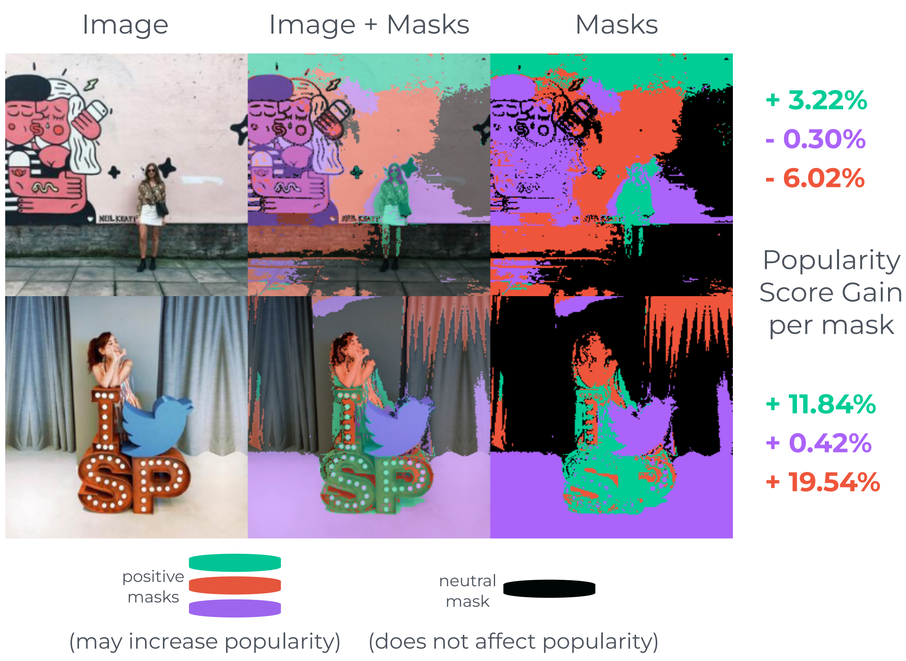
VIP: Visual Insights on Popularity

Anna Klimiuk Michał Szachniewicz

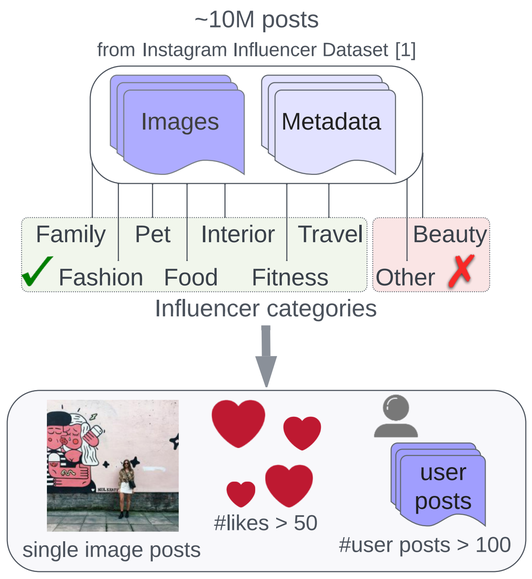
Given: only image.

Goal: estimate its intrinsic popularity and define the contribution of individual image components in a measurable way.

How: predict multiple masks covering meaningful parts in a way that lack of those masks decreases overall popularity.



Dataset preparation

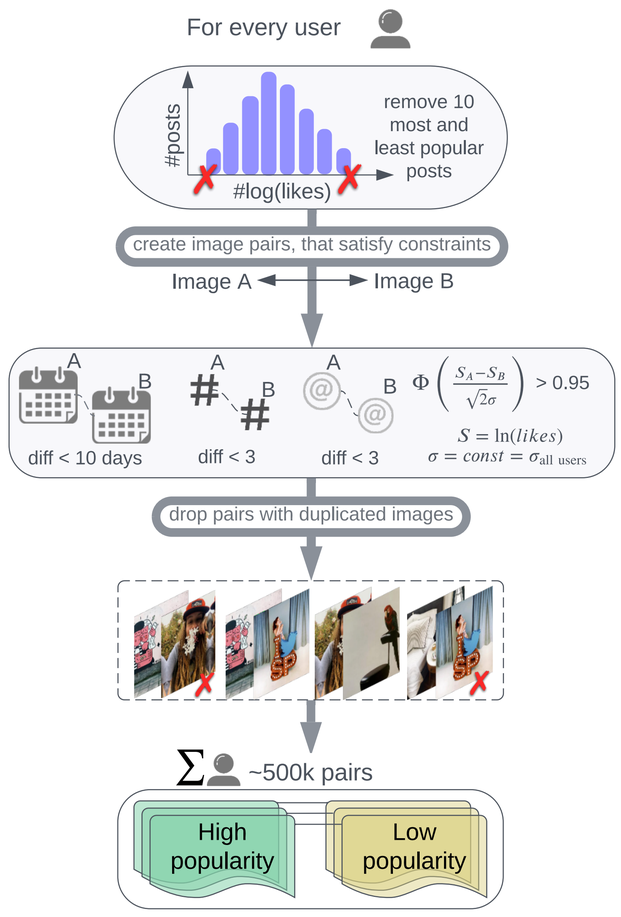


Assuming Thurstone's model [2], the logarithm of post-likes, popularity score S obeys normal distribution and can be considered ground truth for absolute image popularity. Thus, we derive the probability that image A is intrinsically more popular than image B as:

$$\Phi\left(\frac{S_A - S_B}{\sqrt{2}\sigma}\right) > 0.95$$

where σ describes variation of likes and is constant between users.

High - Low popularity pairs

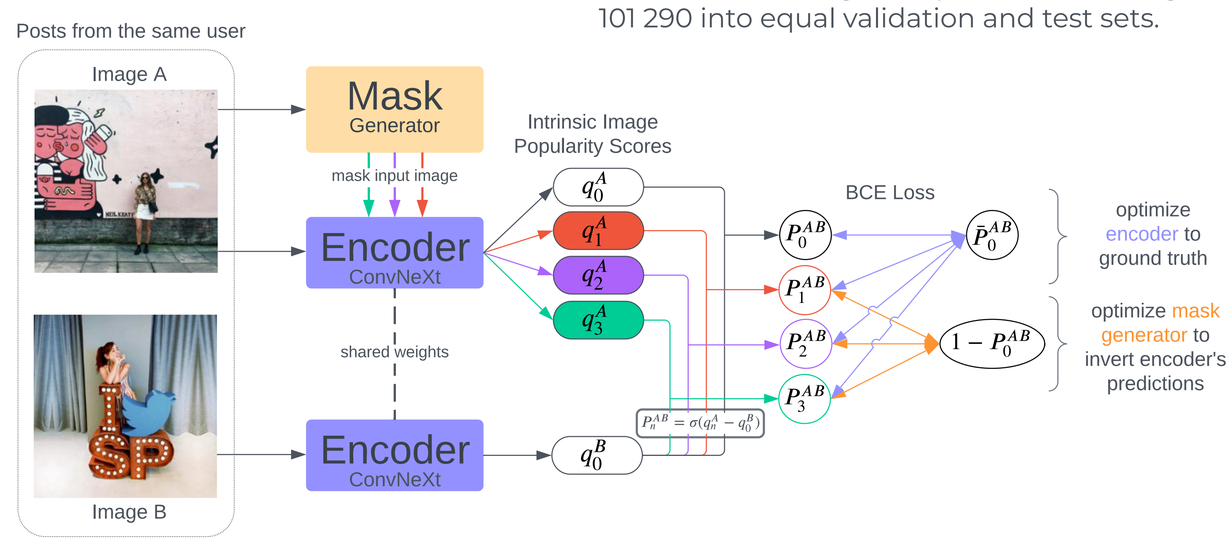


From the gathered 506 448 pairs, we used 405 158 for training and split the remaining 101 290 into equal validation and test sets.

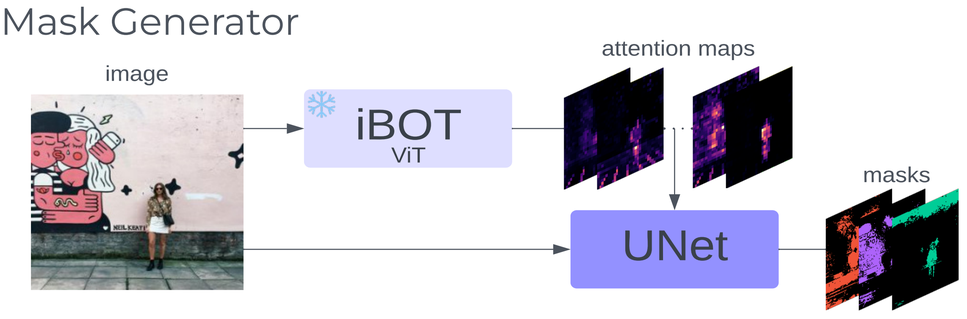
Training procedure

For images A and B satisfying given constraints:

1. generate masks for image A with UNet using attention maps extracted from the pretrained and freezed iBOT model [3] as well as input image,
2. encode image A, its masked versions and image B with the same ConvNeXt [4] encoder and obtain Intrinsic Image Popularity Scores[5]:
   - q_0^A and q_0^B for raw images,
   - q_n^A for image A with positive mask n removed
3. optimize separately:
   - Encoder to ground truth with additional resistance to masking,
   - Mask Generator to invert Encoder's predictions.



Mask Generator

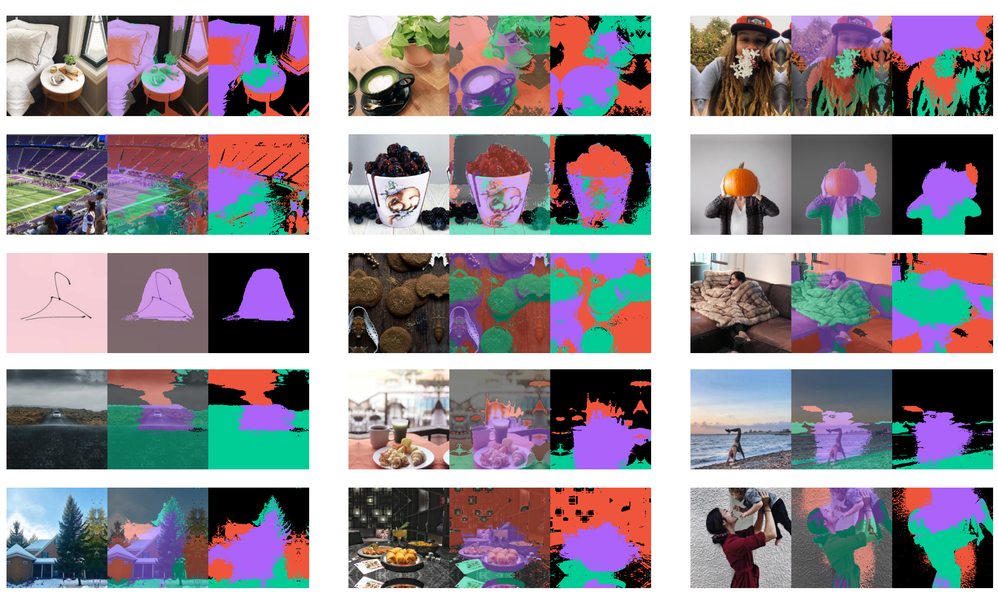


In order to generate masks that cover meaningful parts of an image, we extract attention maps from Vision Transformer pretrained in the iBOT manner and feed them to simple UNet architecture. We also generate one additional mask that is not optimized during training and expect it to behave as a neutral mask - a mask that does not influence image popularity. Additionally, we apply a sparsity loss function over masks to prevent them from collapsing.

Our adversarial training is based on the idea, that popularity score can be changed by masking meaningful parts from an image. Thus, we train ConvNeXt encoder to predict intrinsic popularity for raw as well as masked images and the UNet-based mask generator to produce masks, which occlusion can change encoders predictions. Hereby we obtain the model that is more robust to noisy popularity signals.

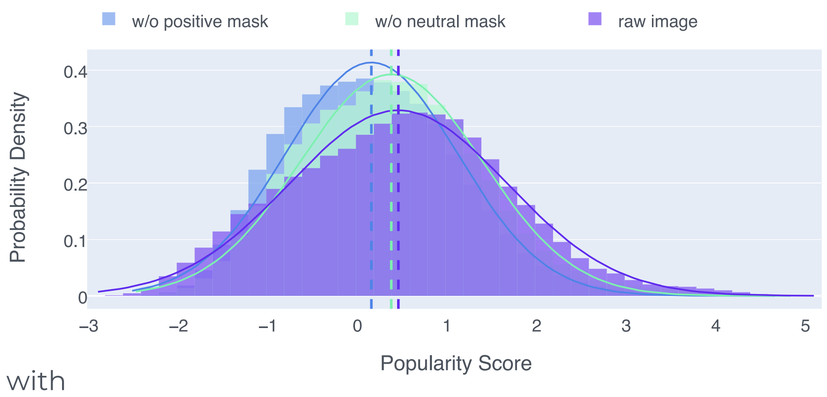
Masks visualization

We present the sample of generated masks, post-processed with fully connected conditional random fields [6].



Results

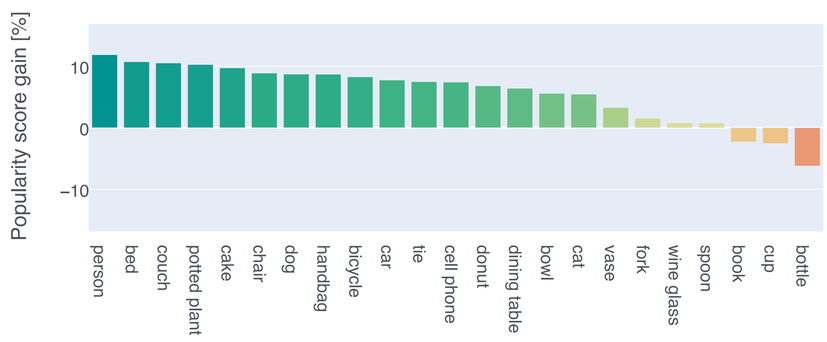
As intended, distribution of popularity scores for images with removed positive masks are left-shifted compared to raw images and for those with neutral mask removed we see that locations of scores distributions are closer.



We compared our method with base one (without adversarial masking). For that purpose we train the same ConvNeXt Nano encoder and evaluate it on raw images, as well as on images with positive and neutral masks removed. As we can see, our training is more resistant to masking, without a loss of quality on raw images.

Masking Variant / Method	Accuracy I	$I - m_1^N$	$I - m_1^P$	$I - m_2^P$	$I - m_3^P$	Matthews Correlation Coefficient I	$I - m_1^N$	$I - m_1^P$	$I - m_2^P$	$I - m_3^P$
ConvNeXt Nano	**75.96%**	70.09%	48.06%	64.60%	59.69%	**0.5191**	0.4017	-0.0387	0.2920	0.1937
+our training	75.70%	**71.92%**	**59.86%**	**68.28%**	**66.12%**	0.5140	**0.4384**	**0.1973**	**0.3657**	**0.3225**

Insights



We conduct object detection using pretrained YOLO v8 model. For every test image we match found objects with their corresponding masks and calculate percentage gain of popularity between raw image and image with objects removed according to masks. Hereby, we can assume that presence of green-shaded objects increases overall image popularity, unlike those marked in red.

References:
[1] Kim, S., Jiang, J.Y., Nakada, M., Han, J., & Wang, W. (2020). Multimodal Post Attentive Profiling for Influencer Marketing. In Proceedings of The Web Conference 2020 (pp. 2878–2884).
[2] Louis L. Thurstone. 1927. A law of comparative judgment. Psychological Review 34, 4 (1927), 273–286.
[3] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T.. (2021). iBOT: Image BERT Pre-Training with Online Tokenizer.
[4] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., & Xie, S.. (2022). A ConvNet for the 2020s.
[5] Keyan Ding, Kede Ma, & Shiqi Wang (2019). Intrinsic Image Popularity Assessment. In Proceedings of the 27th ACM International Conference on Multimedia. ACM.
[6] Krähenbühl, Philipp, and Vladlen Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." In Advances in Neural Information Processing Systems, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Vol. 24. Curran Associates, Inc., 2011.

Katedra Sztucznej Inteligencji
Politechnika Wrocławska