Repurposing Existing Deep Networks for Caption and Aesthetic-Guided Image Cropping

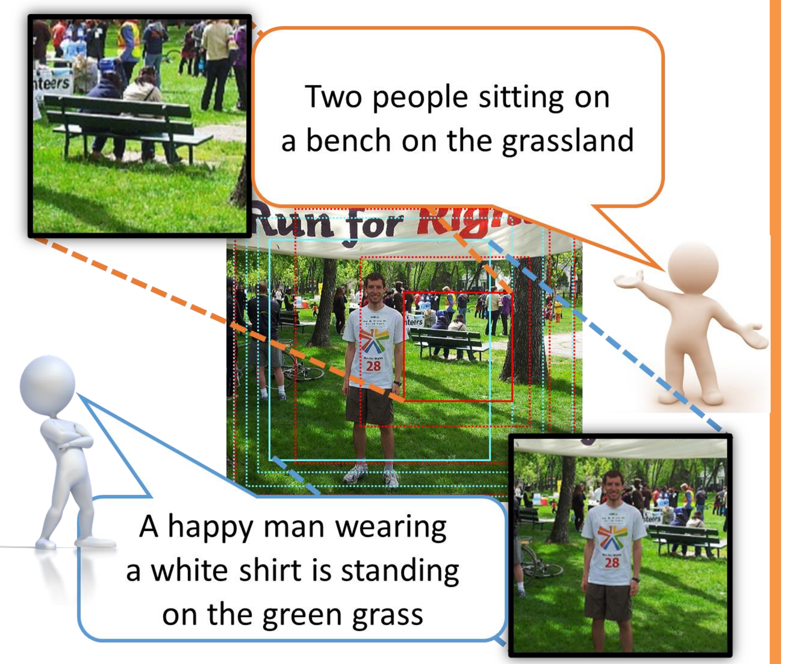
Nora Horanyi[1] Kedi Xia[2] Kwang Moo Yi[3] Abhishake Kumar Bojja[3] Ales Leonardis[1] Hyung Jin Chang[1]

[1]University of Birmingham [2]Zhejiang University [3]University of British Columbia

UNIVERSITY OF BIRMINGHAM

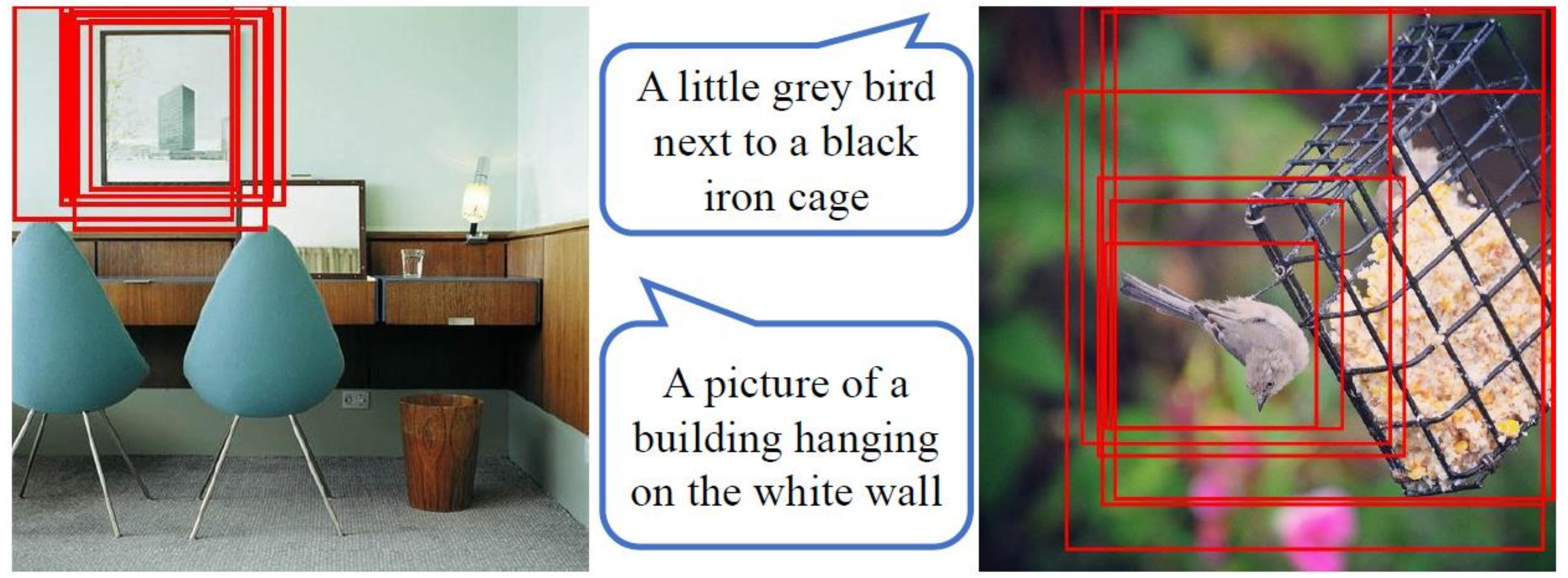
MOTIVATION

- Proposed a novel optimization framework that produces **image crops** that follow **users' descriptions** and **aesthetics criteria**.
- Achieve this without training a specialized network, utilizing two pre-trained networks on related tasks, namely image captioning and aesthetics measuring



CONTRIBUTIONS

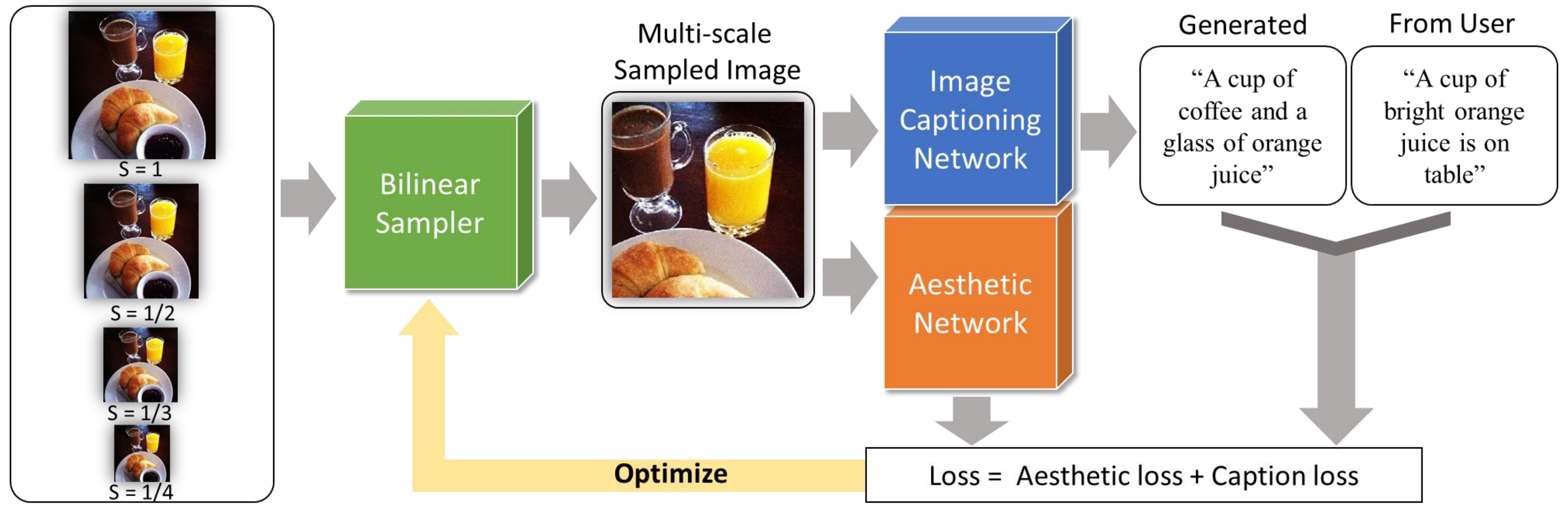
- *We propose a new **deep networks repurposing framework to optimize crop parameters directly** using a bilinear sampler, a pre-trained image captioning network , and a pre-trained aesthetic estimation network.*
- *We optimize to find the crop region that best fits the provided caption in terms of the **image captioning network losses**, as well as maximizes the **aesthetics network scores**.*
- *We generate a **new dataset** with multiple ground truth bounding box annotations for each caption.*
- *With approaches above, we were able to not only outperform state-of-the-art methods but also produce more visually pleasing image crops with well-reflecting user's intention*



ACKNOWLEDGMENT

Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant "Deep Visual Geometry Machines" (RGPIN-2018-03788), and Compute Canada; MoD/Dstl and EPSRC, Department of Defense funded MURI project through EPSRC grant EP/N019415/1; IITP grant funded by the Korea government (MSIT) (IITP-2020-0-01789, ITRC support program; NVIDIA; Institute of Information and communications Technology Planning and evaluation (IITP) grant funded by the Korean government (MSIT) (2021-0-00537, Visual common sense through self-supervised learning for restoration of invisible parts in images).

CAGIC (Caption and Aesthetic-Guided Image Cropping)



$$\mathcal{L}_{total}\left(\mathbf{I},\mathbf{y},\boldsymbol{\theta}\right) = \mathcal{L}_{caption}\left(\mathbf{I},\mathbf{y},\boldsymbol{\theta}\right) + \lambda\mathcal{L}_{aesthetic}\left(\mathbf{I},\boldsymbol{\theta}\right)$$

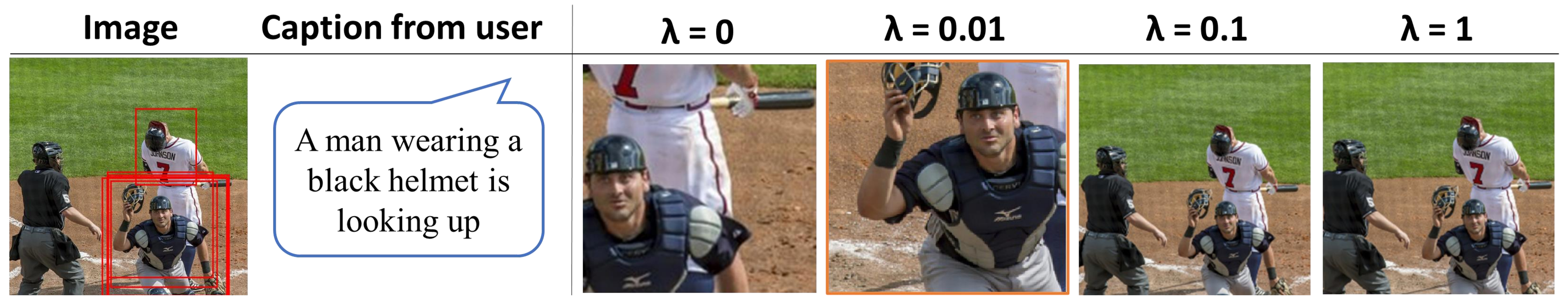
$$\mathcal{L}_{caption}\left(\mathbf{I},\mathbf{y},\boldsymbol{\theta}\right) = H\left(\frac{1}{T_u}\sum_{t=1}^{T_u}\mathbf{y}_t, \frac{1}{T_c}\sum_{t=1}^{T_c} f\left(\mathbf{I}_{crop}\left(\boldsymbol{\theta}\right)\right)_t\right)$$

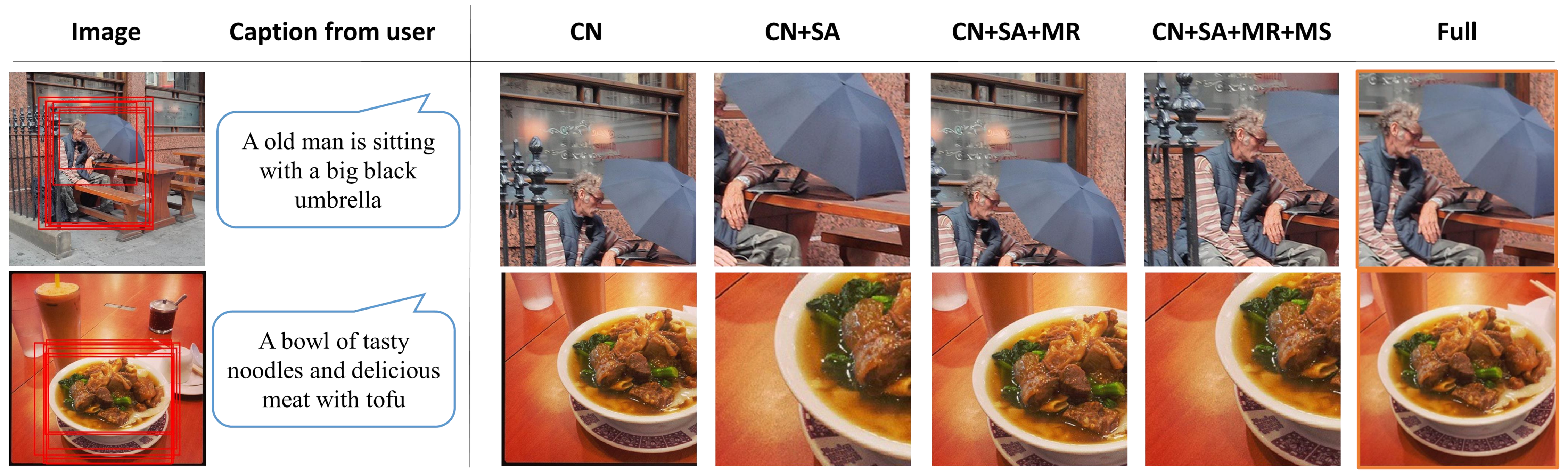
$$\mathcal{L}_{aesthetic}\left(\mathbf{I},\boldsymbol{\theta}\right) = -g\left(\mathbf{I}_{crop}\left(\boldsymbol{\theta}\right)\right)$$

Optimization and stabilization:

- Scale annealing
- Multiple restart technique

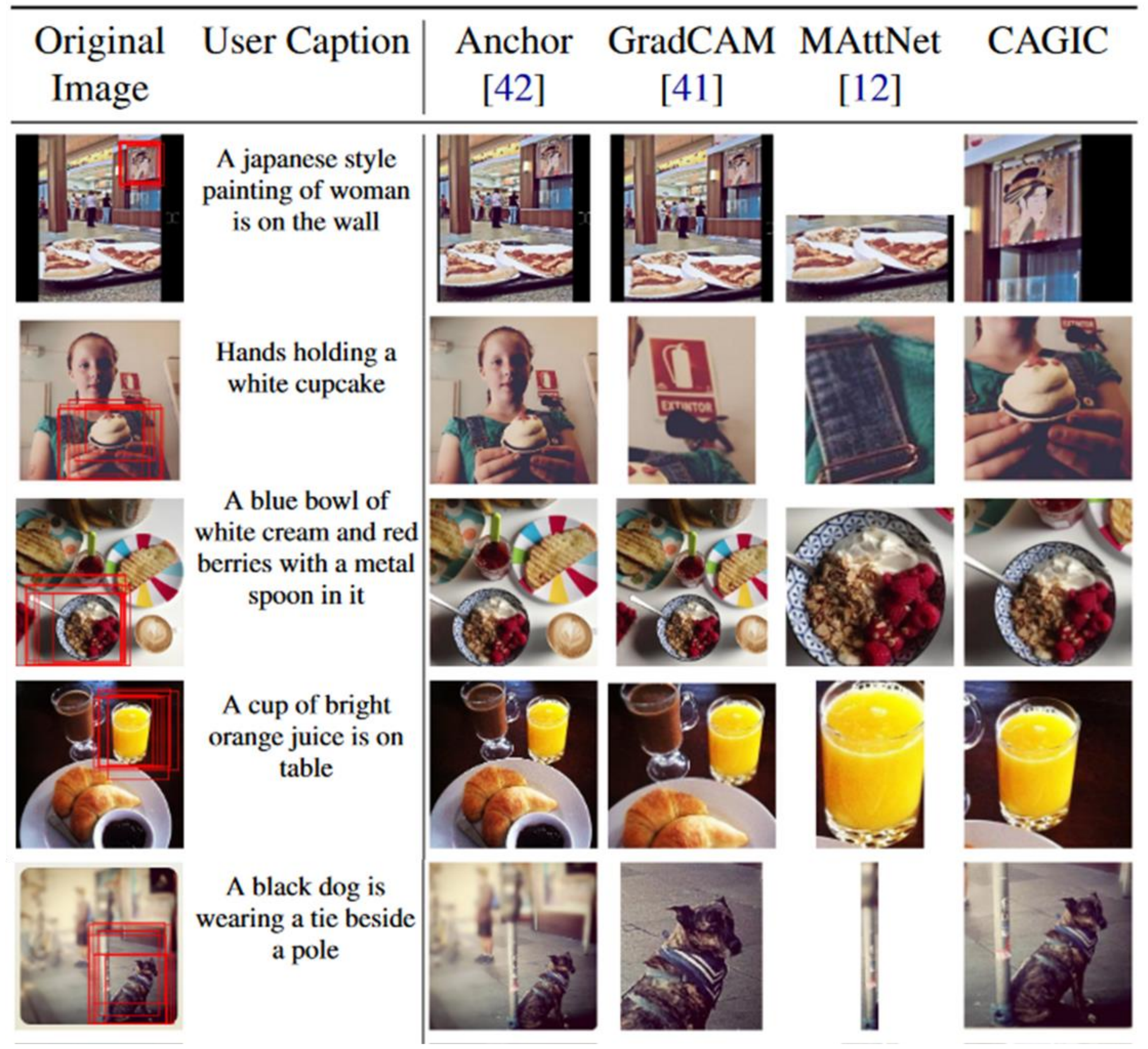
ABLATION STUDIES





RESULTS

Qualitative evaluation



Quantitative evaluation

Intersection over Union measure

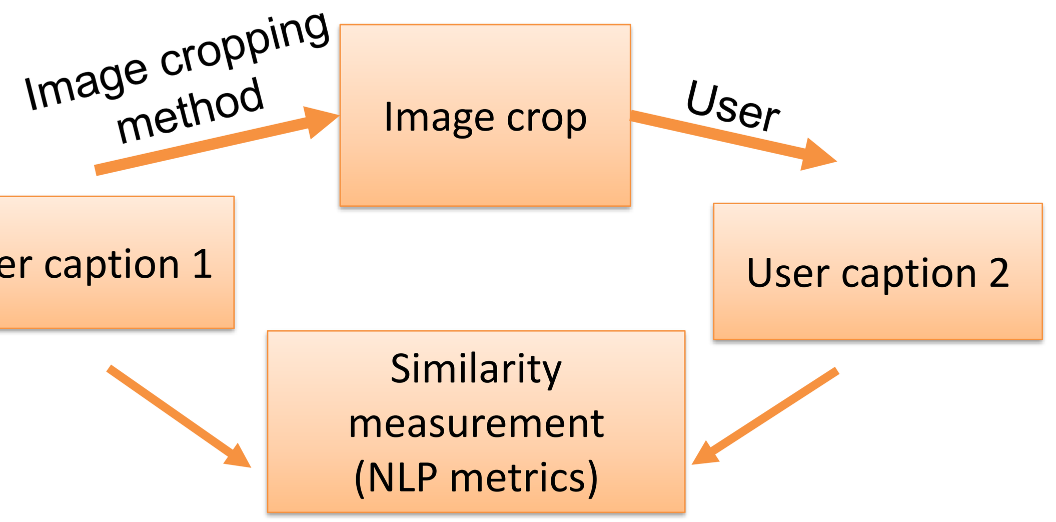
Method	Mean ± Std.
Original	0.2869 ± 0.0280
Anchor [42]	0.3325 ± 0.0236
GradCAM[41]	0.3597 ± 0.2017
MAttNet[12]	0.3851 ± 0.2607
CAGIC	**0.4160 ± 0.0129**

Which output do you like the most?

Users preferred the output crop of CAGIC over the state-of-the-art methods' crops and the original image

	Original Image	MAttNet[12]	GradCAM[41]	CAGIC
Aggregated percentage (%)	21.04	23.93	25.51	**29.52**

Is this the crop we were looking for?



- We asked users to describe the output crop and compared their caption to the original caption
- We used 6 different NPL metrics to calculate the similarities between them
- The captions describing CAGIC's output were most similar to the original caption