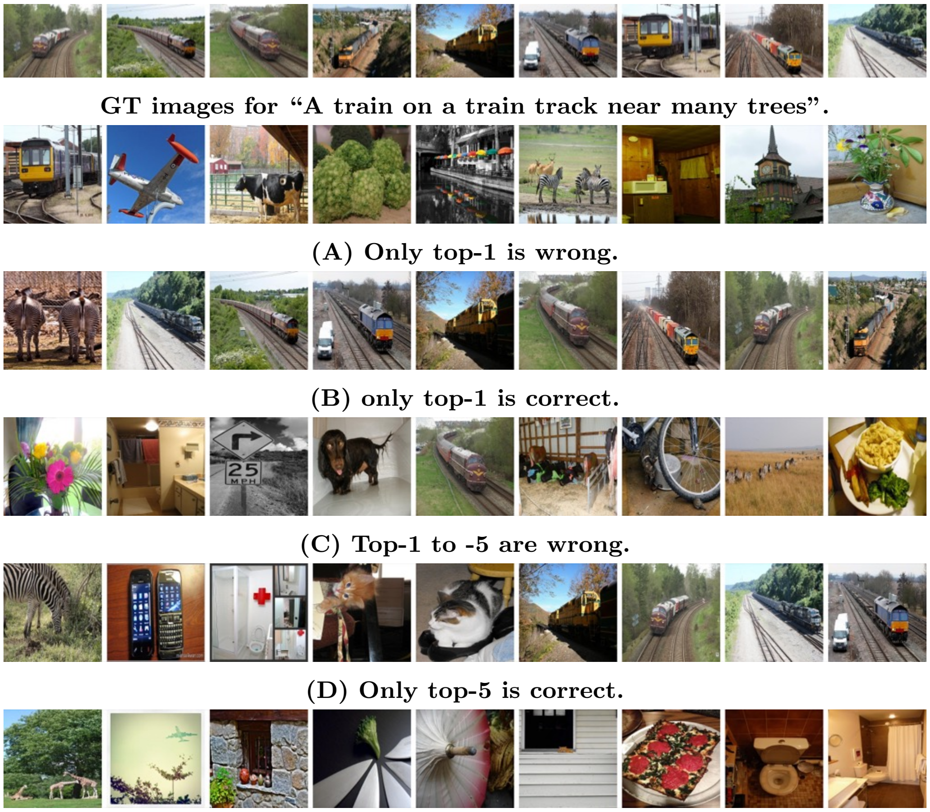


Fig. C.2: **Example sample from ECCV Caption**. Positive captions and images in ECCV Caption. Red: original positive. Green: annotated as “100% Yes”. Blue: annotated as “Weak Yes”.



GT images for "A train on a train track near many trees".

(A) Only top-1 is wrong.

(B) only top-1 is correct.

(C) Top-1 to -5 are wrong.

(D) Only top-5 is correct.

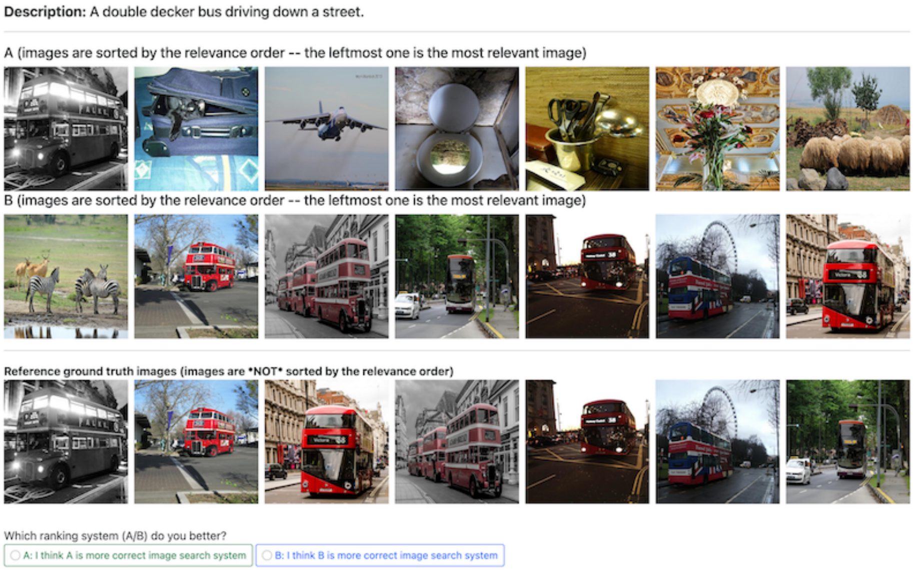
(E) All items are wrong.

Fig. D.1: **Examples of five ranking systems compared by our user study.**

	A	B	C	D	E
A: only top-1 is wrong	0	231	66	111	316
B: only top-1 is correct	89	0	21	79	328
C: top1 to 5 are wrong	254	299	0	273	343
D: only top-5 is correct	185	217	23	0	287
E: all items are wrong	28	16	1	9	24

Table D.1: **Binary preferences for five ranking systems.** Each number in row i and column j denotes that the number of preferences $i > j$. For example, 231 responses preferred "(A) only top-1 is wrong" than "(B) only top-1 is correct", while the number of the converse case is 89.

parameters). The models are pre-trained during 30 epochs and fine-tuned during 30 epochs. For the improved PCME, we use the ResNet-152 model trained with the CutMix [40] augmentation for achieving better R@1 accuracies.



Description: A double decker bus driving down a street.

A (images are sorted by the relevance order -- the leftmost one is the most relevant image)

B (images are sorted by the relevance order -- the leftmost one is the most relevant image)

Reference ground truth images (images are *NOT* sorted by the relevance order)

Which ranking system (A/B) do you better?

○ A: I think A is more correct image search system

○ B: I think B is more correct image search system

Fig. D.2: **Example question in the user study for evaluation metric comparisons.** The question asks which ranking system A or B looks more correct by humans. There are 40 of such questions in each HIT. We collect 8 participants per each question.

D.3 Full table

Table D.2 and Table D.3 show the full results of each model for image-to-text retrieval tasks and text-to-image retrieval tasks, respectively.

E Analysis of Biases in MITL

In this section, We explore and quantify the effect of the choice of multiple machine annotators to the dataset quality. We delve into the effect of the choice of machine annotators used for machine-in-the-loop (MITL) labeling paradigm to the dataset quality. Specifically, we are interested in the model bias, the type of bias that arises because of the pre-selection of plausible samples by the model in the annotation pipeline. We discuss the generalizability of our framework to general annotation tasks in Appendix E.1, and the definition of "bias" in our dataset process in Appendix E.2. We measure the model bias by employing the crowdsourced data as the source for evaluating the ITM models. For a perfectly unbiased data, we shall expect the identical rankings across the version of datasets collected with different models. We show the performances on different versions of the datasets using only one MITL model and provide discussions (Appendix E.3).

	ECCV			CxC	COCO 1K			COCO 5K			
	mAP@R	R-P	R@1	R@1	R@1	R@5	R@10	R@1	R@5	R@10	PMRP
ResNet-152 [15] image encoder + Bi-GRU [6] text encoder											
VSE0†	14.92	26.05	44.73	26.92	48.50	81.74	89.26	24.76	53.82	67.98	51.54
VSE++†	24.45	36.51	64.31	43.66	66.86	91.34	95.80	41.52	72.10	82.88	59.58
PVSE K=1	23.40	35.56	62.57	43.88	66.70	90.94	95.60	41.76	72.96	82.90	58.80
PVSE K=2	26.72	39.24	65.03	46.08	68.42	91.26	96.18	44.10	73.38	83.68	60.69
PCME	26.24	38.65	65.50	46.22	68.78	91.42	96.38	44.26	73.52	83.44	60.87
PCME (CutMix)†	28.55	41.23	68.75	47.14	68.72	92.58	96.36	45.04	75.12	84.66	61.85
Region features based on Bottom-up Attention [1] and SCAN [23]											
VSRN	30.77	42.89	73.83	55.06	76.20	94.76	97.86	53.02	81.12	89.42	62.41
VSRN + AOQ	30.70	42.61	73.12	56.86	77.50	95.44	98.38	55.14	83.30	90.80	62.47
CVSE	28.11	40.14	69.23	52.92	74.14	94.94	98.04	51.00	79.58	89.36	62.23
SGR	27.24	39.15	71.05	58.80	77.26	95.94	98.26	57.24	83.18	90.64	61.98
SAF	27.36	39.30	71.05	56.98	78.06	95.84	98.20	55.48	83.82	91.82	62.00
VSE infty (BUTD region)	31.36	42.78	74.86	60.12	79.64	96.38	98.60	58.34	85.32	92.34	62.98
VSE infty (BUTD grid)	31.68	43.08	76.53	60.64	80.42	96.78	98.86	59.10	85.90	92.82	62.79
VSE infty (WSL grid)	34.80	45.41	81.05	67.88	84.50	98.06	99.38	66.38	89.34	94.60	64.13
Large-scale Vision-Language pre-training											
CLIP ViT-B/32	22.39	32.61	66.06	51.68	69.26	90.92	95.00	50.14	75.00	83.42	54.40
CLIP ViT-B/16	23.67	33.96	68.68	53.84	71.56	91.40	95.82	52.30	76.78	84.64	56.46
CLIP ViT-L/14	24.00	33.84	71.37	57.98	74.20	92.84	96.58	56.32	79.32	86.60	59.79
VinVL (zero-shot)	16.17	27.31	49.64	37.06	58.06	87.54	93.70	35.18	64.36	76.30	36.80
VinVL	34.56	44.39	83.35	75.74	88.14	98.32	99.40	74.66	92.58	96.34	75.73
ViLT (zero-shot)	23.22	33.64	65.34	58.24	77.10	95.12	97.92	56.70	82.54	89.58	57.26
ViLT	27.50	38.57	73.35	62.76	80.76	96.28	98.32	61.50	86.26	92.66	62.65
BLIP	36.02	44.56	88.50	82.68	92.28	98.96	99.58	81.90	95.38	97.80	82.32
Different negative mining (NM) strategies											
PVSE K=1, Sum triplet†	22.26	34.97	56.70	37.18	60.42	89.24	94.84	35.16	66.32	79.22	37.27
PVSE K=1, SHM†	25.24	37.67	64.00	43.24	66.12	91.36	96.18	41.14	71.64	82.68	42.90
PVSE K=1, HNM†	25.24	37.75	64.55	44.78	67.18	91.94	96.16	42.70	73.56	84.04	45.36

Table D.2: **Re-evaluating VL models: Image-to-text retrieval results.**

E.1 Image caption matching problem to general annotation tasks

Many real-world applications are powered by state-of-the-art machine learning (ML) models shown to exceed human-level performances in tasks such as natural language understanding [10] and image classification [14, 15]. However, previous studies have shown that two conditions must be met for these models to perform well: massive training data and quality annotations. For example, large datasets that consist of well-curated 1M images [34], 3.5B Instagram photos [27,37], 300M web photos [11], 400M captioned images [32], 1.8B noisy captioned images [17], 15M hierarchically structured images [11, 41], and synthetic images [8, 9, 40, 42] are the key factors behind the corresponding models' success. Moreover, the annotation quality is equally important for the model performances. Mahajan *et al.* [27] showed that training on 940M images with well-processed 1.5k labels results in a model comparable to training on 3.5B images with noisy and weak 17k labels; both models show 84.2% ImageNet-1K top-1 accuracy.

An emerging pattern for obtaining quality labels for a large dataset is pipelining a machine learning model and human annotators. Expert human annotators are reliable and produce high-quality label, but they are costly to accommo-

	ECCV			CxC	COCO 1K			COCO 5K			
	mAP@R	R-P	R@1	R@1	R@1	R@5	R@10	R@1	R@5	R@10	PMRP
ResNet-152 [15] image encoder + Bi-GRU [6] text encoder											
VSE0†	30.41	40.48	66.37	21.56	39.95	74.71	84.71	19.78	46.12	59.87	42.36
VSE++†	45.57	54.49	81.91	32.24	52.76	84.79	91.92	30.05	60.11	72.95	48.94
PVSE K=1	44.55	53.41	83.93	32.88	53.49	85.11	92.08	30.64	61.37	73.62	48.32
PVSE K=2	53.80	60.6	88.44	34.27	54.91	86.50	93.12	32.15	62.79	74.84	50.36
PCME	47.97	56.99	84.08	33.95	54.65	86.25	93.17	31.80	62.17	74.59	52.54
PCME (CutMix)†	54.92	61.66	88.59	36.26	56.70	87.08	93.84	33.97	63.77	75.85	53.46
Region features based on Bottom-up Attention [1] and SCAN [23]											
VSRN	53.78	60.78	89.19	42.63	62.77	89.71	94.62	40.46	70.58	81.10	48.47
VSRN + AOQ	51.17	58.68	89.94	43.34	63.46	90.50	95.40	41.14	71.50	81.96	50.36
CVSE	46.59	54.88	84.16	38.71	59.88	89.29	94.81	36.60	68.04	79.57	50.74
SGR	44.35	52.93	86.49	42.39	62.06	89.56	94.90	40.47	69.64	80.28	51.84
SAF	44.55	53.07	85.66	42.18	62.24	89.53	94.98	40.12	69.73	80.38	52.42
VSE infty (BUTD region)	49.56	57.16	90.17	44.67	64.79	91.41	95.95	42.41	72.67	83.19	50.31
VSE infty (BUTD grid)	49.12	57.10	89.49	46.29	66.42	92.10	96.36	44.09	74.10	84.01	50.94
VSE infty (WSL grid)	50.02	57.45	91.82	53.70	72.04	93.92	97.19	51.64	79.34	87.58	51.16
Large-scale Vision-Language pre-training											
CLIP ViT-B/32	31.11	41.2	68.09	32.26	49.68	79.29	87.70	30.42	55.96	66.89	50.69
CLIP ViT-B/16	34.82	44.01	73.42	34.67	52.47	80.86	88.90	33.07	58.42	68.99	51.80
CLIP ViT-L/14	31.96	41.76	72.97	38.30	55.46	82.29	90.18	36.55	61.05	71.14	52.82
VinVL (zero-shot)	28.19	38.54	60.74	30.41	50.15	80.37	87.19	28.96	56.99	68.79	37.55
VinVL	47.06	54.71	92.19	59.77	76.62	95.15	97.91	58.12	83.20	90.08	46.79
ViLT (zero-shot)	30.46	39.97	72.65	42.45	62.26	90.70	96.21	40.56	70.00	81.07	51.90
ViLT	41.66	49.97	82.27	44.68	64.73	91.84	96.68	42.85	72.76	83.00	53.11
BLIP	45.01	52.30	93.47	65.92	79.95	95.80	97.82	64.32	85.73	91.57	53.15
Different negative mining (NM) strategies											
PVSE K=1, Sum triplet†	44.41	53.91	79.28	28.20	49.30	83.68	91.87	26.13	56.40	70.28	52.42
PVSE K=1, SHM†	48.02	57.05	83.93	33.09	53.57	85.45	92.43	30.86	60.83	73.71	49.91
PVSE K=1, HNM†	46.28	55.24	82.81	33.25	54.02	85.40	92.24	31.06	61.26	73.76	48.99

Table D.3: **Re-evaluating VL models: Text-to-image retrieval results.**

date. Strong machine annotators are relatively inexpensive, but they result in low-quality and unreliable annotations. One popular method of combining the two is feeding human annotators with machine learning model's outputs. For example, a model suggests annotations (*e.g.*, candidates of labels [22], estimated boxes [22], estimated segmentation maps [2], estimated descriptions [19]) for the given data point, and the annotators only need to confirm or fix the labels given by the machine annotators. This approach is commonly used for building a large-scale dataset, such as OpenImages [2, 22], e-Vil [19].

While rich body of discussions are available for building annotation interfaces and crowdsourcing workflows, we still lack a good understanding of the impact the underlying machine learning models have on the annotators and on the annotation results. In this research, we specifically examine the downstream effects of a common practice where researchers and practitioners consider only one "strong" model in the machine-in-the-loop annotation pipeline. This can be problematic because different models not only show different results to the human annotators, but also in different orders, bring the impartialness of the annotated dataset towards any particular model to the surface. In other words, the machine-in-the-loop annotations are not stable across model choices.

As a realistic scenario for utilizing ML models to aid annotators, we consider the COCO Caption matching task [5,25] that matches each image with sentences in a large database of captions. Due to the sheer bulk of the involved databases (123,287 images and 616,767 captions), it is infeasible for annotators to search through the matching caption. Instead, for each image, we use model-based ranking of possible captions to greatly reduce the search space for annotators. We have conducted studies with five state-of-the-art image-text matching models. Our COCO Caption matching task can be seen as the extreme version of the class label selection task where the number of possible classes is as large as the number of possible descriptions.

A decent overview of the image annotation tools is provided by Sager *et al*. [35]. The annotation task is determined along two axes: the types of inputs and the expressiveness of the labels. This paper focuses on the image inputs, one of the most frequently annotated type of data. The expressiveness and complexity of labels are directly related to the learning task being addressed. Tagging images with class labels is arguably the most common and basic form in the spectrum of label expressiveness. In the other extreme, we have the *image-caption matching task*: given an image, annotator has to search through the database of descriptions to find the one that best matches the image [5]. The caption matching task is of the same nature as image tagging: one needs to find the correct label in a list of possible labels. However, the candidate space for the possible labels is exponentially greater for the caption matching task. If the vocabulary size is V and the lengths of caption sequences are generally L, the size of the candidate space is as large as $O(V^L)$. This contrasts against the number of possible class labels that are generally far smaller than V.

We consider the image-caption matching as a testbed for analyzing annotation pipelines for two reasons. First, it highly relevant for the MITL annotation paradigm because it is downright infeasible for humans to browse through the database. Second, it inherits the same tool as image tagging, making our experimental results and analyses transferable to general image tagging tasks.

E.2 What Do We Mean by "Bias"?

Bias is an overloaded term with multiple senses. We briefly explore its use in relevant fields and make a definition relevant to our paper.

In statistics and machine learning, bias of an estimator or model refers to the mismatch between its average behavior and the true parameter or underlying function [3, 18]. We partially adopt this definition of bias in a broad sense. The annotation pipeline as a whole can be regarded as a mechanism for assigning plausible labels to a given set of images. When we say that the annotation pipeline is "biased", we refer to the discrepancy between the resulting annotations and the true, underlying labels for the samples.

In human-related studies, like psychology, neuroscience, human-computer interaction, and increasingly in machine learning, the use of "bias" often points to its underlying human factor. Examples include "confirmation bias" where humans favorably select data that serve their purpose [31], "reporting bias" where crucial commonsense knowledge is overlooked [12], and "survivorship bias" where non-surviving cases are under-represented [28]. In our MITL annotation pipeline, we study the *model bias* where models hinder humans from generating an unbiased set of labels by presenting humans with only a selection of the candidate labels deemed correct by the models.

E.3 Biases in ECCV Caption

Given the crowdsourced labeled image-caption data of ECCV Caption, our aim is to analyze the degrees of bias in them, depending on the underlying model used for machine-in-the-loop (MITL) labeling paradigm. Specifically, we are interested in the model bias, the type of bias that arises because of the pre-selection of plausible samples by the model in the annotation pipeline. We measure the model bias by employing the crowdsourced data as the source for evaluating the cross-modal retrieval models. For a perfectly unbiased data, we shall expect the identical evaluation results (*i.e.*, in terms of the ranking of the methods) across the version of datasets collected with different models. In this section, we introduce the strategy to measure the model bias and present experimental results and analyses.

We measure the model bias in a labeled dataset by examining whether certain versions of datasets behaves favorably to certain models when they are used as the evaluation benchmark. To measure this, we need to introduce the specific evaluation metrics used for measuring the cross-modal retrieval performances. We use *Recall@1* and *R-Precision*.

Recall@1 is the most widely-used metric for reporting the performances of cross-modal retrieval models. To compute it, we first let $m_i(x)$ be the indicator whether the i-th retrieved item of the input x is a positive match.

$$\begin{aligned} m_i(x) &= 1 \quad \text{if } i\text{-th retrieved item of } x \text{ is a positive match;} \\ m_i(x) &= 0 \quad \text{otherwise.} \end{aligned} \tag{E.1}$$

R@1 measures whether the top-1 retrieved item is a positive match on average:

$$\text{R@1} = \frac{1}{N} \sum_{n=1}^{N} m_1(x_n). \tag{E.2}$$

Despite its popularity, Recall@1 has a serious shortcoming. As argued by Musgrave *et al.* [29], a high Recall@1 does not always guarantee high-quality retrieval results. Musgrave *et al.* have proposed to use the R-Precision as an alternative metric. Let $R(x)$ be the total number of matched items for the input x. Then, R-Precision is defined as follows:

$$\text{R-P} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{R(x)} \sum_{i=1}^{R(x)} m_i(x_n). \tag{E.3}$$

Despite its good properties, in practice, it is impossible to use R-Precision for cross-modal retrieval benchmarks because many cross-modal retrieval benchmarks only a few number of (*e.g.*, 1) positive pairs in the dataset. However, as

	Models used for annotations					
Models	PVSE	VSRN	PCME	ViLT	CLIP	All
PVSE	**76.5**	63.6	67.6	45.3	42.0	76.6
VSRN	68.0	**80.1**	69.0	51.1	47.2	**80.1**
PCME	67.7	64.3	**77.3**	46.0	44.1	77.4
ViLT	59.5	59.8	58.8	**62.0**	**49.3**	72.4
CLIP	51.7	51.5	52.6	42.8	**49.3**	64.3

(a) **Text-to-Image Recall@1.**

	Models used for annotations					
Models	PVSE	VSRN	PCME	ViLT	CLIP	All
PVSE	**53.5**	39.4	42.1	30.1	31.3	42.1
VSRN	41.8	**55.9**	41.6	33.8	35.9	**43.2**
PCME	42.9	39.9	**54.8**	31.8	33.8	43.1
ViLT	34.0	34.5	33.9	**45.5**	37.0	35.8
CLIP	31.0	31.5	31.9	29.6	**39.6**	32.1

(b) **Text-to-Image R-Precision.**

Table E.1: **Model performances vs. different annotation processes.** Each row indicates performances of the same model by different annotation strategies: using the annotations filtered by a specific model. For example, the first column of the tables shows the model performances by only using "PVSE" filtered annotation. "All" denotes the full annotations are used. The bold numbers denote the best model performances for each annotation strategy, where the best performed model and the model used for the annotation strategy are the same in all experiments.

shown in Figure A.1, there actually are many plausible positive pairs missed by the original positive pairs.

We further refurbish the metrics by using the fine-grained degrees of pair positivity provided by the annotators (Appendix B.2), which are not available for conventional cross-modal retrieval datasets. We update the matching function (Equation (E.1)) as follows:

$$\begin{aligned} &m_i(x) = 1 \quad \text{if } i\text{-th retrieved item of } x \text{ is annotated as "100\% YES"} \\ &m_i(x) = 0.5 \quad \text{if } i\text{-th retrieved item of } x \text{ is annotated as "Partially YES"} \\ &m_i(x) = 0 \quad \text{otherwise.} \end{aligned} \tag{E.4}$$

We report the modified Recall@1 and R-Precision using the matching function above as the final performance metric for cross-modal retrieval models on our crowdsourced datasets.

Table E.1a and Table E.1b show the performances on different versions of the datasets using only one MITL model. Not surprisingly, we observe that the

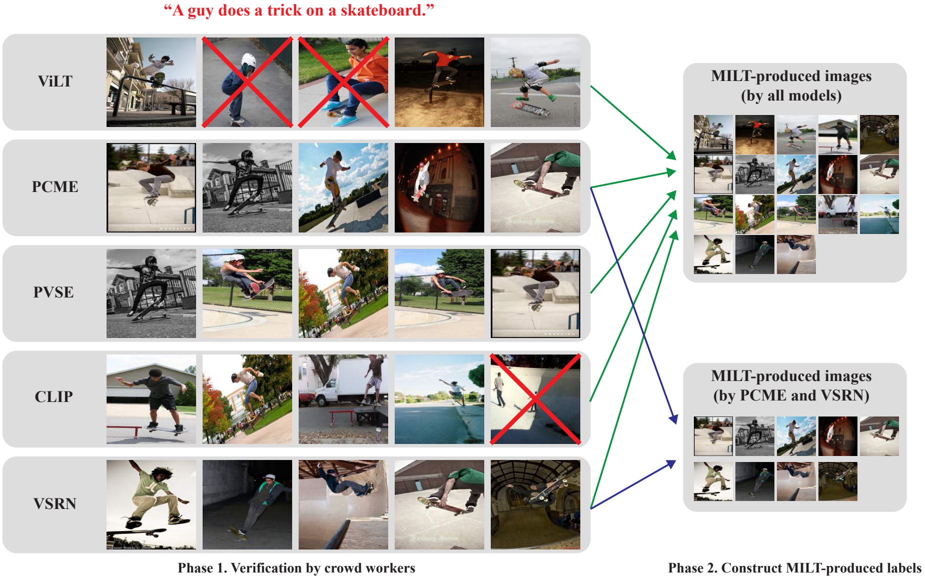

"A guy does a trick on a skateboard."
ViLT
PCME
PVSE
CLIP
VSRN
MILT-produced images (by all models)
MILT-produced images (by PCME and VSRN)
Phase 1. Verification by crowd workers
Phase 2. Construct MILT-produced labels

Fig. E.1: **Overview of our machine-in-the-loop annotation process.** We choose the subset of the verified image caption pairs by crowd workers to control the effect by the models to final annotations.

best-performing model for each dataset coincides with the model used for the MITL label proposal (*i.e.*, the diagonal elements in Table E.1). Other models tend to show quite some drop in performance. This strongly corroborates the existence of model bias in datasets collected with the aid of machine filtering. We further observe that even for models not used for generating the label proposals (*i.e.*, the non-diagonal elements in Table E.1a and Table E.1b), the rankings shift with respect to the underlying MITL model. This suggests that even when one avoids the direct use of the MITL model for evaluating its own performance, one may still observe unstable evaluation results, where different MITL models arbitrarily favor different models.

This suggests that even when one avoids the direct use of the MITL model for evaluating its own performance (*i.e.*, avoiding the diagonal results), one may still observe unstable evaluation results, where different MITL models arbitrarily favor different models.

A crucial limitation in an analysis of this type is the lack of the *true labels*. The obtained versions of datasets are clearly enhanced versions compared to the original COCO Caption dataset, but they are still heavily affected by the model biases as seen above. To make a better estimate of how far the datasets are from the true labels, we introduce the *multi-model strategy* where the workers verify label proposals generated by more than two of the models involved. More specifically, a multi-model strategy involving models $\Theta = \{\text{PVSE}, \text{PCME}\}$ pools the label proposals from both PVSE and PCME and present them to the hu-

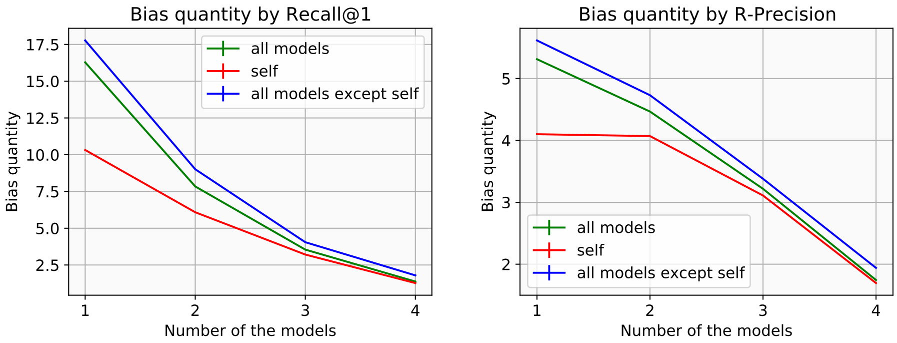


Fig. E.2: **Number of models vs. bias quantity.** The bias quantities (Eq. (E.5)) with changing the number of models for the annotation filtering process are shown. For both Recall@1 and R-Precision metrics, we observe that using more models reduce the severity of bias; the discrepancy between the resulting annotations by models and the underlying labels for the samples decreases.

man annotators. The rest of the verification process is identical as before. The intuition is that the dataset built with the multiple models will be much closer to the true labels for the image-caption matches. In the extreme case, we have the *all-model strategy* involving all five models considered in this work. See the "All" columns in Table E.1a and Table E.1b for the corresponding results.

Based on this intuition, we additionally measure the distance between a version of dataset and the all-model dataset, which is deemed to contain the labels that are closer to the true labels. We define $s_\Theta(\phi)$ as the performance of the model ϕ evaluated upon the dataset built with the multi-model strategy with MITL models Θ. We write $s_{\text{All}}(\phi)$ as a good proxy for the true performance of the model ϕ. We define the model bias incurred by a subset of models Θ as

$$\mathcal{B}_\Theta := \frac{1}{5} \sum_{\phi \in \text{All}} |s_\Theta(\phi) - s_{\text{All}}(\phi)| \tag{E.5}$$

where All $= \{$PVSE, VSRN, PCEM, ViLT, CLIP$\}$. For example, $\mathcal{B}_{\{\text{PVSE}\}}$ using Recall@1 (Table E.1a) is computed as follows:

$$\begin{aligned}\mathcal{B}_{\{\text{PVSE}\}} &= (|76.5 - 76.6| + |68.0 - 80.1| + |67.7 - 77.4| \\ &\quad + |59.5 - 72.4| + |51.7 - 64.3|)/5 = 9.5\end{aligned} \tag{E.6}$$

We break down the degree of bias $\mathcal{B}_\Theta$ into the bias incurred onto oneself ("self-bias") and the one incurred onto the other models ("non-self-bias"). We quantify the self-bias for a set of models Θ to measure the amount of the bias incurred onto oneself: $\frac{1}{|\Theta|}\sum_{\phi\in\Theta}|s_\Theta(\phi) - s_{\text{All}}(\phi)|$. For example, self-bias for PVSE is $|76.5 - 76.6| = 0.1$. The complementary amount of bias, the non-self-bias is computed similarly. For example, PVSE's non-self-bias is computed as $|68.0-80.1|+|67.7-77.4| + |59.5 - 72.4| + |51.7 - 64.3|)/4 = 11.8$.

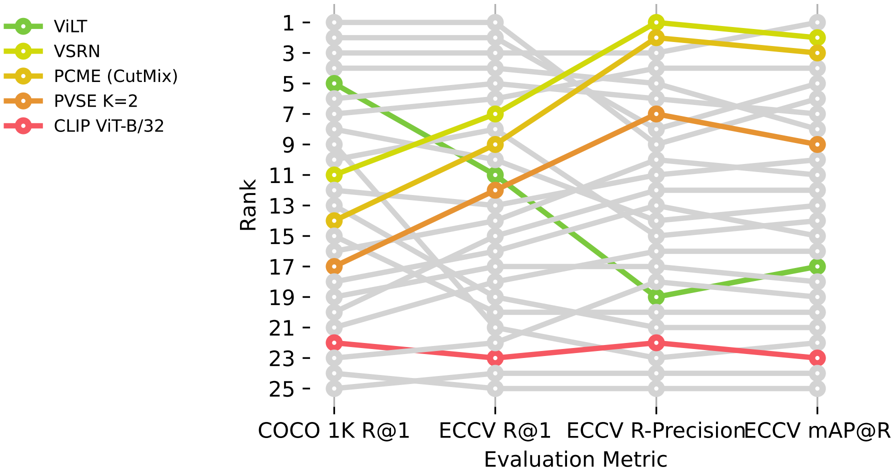


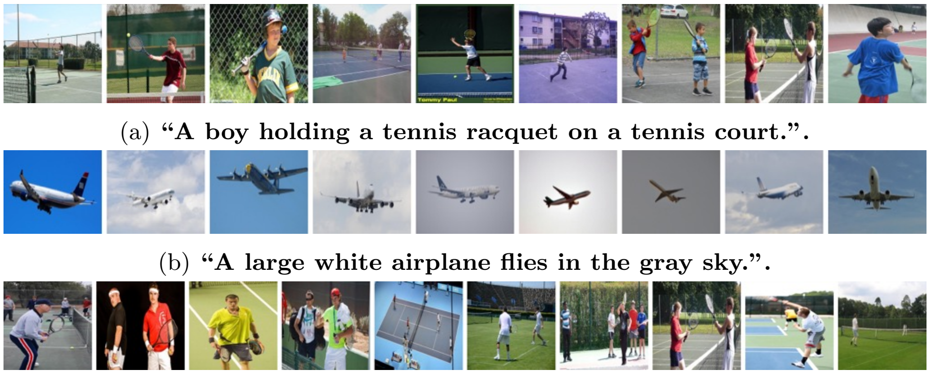
Fig. E.3: **Bias quantity in the dataset.** The full rankings of the chosen MITL annotators on our benchmark.

We plot the degrees of the biases, measured with $\mathcal{B}_\Theta$, self-bias, and non-self-bias in Figure E.2. We have experimented with changing the numbers of models $|\Theta|$. All numbers are averaged over all possible subsets of models of size $|\Theta|$ (*e.g.*, if $|\Theta| = 2$, the result is the average of $n(n-1)/2$ numbers). We omit the case with $|\Theta| = 5$ because all metrics are zero by definition.

We have two observations. First, the smaller number of MITL models make the gap between “self-bias” and “model-bias” larger. This implies that if we use a single model for the MITL annotations, the resulting dataset is highly unlikely to treat the methods differently. Second, the general bias measurements decrease with the number of involved models in the MITL annotation process. For the practical applications, hence, it is advisable to use multiple models to collect the label candidates to verify. In practice, we observe that only three machine annotators (PVSE, PCME and VSRN) achieve better rankings on ECCV mAP@R compared to the COCO R@1 ranking (Figure E.3). That suggests that our ECCV Caption is not fully biased towards the selected machine annotators.

F Noisy Annotations

Our annotations are built upon crowdsoure annotations. Due to the nature of the noisiness of crowdsource annotations, ECCV Caption contains some wrong annotations. Also, our annotations are chosen not only from “100% YES” but also from “Partially YES”. Note that our HIT is designed for specifying the details of what makes the annotation “partially correct” – See Figure B.1. We illustrate some false positive cases in Figure F.1. The false positives can be happened due to (1) wrong object, *e.g.*, “baseball bat” instead of “tennis racquet”



(a) **"A boy holding a tennis racquet on a tennis court.".**

(b) **"A large white airplane flies in the gray sky.".**

(c) **"Two men of some sort on a tennis court.".**

Fig. F.1: **Examples of noisy annotations in ECCV Caption.** Examples of false positive images are shown. Each of false positive contains (a) wrong object, *e.g.*, "baseball bat" instead of "tennis racquet" (b) wrong color, *e.g.*, "blue" instead of "gray" (c) wrong quantity, *e.g.*, "one" instead of "two".

(Figure F.1a), (2) wrong color, *e.g.*, "blue" instead of "gray" (Figure F.1b) (c) wrong quantity, *e.g.*, "one" instead of "two" (Figure F.1c). However, although there exist some false positives in our dataset, we strongly encourage to use ECCV Caption and mAP@R for evaluating a new VL model. Even if there exist some false positives, they are not 100% wrong examples; if a model learns good global ranking, then the partially correct examples should be located in the top rankings than other random items. Therefore we strongly encourage to use mAP@R instead of Recall@k; mAP@R can mitigate the error by false positives, while Recall@k can amplify errors by noisy annotations by only checking whether the top-K retrieved items are in the true items or not.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proc. CVPR (2018) 1, 13, 14
2. Benenson, R., Popov, S., Ferrari, V.: Large-scale interactive object segmentation with human annotators. In: Proc. CVPR. pp. 11700–11709 (2019) 14
3. Bishop, C.M.: Pattern recognition. Machine learning **128**(9) (2006) 15
4. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952) 9
5. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 15
6. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014) 13, 14

7. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Proc. CVPR (2021) 1, 9
8. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proc. CVPR. pp. 113–123 (2019) 13
9. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proc. CVPR Worshops. pp. 702–703 (2020) 13
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, https://doi.org/10.18653/v1/n19-1423 13
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. ICLR (2021), https://openreview.net/forum?id=YicbFdNTTy 13
12. Easterbrook, P.J., Gopalan, R., Berlin, J., Matthews, D.R.: Publication bias in clinical research. The Lancet **337**(8746), 867–872 (1991) 16
13. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. In: Proc. BMVC (2018) 1, 9
14. Geirhos, R., Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. In: Advances in Neural Information Processing Systems. pp. 7538–7550 (2018) 13
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR (2016) 13, 14
16. Heo, B., Chun, S., Oh, S.J., Han, D., Yun, S., Kim, G., Uh, Y., Ha, J.W.: Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In: Proc. ICLR (2021) 9
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. ICML. pp. 4904–4916. PMLR (2021) 13
18. Johnson, R.A., Miller, I., Freund, J.E.: Probability and statistics for engineers, vol. 2000. Pearson Education London (2000) 15
19. Kayser, M., Camburu, O.M., Salewski, L., Emde, C., Do, V., Akata, Z., Lukasiewicz, T.: e-vil: A dataset and benchmark for natural language explanations in vision-language tasks (2021) 14
20. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Proc. ICML (2021) 2
21. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017) 2
22. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. IJCV **128**(7), 1956–1981 (2020) 14
23. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proc. ECCV (2018) 13, 14
24. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proc. ICCV. pp. 4654–4662 (2019) 1

25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. ECCV (2014) 15
26. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: Proc. ICLR (2017) 9
27. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proc. ECCV. pp. 181–196 (2018) 13
28. Mangel, M., Samaniego, F.J.: Abraham wald's work on aircraft survivability. Journal of the American Statistical Association **79**(386), 259–267 (1984) 16
29. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Proc. ECCV (2020) 16
30. Parekh, Z., Baldridge, J., Cer, D., Waters, A., Yang, Y.: Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. arXiv preprint arXiv:2004.15020 (2020) 7
31. Plous, S.: The psychology of judgment and decision making. Mcgraw-Hill Book Company (1993) 15
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proc. ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), http://proceedings.mlr.press/v139/radford21a.html 2, 13
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proc. NeurIPS. pp. 91–99 (2015) 1
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 13
35. Sager, C., Janiesch, C., Zschech, P.: A survey of image labelling for computer vision applications. Journal of Business Analytics pp. 1–20 (2021) 15
36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proc. CVPR. pp. 815–823 (2015) 1
37. Singh, M., Gustafson, L., Adcock, A., Reis, V.d.F., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., van der Maaten, L.: Revisiting weakly supervised pre-training of visual perception models (2022) 13
38. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: Proc. CVPR. pp. 1979–1988 (2019) 1, 9
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. NeurIPS. pp. 5998–6008 (2017) 2
40. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proc. ICCV (2019) 2, 11, 13
41. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. In: Proc. CVPR (2021) 13
42. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: Proc. ICLR (2018) 13