Neural Image Popularity Assessment with Retrieval-augmented Transformer

Liya Ji*
HKUST

Chan Ho Park*
HKUST

Zhefan Rao*
HKUST

Qifeng Chen
HKUST

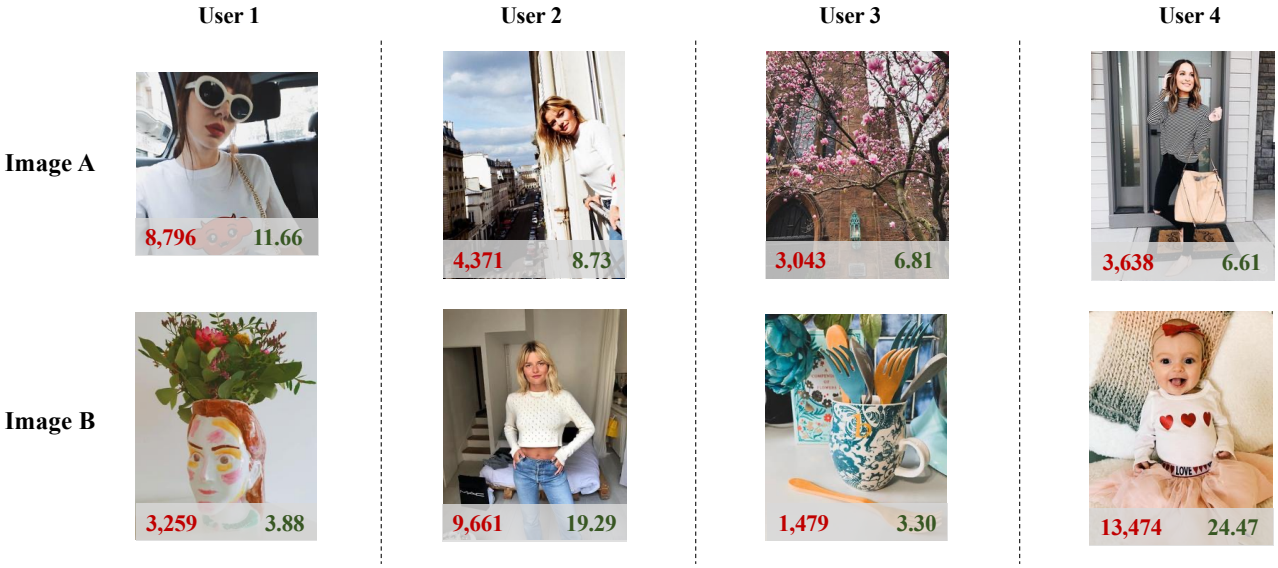

Figure 1: Neural image popularity assessment results on the Instagram dataset [29]. We show four pairs of images horizontally. Each pair of images are selected from the posts of the same person within one week. The ground-truth number of likes is stated on the bottom-left of the image. Ours predicted popularity scores (the higher, the better) are provided at the bottom-right of the images.

ABSTRACT

Since the advent of social media platforms, image selection based on social preference is a challenging task that all users inherently undertake before sharing images with the public. In our user study for this problem, human choices of images based on perceived social preference are largely inaccurate (58.7% accuracy). The challenge of this task, also known as image popularity assessment, lies in its subjective nature caused by visual and non-visual factors. Especially in the social media setting, social feedback on a particular image largely differs depending on who uploads it. Therefore social preference model should be able to account for this user-specific image aspect of the task. To address this issue, we present a retrieval-augmented approach that leverages both image features and user-specific statistics for neural image popularity assessment. User-specific statistics are derived by retrieving past images with their statistics from a memory bank. By combining these statistics with image features, our approach achieves 79.5% accuracy, which significantly outperforms human and baseline models on the pairwise ranking of images from the Instagram Influencer Dataset. Our source code will be publicly available.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Human-centered computing** → **Social media**.

KEYWORDS

Image popularity assessment, Deep neural networks, No-reference image assessment, Retrieval-augmented model

ACM Reference Format:
Liya Ji, Chan Ho Park, Zhefan Rao, and Qifeng Chen. 2023. Neural Image Popularity Assessment with Retrieval-augmented Transformer. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611918

*Joint first authors

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611918

1 INTRODUCTION

Image popularity assessment (IPA) has a wide-ranging impact on content creation, marketing, and user experience on social media. The aim of popularity assessment is to determine the level of popularity of a given image and provide valuable insights that can be utilized to improve content quality and generate business value. In the field of content creation, especially with the rising of AIGC (AI Generated Content), assessing the generated images helps individuals and companies to improve their influence and visibility online. Additionally, for marketers to make informed decisions on which images resonate with the target audiences, analysis, and utilization of social feedback on previously uploaded photos on social media is important. Hence through the use of the popularity assessment model, content creators will be able to understand which aspects of photos certain groups of users enjoy and improve their online experience accordingly.

Image popularity assessment is challenging for its subjective nature, and popularity on social media is affected by non-visual factors. The first challenge of any IPA-based problem is the subjective nature of perceptual assessment because everyone focuses on different aspects while looking at a photo. Figure 1 shows some examples from the Instagram dataset [28]. Our user study shows that in most cases, people could not select the one with higher likes despite the fact that the large difference in ground truth likes (around 2 to 3 times). Unlike other computer vision tasks like object detection or image classification, where the ground truth can be inferred objectively by humans, the subjective nature of social preference leads humans to make wrong guesses most of the time.

The second challenge is that non-visual factors affect the popularity we obtain in image-based social network platforms like Instagram. As shown in [4, 11, 15, 22], in addition to other factors like hashtag, title, and description appended to the image post, the popularity of an image is highly dependent on the user uploading the image. In other words, since the distribution of followers varies depending on the user, even with the same image, the social feedback can differ largely depending on the group of people viewing the content. Therefore constructing a methodology modeling this factor is one of the crucial elements in social media image popularity assessment.

Solving image popularity assessment with metadata achieves impressive performance but faces the limitation in some scenarios, where metadata is not available. Image popularity assessment (IPA) is addressed by multiple works [1, 12, 23, 30, 36, 38, 43, 51] under the setting where meta-data, such as the number of followers, post category, geo-location, or the number of posts are available. Through the effective fusion of multi-modal features, these models successfully model the popularity of a given post of datasets, including Flickr [45] and social media prediction challenge dataset [44]. Despite the effectiveness of image popularity assessment, models based on meta-information and multi-modalities limit the possible application where this information is unavailable. For example, in the context of content creation, the popularity assessment should be done before uploading the image to social media, ideally without any dependency on features such as hashtag, title, and length of description.

Inspired by Ding *et. al* [11], our proposed work only uses images, the timestamp, and the number of likes in the training set, having a more broad application with comparative results as the meta-data models. We propose a retrieval-augmented pipeline to extract user-specific image features and combine these features effectively utilizing a transformer-based block. Inspired by works in natural language processing [48] and image synthesis [3], utilizing retrieval-augmented techniques that provide local content, we leverage the local popularity score per train image stored in the memory bank as a mode of providing a user-specific image feature to the network. More specifically, user-specific image features of a subject image are obtained by retrieving similar images and their distribution of likes from the memory bank. Intuitively, given a subject image, appending the social feedback of the previously-posted similar images provides user-specific context for popularity assessment. This accounts for the different follower distribution and posting styles amongst a large number of users. The main difference with Ding *et. al* [11] is that we make use of the training set as a memory bank for extracting user-specific image features, which allows the model to consider both visual content and user-specific image features.

In addition, we also explore the 3D-aware features by using the depth estimation model with the motivation that the attention of users varies among the geometry of the images. In order to combine multiple features effectively, we utilize a feature aggregation block, which is derived from the self-attention module [42]. This block aggregates the different high-level features and fuses the components from images separately.

We show that our retrieval-augmented model based on the transformer architecture achieves state-of-the-art performance on the Instagram Influencer Dataset [29]. Our contributions can be summarized as:

- We approach the problem of image popularity assessment without using any meta-data at test time with comparable performance as those meta-data models.
- We propose a retrieval-augmented transformer model for neural image popularity assessment, which includes a non-parametric component that extracts user-specific image features.
- We extensively conduct the experiments on the Instagram Influencer Dataset [29] and show state-of-the-art performance with the pairwise accuracy 79.48%. We also demonstrate the empirical analysis of different datasets and potential applications of the trained model.

2 RELATED WORK

2.1 Image Popularity Assessment

Image popularity assessment (IPA) focuses on the social feedback and preference of an image that has been made public online. Multiple datasets and settings [11, 44] address this task with varying types and numbers of additional information along with the image. Social Media Prediction (SMP) challenge [44] specifically aims to understand the change of popularity over time in varying levels of temporal granularity. SMP challenge datasets contain temporal information such as timestamp information, user-specific features such as average view, title length, and other information such as

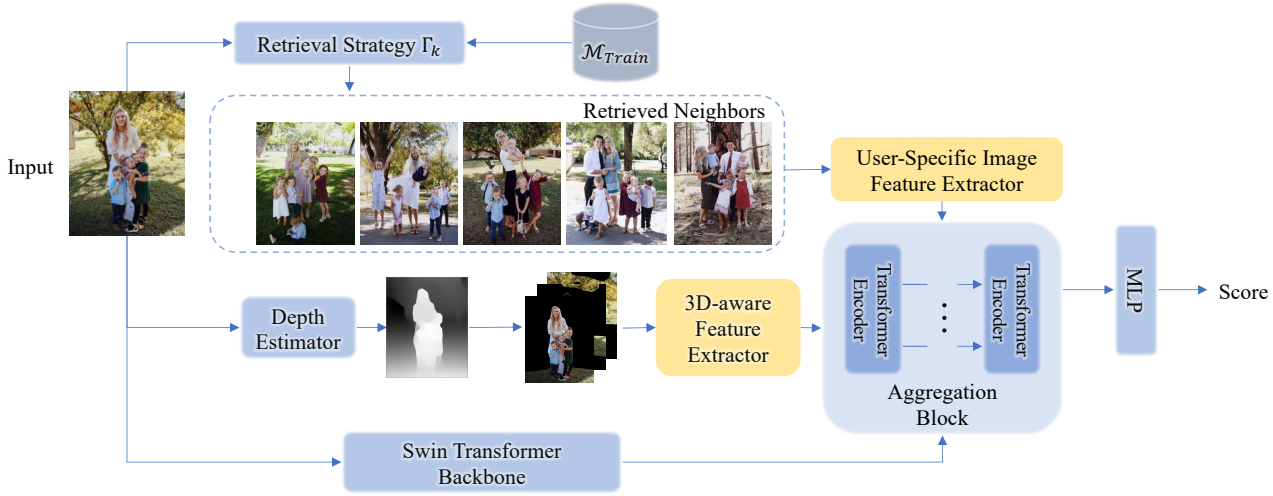

Figure 2: The framework pipeline. (a) Along with the Swin Transformer backbone, we have two auxiliary branches to extract features. One is to extract the retrieval-augmented scores, and the other is to extract the 3D-aware features using the monocular depth estimation model. $\mathcal{M}_{Train}$ stands for the memory bank. (b) The aggregation block contains a stack of transformer encoder blocks that aggregate the feature from the extractors.

geolocation and category of the post. With this dataset, multiple works [12, 24, 26] have designed a way to extract useful textual, visual, and meta-features from the dataset and fuse them together to predict the normalized log number of likes directly. Additionally, in order to extract short-term dependent time-series features, methods like [41, 43] also employ the sliding window averaging method. Some of the best-performing methods like [30, 47] additionally mine from external data sources to make the user information more comprehensive.

Regarding the fusion of features, generally, this method made use of boosting-based models like XGBoost, Catboost, LightGBM, or ensembling of multiple models to aggregate the handcrafted and deep features. Similar to different textual and visual encoders, recent works [6, 7, 22] make use of recurrent networks or attention mechanism to aggregate features of different modalities together.

In contrast to the setting above, Ding *et al.* [11] propose a method to explore the intrinsic image popularity assessment only relying on the image feature. Furthermore, the aforementioned models' target is to predict the exact number of likes or binary order of image pair while we predict the relative popularity score between two images. To our best knowledge, (I^2PA) [11] is the only work that focuses on using only visual features. By leveraging on the available data, our approach outperforms I^2PA [11] by a significant margin and under the same setting.

2.2 Image Aesthetics and Quality Assessment

Image Quality and Aesthetics Assessment are both tasks that aim to bridge the relationship between human perception and digital image. The main idea of image quality assessment (IQA) is to predict the quality of an image from a human's quality assessment point of view for downstream tasks such as image restoration and image quality monitoring systems. In contrast, Image aesthetics assessment (IAA) relates to aspects such as subject placement and artistic values that make an image look visually pleasing.

The development of IQA can be divided into two branches: reference-based methods [10] and non-reference methods [14, 62, 63]. In reference-based algorithms, we require a high-quality image as a reference to calculate the quality score. However, the ground truth high-quality reference images may not be available in practice. Thus, non-reference image quality assessment methods have been proposed to resolve this limitation. Similarly, other extensive datasets [20, 56] allow the construction of image quality assessment models in various settings such as smartphone photography or patch-level quality label available.

Works in image aesthetics assessment (IAA) can also be subcategorized into two branches. The first is the generalized image aesthetics assessment (GIAA) model, and the second is the personalized image aesthetics assessment (PIAA). Datasets such as [54, 55] enable the modeling of general and personal preference through a collection of opinion scores for a large number of images. Methodologies like [31, 35, 40] have been proposed to model aesthetic scores. NIMA [40] is one of the early deep-image feature-based generalized aesthetic score prediction models that have been trained with the AVA dataset [37], which contains a large number of images with corresponding aesthetic scores annotated by human experts. The training is done by predicting the probability mass function over the ratings 1 to 10. Recently, [31] proposed the use of a meta-learning-based strategy to generate a regression model to substitute for fine-tuning, which is a common mode of personalization. Cao *et al.* [5] utilize the cascading effect on the social network for popularity prediction by proposing a novel structure CoupleGNN. The difference is that the influence in our work is only calculated based on the similarities of the images' contents instead of the social relationship.

Though effective in representing the human perception of the digital image, image aesthetic and quality assessment has only a moderate correlation with the popularity assessment. This relation between popularity assessment and quality assessment is also

shown in [54]'s correlation between the image's quality and aesthetics attributes against the image's "willingness to share" and "content preference" attributes.

2.3 Image Retrieval

Image retrieval is the task of finding similar images in a database from a given query image or text. Due to its downstream application, the retrieval should consider not only visual similarities but also the semantic understanding of the query and efficiency. There are various datasets [25, 46] and settings such as sketch-based retrieval [33, 50], multi-modality based retrieval [53, 59], hash-based retrieval [9] and also multi-label based retrieval [32, 49]. Another branch of retrieval tasks that are more relevant to our retrieval method is image-to-image retrieval methods that target to retrieve similar image content such as landmark retrieval [57], semantics-based image retrieval dataset [17] and identical product recognition [52, 58].

However, in the popularity assessment setting, where query text or labels are absent, our aim is to retrieve contextually similar images in order to extract user-specific image features. Therefore our module not only needs to consider background semantics but also the poses of the person in the image and the context of the scene rather than focusing on retrieving images with identical people or backgrounds. Additionally, since there is no exact target for the retrieval, our retrieval module does not involve any parameter update during training and only makes use of pretrained models to query image-to-image from the training dataset. Details on the retrieval process are described in Section 3.2.

3 METHOD

3.1 Overview

Given an input image, we aim to predict its popularity score. A higher score indicates the image to be more popular. Unlike the other existing methods [12, 19, 21, 24, 26] which directly predicts likes or views in the log scale, we have trained our model f to predict a score where the ratio between two images represents the ratio between the likes.

Let (x_a, x_b) denote the input image pair and (l_a, l_b) be the corresponding number of likes of each image. Our task is to train a model f such that $f(x_a), f(x_b)$ predict the popularity scores of image a and image b which is independent of the time of post but dependent on the user uploading the images. In order to extract relevant features and incorporate them into the model in a scalable manner, our pipeline is divided into two main parts: user-specific image feature extractor and intra-image feature aggregation block.

The feature extractor consists of an image backbone model, Swin-Transformer [34], and two auxiliary branches extracting for user-specific features and 3D-aware features. As for the image backbone model, we adopt Swin-B and the feature vector before the final average pooling is chosen. The two motivations behind these auxiliary branches are as follows.

Firstly, given an image with particular content, we assume that the social feedback is correlated to a previously-uploaded similar image. For example, let's assume that there's a user that rarely uploads a photo of their children and receives a higher number of likes relative to the nearby posts (within one month) whenever the user uploads his or her children. Then one can expect that uploading a photo of their children in the future will also result in a relatively higher number of likes in the future as well. Therefore, we introduce a training dataset-based image retrieval and user-specific image feature extractor that captures the relation between the subject image to the previous photos uploaded by the user and the social preference of the user's followers.

Secondly, people are likely to focus more on different aspects of the foreground objects and on the background or far objects. In other words, the image features in different depth levels will provide the model with additional information. In order to extract features at different depth levels, we introduce a 3D-aware feature extractor based on monotonic depth estimation, which can make up for the shortage of a single model lacking 3D awareness from 2D images.

Subsequently, as the user-specific image feature extractor generates multiple features containing different aspects of the photo, the aggregation block is designed to learn the intra-image feature interaction. This feature aggregation module is composed of the transformer encoder and self-attention block [42] for effective and efficient feature aggregation. At last, we make use of the final embedding to predict the popularity score. An overview of our model pipeline is shown in Figure 2.

3.2 User-Specific Image Feature Extractor

Due to the diversity in the audience and the user upload pattern, how social media perceives the image largely differs based on who uploads the image. Previous image popularity assessment methods have either used image features or introduced meta-data such as the number of followers, caption, and upload time as additional features to provide more information about the relation between uploaded image and user. However, due to its limited accessibility, meta-data features are undesirable as input. Therefore, for incorporating user-specific features, our model only utilizes images, the timestamp, and the popularity score (i.e., the corresponding number of likes) in the training dataset. One should note that the timestamp is only utilized to find the nearby post and is not used as input for the model.

To introduce user-specificness, we focus on the past uploaded content and corresponding social feedback already available in the training data. More specifically, given an image, we aim to leverage the social preference of previously uploaded similar images. The social preference of an image is measured by the local score, which equals the image's popularity against the same user's nearby average image popularity. As the retrieval result is dependent on both the subject image and also the user, we call this module a user-specific image feature extractor, as shown in Figure 3.

3.2.1 Retrieval Strategy. The core of the retrieval strategy is to find a proper metric to calculate the similarity between two image posts. To capture the semantic context and relation between objects in the image, each image is analyzed in three aspects: pose, background, and caption. For each similarity calculation, we first extract the background and caption features and detect human poses, if any, presented in the image. Then we compute three similarity scores using ε_p, ε_b, and ε_c between two image posts. We set the pose similarity function(ε_p) as Object Keypoint Similarity (OKS) and

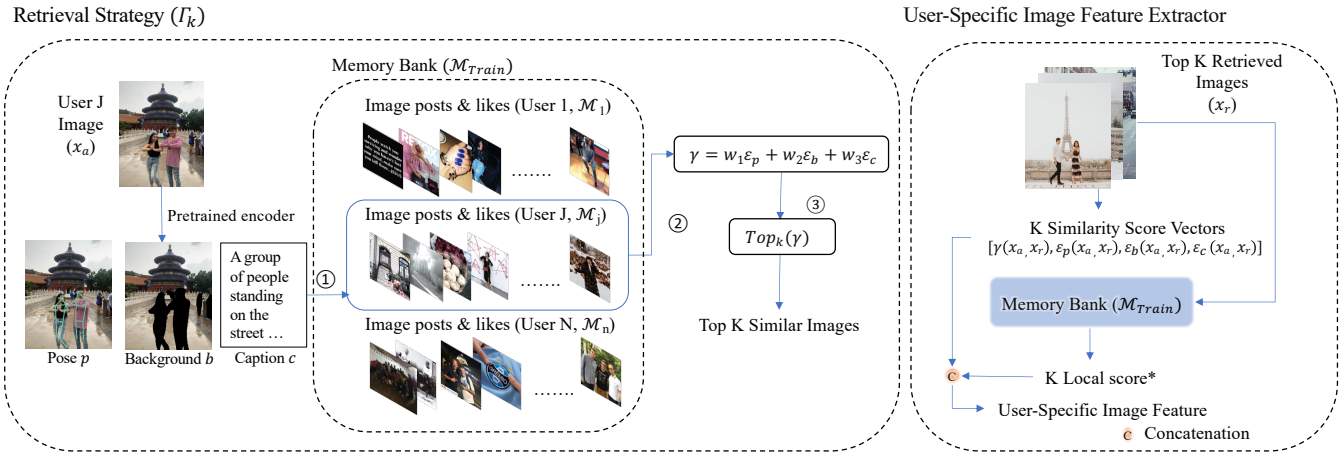

Figure 3: Illustration of retrieval strategy and user-specific image feature extractor. Image retrieval and features are processed based on the user's previous post and corresponding likes. γ stands for the similarity function between the query image and the retrieved image. Given the retrieved image, local popularity score* is calculated by Equation 3 based on images within a time window to avoid the change in the number of followers.

background similarity(ε_b) and caption similarity(ε_c) function both as cosine similarity. Therefore, give a pair of images (x_a, x_b). The final similarity score is:

$$\gamma(x_a, x_b) = w_1\varepsilon_p(x_a, x_b) + w_2\varepsilon_b(x_a, x_b) + w_3\varepsilon_c(x_a, x_b) \quad (1)$$

where w_1, w_2, and w_3 stand for the similarity score weights that sum up to one.

With the similarity score between the subject image (x_a) and user u's previous posts, we retrieve Top K images from user u's memory bank ($\mathcal{M}_u$), which refers to the training set of the same user containing information about images and the corresponding number of likes, and the uploaded timestamp. Two examples of the retrieval can be found in Figure 4. The range of the total similarity score is $[0, 1)$, and for cases where no people are detected in the query image, we set the pose similarity function as zero since there is no similarity to be calculated. We denote these retrieved images by subject image x_a as denoted as $\{x_r\}^a_{1\le r\le K}$.

3.2.2 User-Specific Image Feature Extractor. After obtaining the Top K retrieved images, we use the user-specific image feature extractor to get the local popularity scores. Local popularity score can be thought of as a performance metric where it compares the popularity score (i.e., number of likes) to the average popularity score of nearby posts. Due to the change of followers at different timestamps and in order to represent how well the retrieved image was perceived at the time of posting, we cannot directly use the popularity score of the retrieved images. As a result, we calculate the local score, which is the ratio of the retrieved image's popularity score to the average popularity score of the nearby images.

To calculate the local popularity score, the first step is to find the nearby images given the retrieved image x_r posted by user u. Nearby images C_r is the set of images posted by him or herself within the time window δ centered on the post timestamp of image x_r. In other words, C_r is defined as :

$$C_r = \{x_m : x_m \in \mathcal{M}_u, |t_m - t_r| < \delta, x_m \neq x_r\}, \quad (2)$$

where $\mathcal{M}_u$ denotes the user u's memory bank. δ denotes the value of the time window. t_r and t_m denote the uploaded timestamps of images x_r and x_m separately. During our experiment, we set δ as one month.

The second step is to calculate the local popularity score (p_r) of retrieved image x_r:

$$p_r = \frac{l_r}{\frac{1}{|C_r|}\sum_{x_m \in C_r} l_m}, \quad (3)$$

where l_r and l_m denote the popularity scores of images x_r and x_m separately. The range of local score is $(0, +\infty)$. Based on the concepts above to extract the historical statistics of user u uploading image x_a, the user-specific image feature of image x_a is the concatenation of the local score and the similarity score values of all pairs between the subject image and retrieved images $x^r_{pair} = (x_a, x_r)$, where $x_r \in \mathcal{M}_u$ for all r. As shown in Figure 3, $f_{Stat}(u, x_a)$, of which the dimension is $\mathbb{R}^{5K}$, is calculated by:

$$f_{Stat} = \{[p_r, \gamma(x^r_{pair}), \varepsilon_b(x^r_{pair}), \varepsilon_p(x^r_{pair}), \varepsilon_c(x^r_{pair})]\}_{1\le r\le K}. \quad (4)$$

3.3 3D-aware Feature Extractor

Along with the user-specific image feature, we introduce the 3D-aware feature extractor (f_{3D}) based on the depth map in order to make up for the shortage of 2D images lacking information on the third dimension. Assume x_a is the input, then the monocular depth map will be $D_a = f_{depth}(x_a)$, where $x_a \in \mathbb{R}^{H\times W\times 3}$, $D_a \in \mathbb{R}^{H\times W\times 1}$. Then the depth map is normalized between values 0 and 1. With the normalized depth map D_a, we generate a set of 3D-aware masks $\{M_a\}$ as equation below:

$$m_{ij} = \mathbb{1}(\xi_1 < d_{ij} < \xi_2), \quad (5)$$

where d_{ij} denote the i_{th} row and j_{th} column pixel depth in D_a and ξ_n denote threshold depths. With interval values $\{\xi_1 = 0, \xi_2 = \frac{1}{3}\}$, $\{\xi_1 = \frac{1}{3}, \xi_2 = \frac{2}{3}\}$, $\{\xi_1 = \frac{1}{3}, \xi_2 = 1\}$, we can generate three masks that can be regarded as close, middle, and far region in the given image. Furthermore, we also introduce two other masks $\{\xi_1 = 0, \xi_2 = 0.5\}$

Figure 4: Top 5 retrieved images given a query image. Similarity scores between the query and retrieved images are stated in the bottom left corner of each image. The weighted sum of similarity calculates the similarity in three aspects: pose, background, and caption.

and $\{\xi_1 = 0.5, \xi_2 = 1\}$ to strengthen the connection between each part, which can be regarded as foreground and background masks.

Given the depth-based masked image, we adopt ResNet-50 [18] to extract features from multiple images by:

$$f_{3D}(x_a) = f_{ResNet}(x_a; f_{depth}) = f_{ResNet}(M_a \odot x_a). \quad (6)$$

3.4 Feature Aggregation Block

The feature aggregation block is a module intended to fuse multiple features from the same image. As shown in Figure 2, this part is composed of a stack of transformer encoder blocks. The transformer encoder block consists of a multi-head self-attention (MSA) block, a fully connected feed-forward network (FFN) block, residual structure, and layer normalization (LN) similar to [13, 42]. Also, inspired by CLS token in ViT [13], we introduce a learnable embedding $A \in \mathbb{R}^C$, apart from the linearly projected input features extracted from the three branches.

3.5 Loss

Given a pair of input images, image A and image B, let us denote $\{l_A, l_B\}$ to be the number of likes of the two input images, and $\{s_A, s_B\}$ to be the predicted scores of the two input images.

The loss for training the model is:

$$\mathcal{L} = \|2(\frac{s_A}{s_A + s_B} - \frac{l_A}{l_A + l_B})\|^2 + \|2(\frac{s_B}{s_A + s_B} - \frac{l_B}{l_A + l_B})\|^2. \quad (7)$$

4 EXPERIMENTS

4.1 Implementation Details

4.1.1 Dataset. Instagram Influencer Dataset [29] is used to conduct all experiments. This dataset includes the top 3000 influencers' posts, where the number of posts averages around 523. The reason for choosing this dataset over other Image Popularity Assessment datasets [11, 44] is as follows:

First, [11]'s image pair does not have any information about how much difference there is in terms of popularity but only the order. A large number of image sources were inaccessible. For [44], as stated in [22], 69% of users in the testing set cannot be found on the training set, and the dataset contains 486 thousand posts uploaded by 70 thousand users. This indicates that there are less than seven posts per user on average. This makes it unfeasible for us to construct and test our user-specific image feature model. Therefore in order to construct a similar setting as [11] where the number of posts per user is large and also for access to the number of likes, we have chosen Instagram Influencer Dataset [29].

There are two types of information in the dataset: the image files and the metadata, consisting of timestamps, hashtags, captions, and more. We only use images, the number of likes, user ID, and timestamps. Out of all the influencers in the dataset, we only utilize the top 3000 users with the highest number of followers. For ease of discussion, we name this dataset Top3000.

4.1.2 Experimental Settings. We split the Top3000 dataset into several groups and compared only the images from the same group with each other. For one group with n number of images, the total image pairs will be $C(n, 2) = \frac{n(n-1)}{2}$. Following are the settings we have used for generating the groups.

- These images within one group must belong to the same celebrity. And each group's size is kept between 2 to 30.
- Post timestamp of any pairs is within one week to avoid time-changing factors such as the number of followers.
- Similar to image pair generation criteria (PDIP) proposed by [11], we ensure that the one image is intrinsically more popular than the other image through a constraint

$$\mu = \frac{|l_a - l_b|}{max(l_a, l_b)} \geq 0.5, \quad (8)$$

where l_a, l_b stands for the likes for image a and image b.

We randomly assign the group with the ratio of 8:1:1 to the training, validation, and test sets. The training set has 569,151 images, 115,496 groups, and 521,195 pairs. The validation set has 71,071

Method	Features	Metrics			
		Pairwise Accuracy↑	PLCC↑	SRCC↑	Ratio MSE↓
Human	Image	58.66%	-	-	-
NIMA [40]	Image	51.60%	0.0285	0.0259	2.0720
MUSIQ [27]	Image	52.84%	0.0614	0.0540	2.0830
Ding *et al.* [11]	Image	72.73%	0.3157	0.3019	-
Zhang *et al.* [60]	Image, User ID, Caption	64.86%	0.2149	0.2024	1.8892
Xu *et al.* [51] (Swin-B)	Image, Description, Hashtag Comment, Caption	77.51%	0.4047	0.3804	1.8130
Ours	Image	**79.48%**	**0.4052**	**0.3871**	**1.1388**

Table 1: Quantitive evaluation of our model and baselines on the Instagram Top3000 dataset.

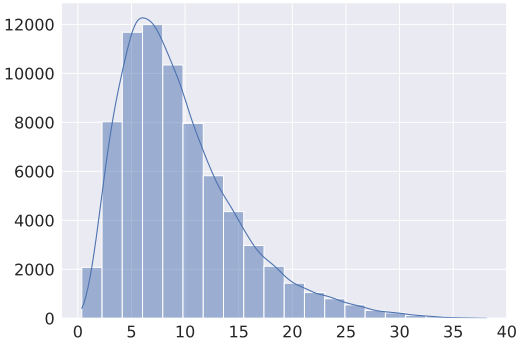

Figure 5: Histogram of the popularity scores for the test dataset.

images, 14,450 groups, and 62,820 pairs. The test set has 71,948 images, 14,504 groups, and 66,536 pairs.

In order to evaluate the popularity score, we report metrics including Pairwise Accuracy, Pearson linear correlation coefficient (PLCC), Spearman's rand-order correlation coefficient (SRCC), and Ratio MSE. For PLCC and SRCC, we report the mean of each group's metric score. And for Pairwise Accuracy and Ratio MSE, we compare popularity scores between image pairs constructed. Give a pair of images (x_a, x_b) and corresponding like and predicted score as (s_a, s_b), (l_a, l_b), the Ratio MSE is defined as:

$$Ratio\ MSE = || \log\frac{s_a}{s_b} - \log\frac{l_a}{l_b} ||^2. \quad (9)$$

4.1.3 Training Setting. We adopt Swin-B [34], which is pre-trained on ImageNet-22K as our image feature extractor. As for the 3D-aware feature's monocular depth estimator and user-specific image feature extractor's pose estimator, caption feature extractor and background segmentation model, we use the pre-trained AdaBins [2], M2 Transformer [8], DEKR [16], and DeepLab V3 [61] respectively. The parameters are frozen during the training stage. We used AdamW as an optimizer with weight decay of 10^{-2} while setting learning rates as 10^{-6} for the pre-trained Swin-B and 10^{-5} for the rest of the trainable parts. The training stage takes around 30 hours on 4 NVIDIA 2080Ti GPUs with a batch size of 12. The latency for inference is 62ms for one batch.

4.2 Quantitative Results

4.2.1 Main results. The main results are shown in Table 1. We have chosen these models to compare our model's performance against

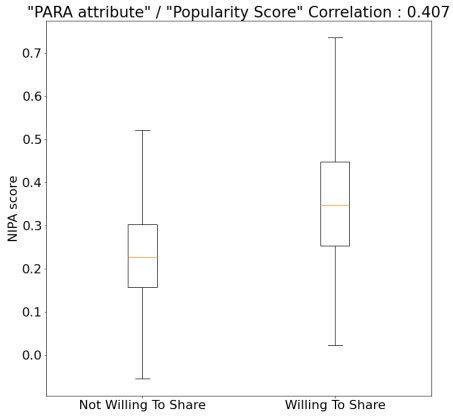

Figure 6: Correlation between willingness to share an image from PARA [54] dataset and our model's popularity score.

the image aesthetic assessment model NIMA [40], image quality assessment model MUSIQ [27] and other image popularity assessment models which use meta-features [51, 60], and I^{2}PA model Ding *et. al* [11], which does not use meta-features. For experiments of Xu *et al.* [51], we have changed the image feature extractor to Swin Transformer for a fair comparison. For NIMA [40], MUSIQ [27], and Ding [11], we use the pretrained model, which is made public by the authors. For other works [51, 60], we trained models with the meta-information incorporated in their model on our dataset by applying directly predicting the log scale's popularity score with MSE Loss, respectively.

We could see from Table 1 that our model has significantly increased the pairwise accuracy by 8.93% compared with the Ding *et al.* [11]. We also show the histogram of our prediction popularity scores in Figure 5 for our test dataset. In addition to the results above, we conducted a user study and asked 22 Instagram users to predict which one of the two images would get more the number of likes when given 30 image pairs. We remove the highest and lowest scores from the candidates and report the mean of the error rate in Table 1.

4.2.2 Empirical Analysis. In order to validate the effectiveness of our trained model, we have checked the correlation between the "willingness to share" attribute of Personalized Aesthetics with Rich Attribute (PARA) dataset [54] and our model's inference on the training dataset. The "willing to share score" represents the

Ablation	Method	Pairwise Accuracy↑
Fused Features	Basic	78.64%
	Basic, Retrieval	79.12%
	Basic, Retrieval, 3D	**79.48%**
Fusion Method	Concatenation Block	68.97%
	Aggregation Block	**79.48%**
Training Loss	Log Number of Likes	68.73%
	Pairwise Likes Ratio	**79.48%**

Table 2: Ablation study on architectures and losses.

responses of 438 subjects, where subjects rated an integer score between 1 to 5 on the question "The willingness to share this photo to social media." Each image was rated by 26 subjects on average. Hence for a valid evaluation of whether the image is "willing to be shared or to not be shared," we narrowed images down to 5079 images by filtering images when less than 33% of the subjects agreed on a common rate and when the average rate is between bottom 10th percentile and upper 10th percentile.

Based on the filtered images, we define "willing to share" images when the average rate is higher than the upper 10th percentile of the average rates of all images and similarly define "unwilling to share" using the bottom 10th percentile value. For user-specific image feature extraction, we sampled 500 random images of the training dataset as a memory bank for the subject image. Based on the 5079 image samples, we achieved a correlation higher than 0.4 without any tuning to the PARA dataset [54].

4.3 Ablation Study

In Table 2, we aim to show the effect of three components involved in training our final model. Firstly, we show that combining auxiliary branch features with the Swin Transformer backbone feature each leads to improvement. Secondly, we show the effect of the feature aggregation block, which consists of a self-attention module and a learnable vector, by comparing the result against a simple concatenation of multiple features. We verified the aggregation block leads to a 10.51% improvement. Lastly, as stated in Equation 7, we have made use of the pairwise number of like ratios as the training target of our model. Since predicting the likes of each image in the log scale is a common way of training a popularity assessment network, we compare the training results of the log scale number of likes target against our training target. We verified the pairwise-like ratio results in 10.75% higher pairwise accuracy.

4.4 Applications

4.4.1 Video thumbnail recommendation. One possible application of our model would be to generate a thumbnail of a user-uploaded video by extracting every 25 frames of a video uploaded by a user and evaluating their popularity score. Figure 7 shows an example of our video thumbnail selection results. More details are shown in the supplementary material.

4.4.2 Popularity assessment on Generative Model. With realistic image generative models like denoising diffusion models, one would still have to undergo the process of choosing an image from a large number of generated images. Hence we use stable diffusion [39]

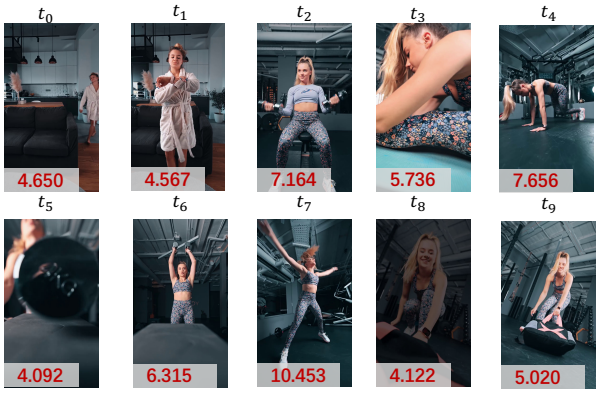

Figure 7: Choosing the best frame from a video as the cover. The predicted popularity score is stated in the bottom left corner of each image.

Figure 8: Selecting the best photo from a group of images generated by Stable Diffusion [39] with prompt "Kendall Jenner on Instagram." The predicted popularity score is stated at the bottom of each image.

with the prompt "Kendall Jenner on Instagram" to generate the images and predicted popularity score using Kendall Jenner's previous Instagram posts. As demonstrated in Figure 8, for content creators, our model can serve as popularity criteria or reference points specific to a particular user.

5 CONCLUSION

This paper proposes a retrieval-augmented approach based on a deep neural network that fuses image features and user-specific image statistics to address the Intrinsic Image Popularity Assessment. However, there are still several limitations and future work. Firstly, various data types and formats, such as images, sidecars (multiple images), and videos, are being uploaded to social media. Assessing the popularity of these other data formats will be a direct extension of the work. Secondly, in the proposed method, we retrieve images from the historical posts of the same user. For those users with few posts on social media, the user-specific image features may not be as meaningful as it is for top accounts with a large number of posts. From that regard, a more general retrieval strategy across users and platforms will be an improvement of the proposed method for broader applications in various areas.

ACKNOWLEDGMENTS

This project was supported by the National Key R & D Program of China under grant number 2022ZD0161501.

REFERENCES

[1] Fatma S Abousaleh, Wen-Huang Cheng, Neng-Hao Yu, and Yu Tsao. 2020. Multimodal deep learning framework for image popularity prediction on social media. *IEEE Transactions on Cognitive and Developmental Systems* 13, 3 (2020), 679–692.

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4009–4018.

[3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Semi-Parametric Neural Image Synthesis. In *Advances in Neural Information Processing Systems*.

[4] Ethem F Can, Hüseyin Oktay, and R Manmatha. 2013. Predicting retweet count using visual cues. In *Proceedings of the 22nd ACM international conference on information & knowledge management*. 1481–1484.

[5] Qi Cao, Huawei Shen, Jinhua Gao, Bingzheng Wei, and Xueqi Cheng. 2020. Popularity prediction on social platforms with coupled graph neural networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 70–78.

[6] Weilong Chen, Chenghao Huang, Weimin Yuan, Xiaolu Chen, Wenhao Hu, Xinran Zhang, and Yanru Zhang. 2022. Title-and-Tag Contrastive Vision-and-Language Transformer for Social Media Popularity Prediction. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 7008–7012. https://doi.org/10.1145/3503161.3551568

[7] Yingying Cheng, Fan Zhang, Gang Hu, Yiwen Wang, Hanhui Yang, Gong Zhang, and Zhuo Cheng. 2021. Block Popularity Prediction for Multimedia Storage Systems Using Spatial-Temporal-Sequential Neural Networks. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3390–3398. https://doi.org/10.1145/3474085.3475495

[8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[9] Hui Cui, Lei Zhu, Jingjing Li, Zhiyong Cheng, and Zheng Zhang. 2021. Two-pronged Strategy: Lightweight Augmented Graph Network Hashing for Scalable Image Retrieval. *CoRR* abs/2108.03914 (2021). arXiv:2108.03914 https://arxiv.org/abs/2108.03914

[10] Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. 2021. Locally Adaptive Structure and Texture Similarity for Image Quality Assessment. *CoRR* abs/2110.08521 (2021). arXiv:2110.08521 https://arxiv.org/abs/2110.08521

[11] Keyan Ding, Kede Ma, and Shiqi Wang. 2019. Intrinsic image popularity assessment. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1979–1987.

[12] Keyan Ding, Ronggang Wang, and Shiqi Wang. 2019. Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2682–2686. https://doi.org/10.1145/3343031.3356062

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[14] Yixuan Gao, Xiongkuo Min, Yucheng Zhu, Jing Li, Xiao-Ping Zhang, and Guangtao Zhai. 2022. Image Quality Assessment: From Mean Opinion Score to Opinion Score Distribution. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 997–1005. https://doi.org/10.1145/3503161.3547872

[15] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image Popularity Prediction in Social Media Using Sentiment and Context Features. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan (Eds.). ACM, 907–910. https://doi.org/10.1145/2733373.2806361

[16] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. 2021. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14676–14686.

[17] Yinzheng Gu, Chuanpeng Li, and Yu-Gang Jiang. 2019. Towards Optimal CNN Descriptors for Large-Scale Image Retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1768–1776. https://doi.org/10.1145/3343031.3351081

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[19] Ziliang He, Zijian He, Jiahong Wu, and Zhenguo Yang. 2019. Feature Construction for Posts and Users Combined with LightGBM for Social Media Popularity Prediction. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2672–2676. https://doi.org/10.1145/3343031.3356054

[20] Vlad Hosu, Hanhe Lin, Tamás Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Trans. Image Process.* 29 (2020), 4041–4056. https://doi.org/10.1109/TIP.2020.2967829

[21] Chih-Chung Hsu, Li-Wei Kang, Chia-Yen Lee, Jun-Yi Lee, Zhong-Xuan Zhang, and Shao-Min Wu. 2019. Popularity Prediction of Social Media based on Multi-Modal Feature Mining. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2687–2691. https://doi.org/10.1145/3343031.3356064

[22] Chih-Chung Hsu, Pi-Ju Tsai, Ting-Chun Yeh, and Xiu-Yu Hou. 2022. A Comprehensive Study of Spatiotemporal Feature Learning for Social Medial Popularity Prediction. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 7130–7134. https://doi.org/10.1145/3503161.3551593

[23] Chih-Chung Hsu, Li-Wei Kang, Chia-Yen Lee, Jun-Yi Lee, Zhong-Xuan Zhang, and Shao-Min Wu. 2019. Popularity prediction of social media based on multi-modal feature mining. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2687–2691.

[24] Feitao Huang, Junhong Chen, Zehang Lin, Peipei Kang, and Zhenguo Yang. 2018. Random Forest Exploiting Post-related and User-related Features for Social Media Popularity Prediction. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 2013–2017. https://doi.org/10.1145/3240508.3266439

[25] Bogdan Ionescu, Alexandru-Lucian Gînsca, Bogdan Boteanu, Mihai Lupu, Adrian Popescu, and Henning Müller. 2016. Div150Multi: a social image retrieval result diversification dataset with multi-topic queries. In *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016, Klagenfurt, Austria, May 10-13, 2016*, Christian Timmerer (Ed.). ACM, 46:1–46:6. https://doi.org/10.1145/2910017.2910620

[26] Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. 2019. Catboost-based Framework with Additional User Information for Social Media Popularity Prediction. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2677–2681. https://doi.org/10.1145/3343031.3356060

[27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*.

[28] Jongyoo Kim and Sanghoon Lee. 2017. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1676–1684.

[29] Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. 2020. Multimodal Post Attentive Profiling for Influencer Marketing. In *Proceedings of The Web Conference 2020*. 2878–2884.

[30] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4565–4569. https://doi.org/10.1145/3394171.3416273

[31] Yaohui Li, Yuzhe Yang, Huaxiong Li, Haoxing Chen, Liwu Xu, Leida Li, Yaqian Li, and Yandong Guo. 2022. Transductive Aesthetic Preference Propagation for Personalized Image Aesthetics Assessment. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 896–904. https://doi.org/10.1145/3503161.3548244

[32] Ying Li, Hongwei Zhou, Yeyu Yin, and Jiaquan Gao. 2021. Multi-label Pattern Image Retrieval via Attention Mechanism Driven Graph Convolutional Network. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 300–308. https://doi.org/10.1145/3474085.3475695

[33] Zhixin Ling, Zhen Xing, Jiangtong Li, and Li Niu. 2022. Multi-Level Region Matching for Fine-Grained Sketch-Based Image Retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 462–470. https://doi.org/10.1145/3503161.3548147

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[35] Hao Lou, Heng Huang, Chaoen Xiao, and Xin Jin. 2021. Aesthetic Evaluation and Guidance for Mobile Photography. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 2780–2782. https://doi.org/10.1145/3474085.3478557

[36] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A multimodal approach to predict social media popularity. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 190–195.

[37] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. IEEE Computer Society, 2408–2415. https://doi.org/10.1109/CVPR.2012.6247954

[38] Christoffer Riis, Damian Konrad Kowalczyk, and Lars Kai Hansen. 2020. On the limits to multi-modal popularity prediction on instagram–a new robust, efficient and explainable baseline. *arXiv preprint arXiv:2004.12482* (2020).

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[40] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing* 27, 8 (2018), 3998–4011.

[41] Yunpeng Tan, Fangyu Liu, Bowei Li, Zheng Zhang, and Bo Zhang. 2022. An Efficient Multi-View Multimodal Data Processing Framework for Social Media Popularity Prediction. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 7200–7204. https://doi.org/10.1145/3503161.3551607

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[43] Kai Wang, Penghui Wang, Xin Chen, Qiushi Huang, Zhendong Mao, and Yongdong Zhang. 2020. A Feature Generalization Framework for Social Media Popularity Prediction. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4570–4574. https://doi.org/10.1145/3394171.3416294

[44] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. SMP Challenge: An Overview of Social Media Prediction Challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2667–2671. https://doi.org/10.1145/3343031.3356084

[45] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 272–278. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11887

[46] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 11307–11317. https://doi.org/10.1109/CVPR46437.2021.01115

[47] Jianmin Wu, Liming Zhao, Dangwei Li, Chen-Wei Xie, Siyang Sun, and Yun Zheng. 2022. Deeply Exploit Visual and Language Information for Social Media Popularity Prediction. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 7045–7049. https://doi.org/10.1145/3503161.3551576

[48] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. *arXiv preprint arXiv:2203.08913* (2022).

[49] Chengyin Xu, Zenghao Chai, Zhengzhuo Xu, Chun Yuan, Yanbo Fan, and Jue Wang. 2022. Hyp^2 Loss: Beyond Hypersphere Metric Space for Multi-label Image Retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 3173–3184. https://doi.org/10.1145/3503161.3548032

[50] Jiaqing Xu, Haifeng Sun, Qi Qi, Jingyu Wang, Ce Ge, Lejian Zhang, and Jianxin Liao. 2021. DLA-Net for FG-SBIR: Dynamic Local Aligned Network for Fine-Grained Sketch-Based Image Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 5609–5618. https://doi.org/10.1145/3474085.3475705

[51] Kele Xu, Zhimin Lin, Jianqiao Zhao, Peichang Shi, Wei Deng, and Huaimin Wang. 2020. Multimodal Deep Learning for Social Media Popularity Prediction With Attention Mechanism. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4580–4584. https://doi.org/10.1145/3394171.3416274

[52] Runming Yan, Yongchun Lin, Zhichao Deng, Liang Lei, and Chudong Xu. 2020. Multi-Feature Fusion Method Based on Salient Object Detection for Beauty Product Retrieval. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4713–4717. https://doi.org/10.1145/3394171.3416272

[53] Yuchen Yang, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Cross-modal Joint Prediction and Alignment for Composed Query Image Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3303–3311. https://doi.org/10.1145/3474085.3475483

[54] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. 2022. Personalized Image Aesthetics Assessment with Rich Attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 19829–19837. https://doi.org/10.1109/CVPR52688.2022.01924

[55] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L. Rosin. 2023. Towards Artistic Image Aesthetics Assessment: a Large-scale Dataset and a New Method. *CoRR* abs/2303.15166 (2023). https://doi.org/10.48550/arXiv.2303.15166 arXiv:2303.15166

[56] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. 2020. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 3572–3582. https://doi.org/10.1109/CVPR42600.2020.00363

[57] Sangwoong Yoon, Woo-Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. 2021. Image-to-Image Retrieval by Learning Similarity between Scene Graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 10718–10726. https://ojs.aaai.org/index.php/AAAI/article/view/17281

[58] Jun Yu, Guochen Xie, Mengyan Li, Haonian Xie, Xinlong Hao, Fang Gao, and Feng Shuang. 2020. Attention Based Beauty Product Retrieval Using Global and Local Descriptors. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4708–4712. https://doi.org/10.1145/3394171.3416289

[59] Feifei Zhang, Ming Yan, Ji Zhang, and Changsheng Xu. 2022. Comprehensive Relationship Reasoning for Composed Query Based Image Retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 4655–4664. https://doi.org/10.1145/3503161.3548126

[60] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 world wide web conference*. 1277–1286.

[61] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *CVPR*.

[62] Zihan Zhou, Yong Xu, Ruotao Xu, and Yuhui Quan. 2022. No-Reference Image Quality Assessment Using Dynamic Complex-Valued Neural Model. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 1006–1015. https://doi.org/10.1145/3503161.3547982

[63] Yunan Zhu, Haichuan Ma, Jialun Peng, Dong Liu, and Zhiwei Xiong. 2021. Recycling Discriminator: Towards Opinion-Unaware Image Quality Assessment Using Wasserstein GAN. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 116–125. https://doi.org/10.1145/3474085.3479234