A Dataset Analysis

A.1 Subreddits Analysis

There are 1.5M members in the `/r/photocritique` subreddit. Since it is not possible to collect demographic information about subreddit members, we report the statistic related to a recent analysis about Reddit [10]. Slight majority of Reddit users are male (61%). 48% of Reddit users are in the US, followed by the UK, Canada, Australia and Germany. People between the age of 18 and 29 make up Reddit's largest user base (64%). The second biggest age group is 30 to 49 (29%). Teenagers below 15 are not very active on Reddit. Only 7% of Reddit users are over 50.

In light of the previous statistics, it is necessary to underline that the data treated in our dataset, therefore the inferred concept of aesthetics, presents a bias due to the limited cultural and geographical integration of the people who produced the information.

Here, it follows a deeper analysis of `/r/photocritique` subreddit. Figure 5 shows the number of posts and comments per year downloaded from the seven subreddits we have selected. We observe that the number of posts and comments increase over the course of time. Data from 2013 could not be retrieved due to problems with Pushshift [11]. Since 2015 there have been a number of posts over 20K and a number of comments that exceeds 100K until reaching the peak of 250K in 2021. Furthermore, although the posts are substantially fewer than the comments, the posts have reached a constant level of over 50,000 per year.

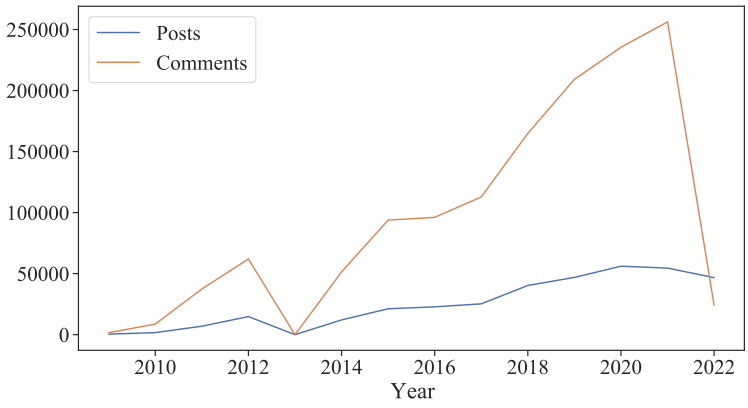

Figure 5: The number of posts and comments between May 2009 and February 2022 for the 6 considered subreddits.

A.2 Image Resolution

We investigate the resolution of the images of three datasets, namely AVA, PCCD, and our RPCD. We categorize images into 4 common image resolutions in still camera photography, namely Standard Definition – SD (720×576 pixels), High Definition – HD (1280×720 pixels), FullHD (1920×1080 pixels), and UltraHD (3840×2160 pixels). In Figure 6 the distributions of the images for the three datasets are plotted with respect to the four considered resolutions. As it is possible to see, our dataset it the only one that has UltraHD images. Most of the images are UltraHD resolution (51.20%), but there are also images for the other three resolutions. On the other hand, all AVA images have a resolution of 720×576 pixels, while most PCCD images (i.e. 89.07%) have FullHD resolution.

A.3 Sentiment Polarity Classification

We delve into the analysis of the sentiment score distributions of our dataset and those of AVA and PCCD. Figure 7 shows the spreads of the sentiment scores for the three datasets. AVA and PCCD have very similar median and standard deviation, namely 0.77 and about 0.15. Our RPCD on the

[10] `https://www.statista.com/topics/5672/reddit/#topicHeader` (Accessed on 22/08/2022)
[11] `https://www.reddit.com/r/pushshift/comments/sb982i/very_recent_data_missing/`

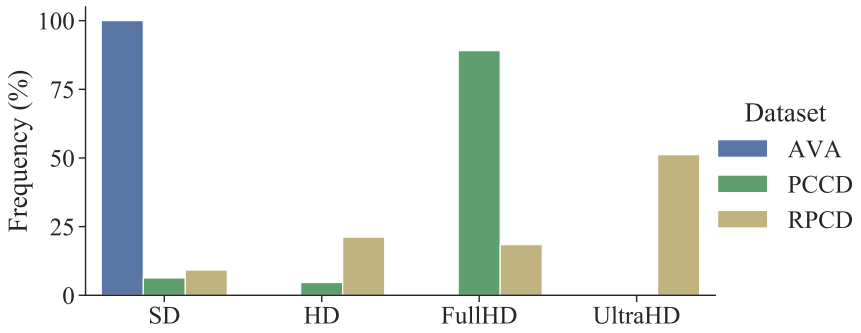

Figure 6: Image distribution of AVA, PCCD and our dataset, RPCD, for various standard image resolutions.

other hand has a median of 0.60 and a larger standard deviation (i.e., 0.25). This difference between ours and the other datasets indicates that RPCD have a richer representation of the whole aesthetic taste spectrum, providing information about why an image have a specific score for high and low sentiment scores.

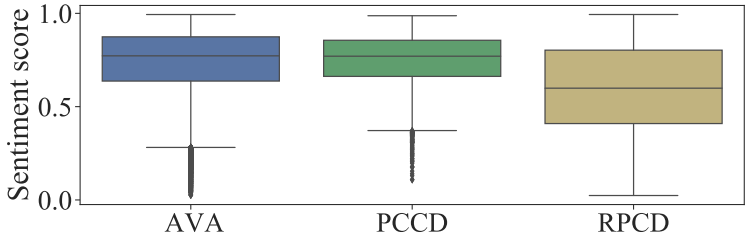

Figure 7: Boxplots of sentiment score distributions for the three considered datasets, namely AVA, PCCD and our RPCD.

Figure 8 reports some samples of the AVA dataset whose aesthetic score given by the human raters is equal to 5 (i.e., average score of the distribution), while our sentiment score span almost the entire range. It can be seen that the comments concern different aspects of photography. For example, for the central image a user has concerns about the pose of the subject "I think this would have been much more effective if the flower was facing the camera.", while another user would have preferred a different optical technique, i.e., "I would like more depth-of-field, so that the furthest petals are in focus also". Sometimes there can be very conflicting opinions in the comments (see the first image on the left). In general, comments reveal many facets of judgment shaped by the polarity of sentiment. This therefore justifies the difference between the annotated aesthetic score and the estimated sentiment score.

A.4 Content Analysis

We automatically analyze image content by using image classifiers for both semantic and composition aspects. In this section we detail the classifiers design and training and some qualitative results on our RPCD dataset.

Semantic Content and Composition Rule. To categorize the semantic content and composition of RPCD images, we use two classifiers based on the same backbone, namely the Vision Transformer (ViT) presented in [8]. In particular, we use the ViT parameters learned on ImageNet (keeping them freezed on the new tasks). The last linear classification layer is peculiar to each task and its parameters are trained. We use the same hyperparameters for the two classifiers, that is SGD with momentum equal to 0.9 and weight decay of $1e$-4. We train using batches of 32 images for 90 epochs with an initial learning rate of 0.01, that is then dropped every 30 epochs by a factor of 0.1.

The semantic content classifier is trained to discriminate six different semantic content, namely `Animal`, `Architecture`, `Human`, `Landscape`, `Plant`, and `Static`. For this purpose, we use 15,981

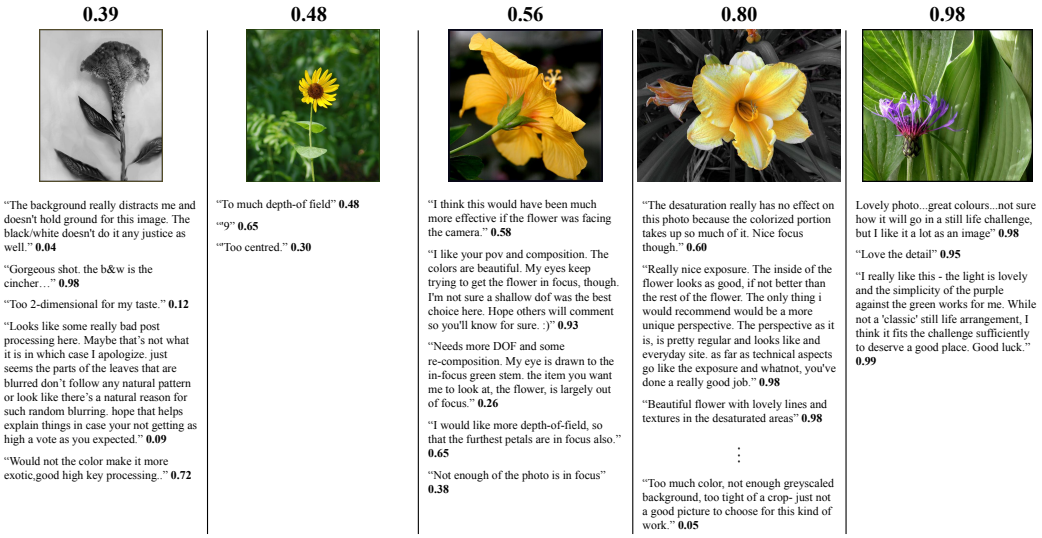

Figure 8: AVA samples annotated with an aesthetic score of 5, whose sentiment score we propose varies between 0.39 and 0.98. For each image we report the overall sentiment score (top of the image) and comments with the corresponding predicted sentiment score in bold.

images of the dataset CUHK-PQ [25] (i.e. all the images of the dataset apart from those of the Night category). We split the whole dataset int 80% training images and 20% test images. The resulting classifier achieved an accuracy of 87.08% on the test set. Figure 9 reports two images from RPCD for each semantic category.

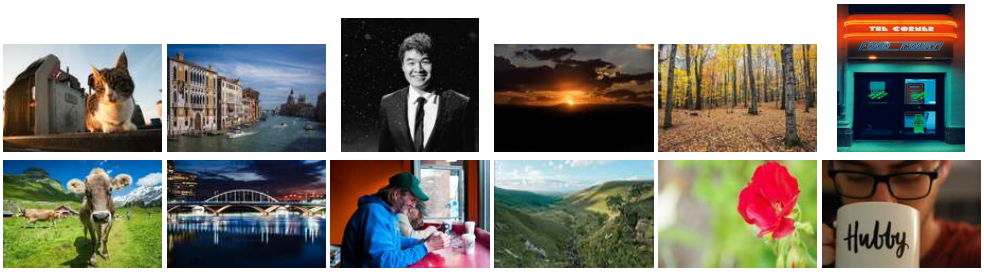

Figure 9: Images from our RCPD dataset categorized with respect to the semantic content.

Our composition classifier is trained on the KU-PCP dataset [21], which consists of 4244 outdoor photographs. We exploit the data splits provided by the authors which comprise of a training set of 3169 images and 1075 validation images. Each image has been annotated by 18 human subject to categorize it into nine composition classes: `Center`, `Curved`, `Diagonal`, `Horizontal`, `Pattern`, `Rule of Thirds (RoT)`, `Symmetric`, `Triangle`, and `Vertical`. Since an image may follow multiple composition rules, each sample is given with one or more (at most 3) composition labels. Following [13], images with more than one rule are trained multiple times for each ground-truth class. This training strategy is shown more effective than multi-label loss. The estimated accuracy on the test set is equal to 33.36%. Figure 10 shows some images from the RPCD categorized for each composition rule.

Shot Scale. We implement a Subject Guidance Network (SGNet) inspired by [32] to perform shot scale classification on images. The key idea is to use a subject map to determine the portion occupied by the subject with respect to the frame. We distinguish among five shot scale types, namely extreme close-up (ECS), close-up (CS), medium (MS), full (FS) and long (LS). The model is trained on the public MovieNet dataset [15] and optimized with stochastic gradient descent using cross-entropy loss.

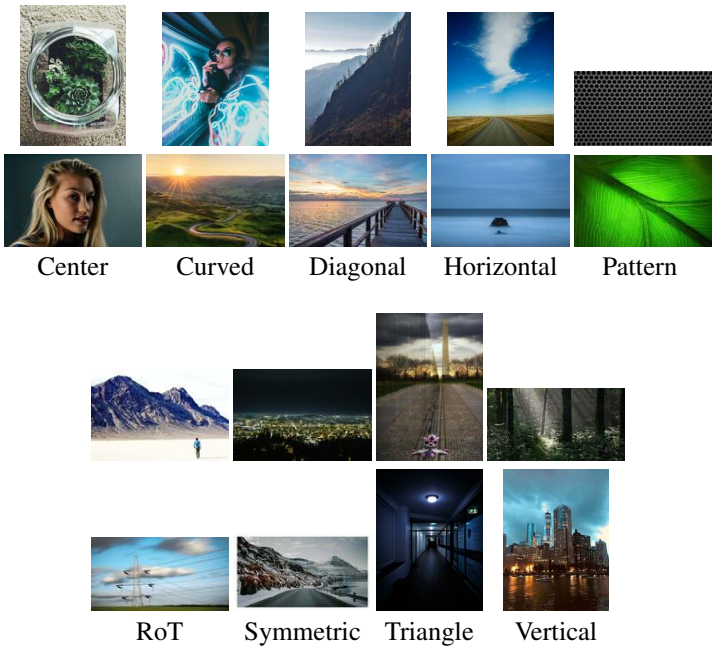

Figure 10: Images from our RCPD dataset categorized with respect to the main composition rule.

We use a learning rate of $1e$-3, batch size of 16 and we train for 60 epochs. We achieved 99.72% accuracy on the test set of the MovieNet dataset and observed a good generalization to the proposed RPCD dataset. The shot scale reveals information of how the photographer used the camera in order to emphasize either a location (long), an event (medium/full) or the identity of a subject (extreme close-up/close-up). Figure 11 reports some images annotated for each shot scale category.

Figure 11: Images from our RCPD dataset categorized with respect to the shot scale.

Aesthetic Aspect Prediction. The aesthetic aspect of each comment is predicted using a transformer model, DistilBERT, implemented using HugginFace's transformers library [39]. This approach differs with previous attempts of automatically labeling the aesthetic attributes of comments, which were based on keywords [17]. Instead, we fine-tune the language model for the text classification task of predicting the aesthetic attribute of a text using the PCCD [5] dataset, where 7 different classes are available: `general_impression`, `subject_of_photo`, `composition`, `use_of_camera`, `depth_of_field`, `color_lighting`, and `focus`. We use a learning rate of $2e$-5, batch size of 16, weight decay of 0.01 and we train for 5 epochs. The rest of the parameters are left to the default ones in the HugginFace Trainer API. We randomly split the whole dataset in two folds: 90% for training, and the remaining 10% for validation and testing. Additionally, we clean URLs and escaped characters from the dataset. The fine-tuning converges at epoch 2, where the weighted metrics over the 7 different classes are: Precision, 0.8771; Recall, 0.8751; F1-score, 0.8755; and Accuracy, 0.8751.

Figure 12 shows the correlation matrix of the classifier performance on the test set. The classifier is available on HuggingFace's model hub [12].

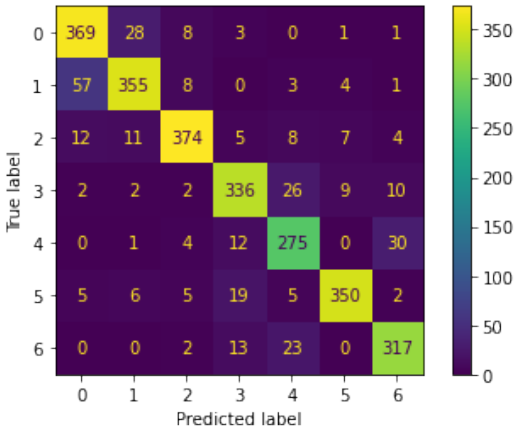

Figure 12: Correlation matrix of the aesthetic aspect classifier.

A.5 Explicit or offensive content

We use Detoxify [13], a library to predict toxic comments, to carry out a preliminary analysis of the presence of offensive content in the dataset. We use the *unbiased* model, a model that recognizes toxicity and minimizes this type of unintended bias with respect to mentions of identities (for example, minimize the bias towards the toxic class when a mention to a minority, which are often the target of offensive comments, is mentioned and the comment is not actually offensive).

In Table 4 we show the results of this preliminary analysis of the presence of offensive content in the dataset using Detoxify to predict the "offensive probability" of the 216K comments in the dataset. We have considered a comment to be offensive if the predictions probability for any of the labels is higher than 0.5. In total, there are 8K comments with a predicted probability of being offensive greater that 0.5, which represent less than the 4% of the total of comments in the dataset.

A.6 Topic Modeling

We use BERTopic [12] to clusterize the comments in all three datasets to compare the main topics being discussed. This method leverages on the document embeddings created using a text encoder to produce clusters after reducing the dimensionality of such embeddings. Then, TD-IDF is applied to the documents of the cluster to get the importance score of each word, obtaining the relevant topics in the cluster.

[12] `https://huggingface.co/daveni/aesthetic_attribute_classifier`
[13] `https://github.com/unitaryai/detoxify`

Table 4: Offensive content analysis of our RPCD using Detoxify.

Offensive label	Predicted Probability Mean	Total
toxicity	2.889%	4369
severe_toxicity	0.019%	0
obscene	1.125%	2385
identity_attack	0.374%	336
insult	0.672%	742
threat	0.418%	287
sexual_explicit	0.259%	439

To generate the topics, we used the automatic topic reduction feature available in the library to reduce the number of topics, starting from the least frequent topic, as long as it exceeds a minimum similarity of 0.915. We additionally sample 100K comments from AVA and Reddit datasets to avoid memory constraints. We describe the datasets topics in the Table 5. It shows the top 30 topics together with the count of documents belonging to each of them and the most important words per topic. Topics related to aesthetic attributes are in bold. We observe that in all of them we can find topics regarding aesthetic aspects such as composition, exposure, focus or color; but also topics related to the subject of the image such as sky, bird or flower.

Table 5: Top 30 detected Topics on AVA, PCCD, and our RPCD.

AVA		PCCD		RPCD	
Count	Name	Count	Name	Count	Name
35798	focus_and_challenge_this	12585	the_and_of_to	43491	and_the_to_is
1417	her_she_face_shes	1256	hi_you_work_image	3441	her_she_face_hair
1170	flower_flowers_petals_leaf	690	flower_flowers_petals_rose	2770	horizon_tree_trees_straighten
1150	**crop_cropped_tighter_cropping**	641	her_eyes_she_face	2423	bird_dog_cat_birds
1081	dog_cat_cats_dogs	497	**exposure_speed_shutter_water**	1971	sky_clouds_cloud_blue
904	sky_clouds_cloud_skies	469	bird_birds_feathers_the	1367	**crop_cropped_square_tighter**
899	title_titles_without_titled	451	**sharp_focus_looks_resolution**	1218	building_buildings_tower_architecture
810	ribbon_red_congrats_deserved	434	**field_depth_shallow_appropriate**	1217	his_him_he_face
786	tree_trees_branches_branch	417	**subject_interesting_matter_choice**	1064	where_taken_live_place
773	water_drops_fog_rain	372	**color_lighting_colors_sky**	888	flower_flowers_petals_focus
747	**composition_composed_shot_nicely**	365	tree_trees_branches_the	821	**water_reflection_exposure_puddle**
714	portrait_self_portraits_candid	347	child_baby_children_daughter	782	boat_boats_ship_water
687	reflection_reflections_mirror_mirrors	326	**iso_noise_speed_shutter**	728	please_titles_examples_specific
684	score_averaged_total_autool	298	**perspective_composition_angle_good**	675	car_cars_truck_front
653	comment_done_knowitall_explaining	295	**dof_diffraction_focus_good**	649	**iso_shutter_speed_noise**
626	**bw_conversion_choice_works**	257	landscape_location_beautiful_landscapes	617	**hdr_range_dynamic_exposures**
620	**sharp_sharpness_sharper_sharpened**	251	**aperture_fdepth_field**	592	**bridge_bridges_leading_lines**
617	**capture_great_wonderful_colors**	225	**auto_manual_mode_settings**	549	mountain_mountains_clouds_foreground
597	**shadow_shadows_light_harsh**	198	good_very_bad_apparently	515	street_road_photography_trails
564	finish_top_congrats	181	spot_on_looks_seems	465	url_thisurl_oneurl_heres
537	bird_birds_beak_eagle	178	**perfect_looks_about_focus**	446	photography_learn_photographer_art
533	**framing_frame_framed_filled**	162	**focus_subject_main_sharp**	440	**bw_conversion_color_version**
526	road_where_city_place	154	boat_boats_pier_horizon	411	leaf_leaves_plant_plants
484	congratulations_congrats_proud_fantastic	154	looks_good_great_very	409	**vignette_vignetting_heavy_strong**
473	**tones_tone_tonemapping_mapping**	137	building_buildings_right_perspective	407	critique_criticism_critiques_no
464	building_buildings_tower_architecture	137	**horizon_line_frame_middle**	406	rock_rocks_foreground_bottom
462	**border_borders_fan_distracting**	134	dog_dogs_fur_eyes	400	beautiful_pic_gorgeous_lovely
431	**focus_out_focused_seems**	125	butterfly_wings_butterflies_wing	390	reflection_mirror_reflections_mirrors
422	meets_challenge_meet_fits	118	animal_animals_wildlife_monkey	388	stars_star_trails_astrophotography
417	**lighting_light_brighter_composition**	112	**bw_contrast_choice_conversion**	371	portrait_portraits_landscape_self

A.7 Informativeness Analysis

We use the definition of informativeness score of a previous work [10] as a proxy of how meaningful are the comments in our dataset and how do they compare to other datasets. This definitions leverages on the relative frequency of unigrams and bigrams respect to the total vocabulary of the corpus. Then, a comment is represented as the union of its unigrams and bigrams and it is assigned an informativeness score ρ as the average of the negative log probabilities of its unigrams ($P(u_i)$) and bigrams ($P(b_j)$):

$$\rho = -\frac{1}{2}[log \prod_{i}^{N} P(u_i) + log \prod_{j}^{M} P(b_j)] \tag{3}$$

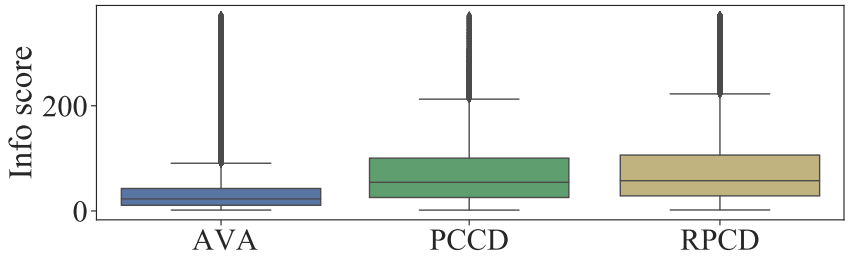

Figure 13: Informativeness score for each of the considered datasets.

The more frequent a word is, the less informative it will be. We observe that the proposed RPCD dataset have a slightly higher informativeness score than PCCD dataset, while both of them have a score twice as high as AVA dataset. The box plot shown in Figure 13 describes the informativeness score distribution among the different datasets, where the dataset with the highest average informativeness score is RPCD (78.61), followed by PCCD (72.96) and finally, with less than half the score, AVA (32.43).

B Experiments and Implementation Details

B.1 Image Aesthetic Assessment

AestheticViT. For ranking the images with respect to aesthetic or sentiment scores, we experiment with different models based on Vision Transformer (ViT) [8] as a baseline as this architecture has proved its effectiveness for several tasks. We run experiments with various versions of ViT in terms of model size (i.e., Tiny, Small, Base, and Large), input patch size, and pre-training dataset. In what follows we use brief notation to indicate the model size and the input patch size: for instance, ViT-L/16 means the "Large" variant with 16×16 input patch size. We also consider the Data-efficient image Transformer (DeiT), a post-ViT model that improves the training process and performance.

On top of the pre-trained transformer, we add a fully-connected layer which is randomly initialized. The whole model is then trained to predict the final score using the mean square error as the loss function. We resize the input images to have the maximum input size of 700 pixels and adjust the other size to preserve the original aspect ratio. To handle images with varying resolutions, we scale input positional embedding accordingly with the image resolution by performing bilinear interpolation. During training we adopt a batch size of 1 because the images can have different resolutions. We finetune the models until convergence, for a maximum of 5 epochs (although convergence usually occurs on epoch 2 or 3). The learning rate is empirically set to $1e$-6 and we use Adam optimizer. We exploit the available model implementations and pre-trained weights of the PyTorch Image Models library[14].

We first perform experiments for estimating the aesthetic score of AVA [28] and PCCD [17] (i.e., the only two datasets with both comments and aesthetic scores). Table 6 reports performance on the test sets in terms of SRCC, and LCC for aesthetic score regression and accuracy for low-/high-aesthetic categorization. The accuracy is computed defining as high quality images those with an score above 5, and poor quality otherwise. The best results are achieved by the `ViT-L/16` pre-trained on ImageNet-21k and other considerations can be made. First, the use of larger patches, that is 32×32 pixels instead of 16×16 pixels, causes a significant drop in performance for the same model size. In fact, we have that `ViT-B/32` achieves 0.446 of SRCC, while `ViT-B/16` obtains 0.759 of SRCC on AVA. Second, the performance increases as the model size grows. On AVA, the SRCC is equal to 0.725 for `ViT-T/16` and 0.793 for `ViT-L/16`. Third, the DeiT models are slightly less performing than the basic ViT versions and pre-training on ImageNet-21k instead of ImageNet results in a minimal increase in results, i.e., about 0.02.

Table 7 reports the comparison with state-of-the-art methods on the AVA dataset (for PCCD, there is no benchmark for aesthetic score assessment). Our `ViT-L/16` pre-trained on ImageNet-21k (in the table named as `ViT-L/16 - 21k`) obtains better performance than Hosu *et al*. [14] for aesthetic

[14] `https://github.com/rwightman/pytorch-image-models`

Table 6: Comparison of various transformers for image aesthetic assessment on AVA and PCCD. In each column, the best and second-best results are marked in **boldface** and underlined, respectively.

Model	Pretrain dataset	AVA			PCCD		
		SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)
DeiT-T/16	ImageNet	0.725	0.731	80.33	0.227	0.262	**98.34**
DeiT-S/16	ImageNet	0.746	0.750	80.90	0.289	0.296	**98.34**
DeiT-B/16	ImageNet	0.765	0.768	81.95	0.203	0.205	98.22
ViT-S/16	ImageNet	0.734	0.738	81.00	0.277	0.293	98.22
ViT-B/16	ImageNet	0.759	0.762	81.38	0.297	0.318	**98.34**
ViT-B/32	ImageNet	0.446	0.464	73.93	0.059	0.075	**98.34**
ViT-B/16	ImageNet-21k	0.773	0.774	81.91	0.282	0.322	**98.34**
ViT-L/16	ImageNet-21k	**0.793**	**0.793**	**82.85**	**0.369**	**0.367**	**98.34**

Table 7: Comparison of our baseline with state-of-the-art methods on the AVA dataset for image aesthetic assessment. In each column, the best and second-best results are marked in **boldface** and underlined, respectively. The "–" means that the result is not available.

Model	SRCC	LCC	Accuracy (%)
Murray *et al*. [28]	–	–	66.70
Lu *et al*. [24]	–	–	74.46
Ma *et al*. [26]	–	–	81.70
Kong *et al*. [20]	0.558	–	77.33
Talebi *et al*. [37]	0.612	0.636	81.51
Chen *et al*. [6]	0.649	0.671	**83.20**
Xu *et al*. [40]	0.724	0.725	80.90
Ke *et al*. [19]	0.726	0.738	81.15
Celona *et al*. [4]	0.731	0.732	80.75
Hosu *et al*. [14]	0.756	0.757	81.72
ViT-L/16 - *21k*	**0.793**	**0.793**	82.85

score regression with an increment of 0.04 on both SRCC and LCC. On the other hand, we are in second place for the aesthetic classification with an accuracy of 0.35% lower than Chen *et al*. [6]. Correlation metrics are more adequate than accuracy [14], and exact score estimation is more challenging and representative of the full range of scores. Therefore, we can claim that we have achieved an excellent result.

We perform experiments considering the previous backbones, apart from `ViT-B/32` which produced the worst results, for the sentiment score estimation. Results on AVA, PCCD and our RPCD are reported in Table 8. We observe the same behavior as the aesthetic assessment, that the larger models outweigh the smaller ones. We also point out that `ViT-L/16 - 21k` achieves slightly higher performance than `ViT-L/16` on AVA, vice versa on RPCD. Finally, on PCCD we get the worst results in terms of correlation and the best results for classification compared to the other two datasets.

ViT + Linear probe. The goal of this experiments is to assess to what extent the results obtained to predict the aesthetic and sentiment scores are due to the knowledge already present in the pre-trained model. We use the pre-trained ViT models as feature extractors and then we fit a linear regressor on those extracted features to predict the aesthetic score. This linear regressor was implemented as a Stochastic Gradient Descent Regressor with Scikit-Learn [30]. In Table 9 are reported the results for image aesthetic assessment on AVA and PCCD. Table 10 presents the results of the same experiment but using the sentiment score instead. Table 11 and Table 12 show the difference in performance between the trained models and the linear probe experiments. We can observe how for every case and every metric (except for the accuracy of Vit-S and ViT-L-21k on PCCD dataset to predict the aesthetic score), training the models outperform the pre-trained models (linear probes). The increase in performance is higher on AVA dataset, while PCCD and RPCD datasets do not benefit that much of further training. This may suggest that there is room for better training procedures on this datasets.

Table 8: Results obtained using ViT for estimating the sentiment score on AVA, PCCD, and RPCD. In each column, the best and second-best results are marked in **boldface** and underlined, respectively.

Model	AVA			PCCD			RPCD		
	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)
DeiT-T/16	0.492	0.507	90.46	0.187	0.220	**93.87**	0.188	0.189	64.68
DeiT-S/16	0.500	0.513	90.48	0.170	0.182	**93.87**	0.190	0.189	64.61
DeiT-B/16	0.529	0.535	90.53	0.202	0.233	**93.87**	0.216	0.218	64.62
ViT-S/16	0.498	0.512	90.41	0.192	0.211	93.75	0.202	0.199	64.65
ViT-B/16	0.527	0.534	90.46	**0.228**	**0.262**	**93.87**	0.230	0.230	65.00
ViT-L/16	0.542	**0.551**	90.50	0.212	0.236	**93.87**	**0.249**	**0.253**	**65.27**
ViT-B/16 - *21k*	0.533	0.534	90.47	0.206	0.225	**93.87**	0.228	0.228	64.73
ViT-L/16 - *21k*	**0.544**	0.550	**90.55**	0.199	0.225	**93.87**	0.246	0.246	65.08

Table 9: Results obtained by using ViT as a feature extractor followed by a linear regressor (we called ViT + Linear probe) for estimating the aesthetic score on AVA and PCCD. In each column, the best and second-best results are marked in **boldface** and underlined, respectively.

Model	AVA			PCCD		
	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)
DeiT-T/16	0.345	0.355	71.66	0.185	0.191	98.34
DeiT-S/16	0.454	0.459	74.27	0.212	0.203	98.34
DeiT-B/16	0.506	0.510	74.89	0.203	0.205	98.22
ViT-S/16	0.484	0.489	74.60	0.163	0.189	98.34
ViT-B/16	0.553	0.557	75.69	**0.254**	**0.272**	97.98
ViT-L/16	0.528	0.534	74.73	0.203	0.222	**98.46**
ViT-B/16 - *21k*	**0.570**	**0.570**	**76.44**	0.241	0.246	98.34
ViT-L/16 - *21k*	0.502	0.505	74.48	0.210	0.222	**98.46**

Table 10: Results obtained by using ViT as a feature extractor followed by a linear regressor (we called ViT + Linear probe) for estimating the sentiment score on the three considered datasets. In each column, the best and second-best results are marked in **boldface** and underlined, respectively.

Model	AVA			PCCD			RPCD		
	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)
DeiT-T/16	0.238	0.235	90.26	0.153	0.151	**93.87**	0.107	0.108	62.56
DeiT-S/16	0.300	0.303	90.27	0.139	0.135	93.40	0.128	0.128	63.09
DeiT-B/16	0.338	0.342	**90.32**	0.136	0.127	93.16	0.129	0.129	63.74
ViT-S/16	0.320	0.322	90.30	0.152	0.162	92.22	0.115	0.115	61.88
ViT-B/16	0.369	0.375	90.27	0.131	**0.166**	93.04	0.144	0.142	61.02
ViT-L/16	0.366	0.366	90.26	**0.156**	**0.166**	93.04	0.136	0.140	62.48
ViT-B/16 - *21k*	**0.392**	**0.395**	90.27	0.111	0.114	93.40	**0.172**	**0.174**	**64.59**
ViT-L/16 - *21k*	0.348	0.348	90.26	0.145	0.158	93.40	0.154	0.155	64.44

Table 11: Performance difference between ViT + Linear Probe and Aesthetic ViT (Table 6 - Table 9) for aesthetic score.

Model	AVA			PCCD		
	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)
DeiT-T/16	+0.380	+0.376	+8.670	+0.042	+0.071	+0.000
DeiT-S/16	+0.292	+0.291	+6.630	+0.077	+0.093	+0.000
DeiT-B/16	+0.259	+0.258	+7.060	+0.000	+0.000	+0.000
ViT-S/16	+0.250	+0.249	+6.400	+0.114	+0.104	-0.120
ViT-B/16	+0.206	+0.205	+5.690	+0.043	+0.046	+0.360
ViT-B/16 - *21k*	+0.203	+0.204	+5.470	+0.041	+0.076	+0.000
ViT-L/16 - *21k*	+0.291	+0.288	+8.370	+0.159	+0.145	-0.120

Table 12: Performance difference between ViT + Linear Probe and Aesthetic ViT (Table 8 - Table 10) for sentiment score.

Model	AVA			PCCD			RPCD		
	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)	SRCC	LCC	Acc. (%)
DeiT-T/16	+0.254	+0.272	+0.200	+0.034	+0.069	+0.000	+0.081	+0.081	+2.120
DeiT-S/16	+0.200	+0.210	+0.210	+0.031	+0.047	+0.470	+0.062	+0.061	+1.520
DeiT-B/16	+0.191	+0.193	+0.210	+0.066	+0.106	+0.710	+0.087	+0.089	+0.880
ViT-S/16	+0.178	+0.190	+0.110	+0.040	+0.049	+1.530	+0.087	+0.084	+2.770
ViT-B/16	+0.158	+0.159	+0.190	+0.097	+0.096	+0.830	+0.086	+0.088	+3.980
ViT-L/16	+0.176	+0.185	+0.240	+0.056	+0.070	+0.830	+0.113	+0.113	+2.790
ViT-B/16 - *21k*	+0.141	+0.139	+0.200	+0.095	+0.111	+0.470	+0.056	+0.054	+0.140
ViT-L/16 - *21k*	+0.196	+0.202	+0.290	+0.054	+0.067	+0.470	+0.092	+0.091	+0.640

NIMA. We compare the previous ViT models with a model from the literature, i.e., NIMA [37], for sentiment score prediction. NIMA is trained by us using the code released by its authors. We use an ImageNet-trained VGG-16 as the backbone. Input images are resized to a fixed spatial resolution of 224×224 pixels. As in [37], for model optimization we exploit the Earth Mover's Distance (EMD):

$$EMD(\hat{\mathsf{q}}, \mathsf{q}) = \left(\frac{1}{N} \sum_{k=1}^{N} |CDF_{\hat{\mathsf{q}}}(k) - CDF_{\mathsf{q}}(k)|^r \right)^{\frac{1}{r}}, \tag{4}$$

where $\hat{\mathsf{q}}$ and q are the ground-truth and the predicted score distributions, respectively. Finally, $CDF_*(k)$ is the cumulative distribution function, r equal to 2 is used to penalize the Euclidean distance between the CDFs. We use the probability distribution on the three sentiment polarity classes p as ground-truth. We optimize the model by using Stochastic Gradient Descent (SGD) with learning rate of $5e$-3 and batch size equal to 64 for 100 epochs. We use an early stopping policy based on validation loss with a patience term of 10 epochs.

Summary. Experiments with `ViT + Linear probe` have shown that pre-trained ViTs for image recognition do not work well for predicting aesthetic and sentiment scores. It is therefore necessary to train the model to learn the characteristics that best encode the various aspects of aesthetics. Table 11 and Table 12 report the difference in performance between `ViT + Linear prob` and `AestheticViT` models for aesthetic score estimation and sentiment score estimation, respectively. This way we highlight the gain obtained thanks to the training of the backbones.

Among the various tested models, `ViT-L/16 - 21k` achieved the best results on both AVA and PCCD for aesthetic assessment. It also outperforms state-of-the-art aesthetic assessment methods on the AVA dataset. On the other hand, for the prediction of the sentiment score the `ViT-L/16` model obtained the best performance regardless of the dataset used for pre-training.

B.2 Image Aesthetic Critique Generation

We verify the use of the proposed dataset for the generation of aesthetic image critique by using Bootstrapping Language Image Pre-training (BLIP) [22]. It is a method for the unified understanding and generation of the visual language. A pre-trained ViT-B/16 on the COCO dataset is finetuned for aesthetic captioning by exploiting the AdamW optimizer with initial learning rate equal to $1e$-5, weight decay of 0.05, and a cosine learning rate schedule. We train for 5 epochs using a batch size of 16 samples. During inference, we use beam search with a beam size of 3, and set the minimum and maximum generation lengths as 20 and 50, respectively.

C Resources Used

In this section we briefly list the resources used to carry out this work:

- Host machines: The machines used by the authors, each of them with access to a GPU NVIDIA GeForce RTX 2080 Ti.

- Access to internal cluster[15] with access to various instances with the following GPUs: NVIDIA GeForce RTX 2080 Ti and NVIDIA TITAN RTX.
- A part of the experiments, but not all, were logged to Weights & Biases[16], which registered the time used for those experiments, summing up a total of 2500 hours.

D Ethical considerations

This section comments on the Ethics Guidelines [17] of NeuroIPS. In particular, we comment on various of the points brought on this guidelines:

- **Personally identifiable information and data collection**. The samples in our dataset are attached to the user ID. While this provides a first level of anonymity to the users, it is fairly straight forward to access the public user profile, which may contain identifiable information the user had previously agreed to share and may be identifiable. Every user consents the collection of this information and accepts the Reddit's Privacy Policy [18], where it is stated that *"[...] Reddit also allows third parties to access public Reddit content via the Reddit API and other similar technologies. [...]"*. Thus, not every user has been directly asked for consent to include data produced by them in this dataset, but this consent is comprised under the Privacy Policy and the Reddit API terms of Use. We expand on this in the Section F. However, we point out that, a priori, disclosing that a person has any activity or belongs to the `r/photocritique` subreddit does not involve degrading or embarrassing such person.
- **Data consent**. As pointed out above, every user consents accepts the Reddit's Privacy Policy, where it is stated that *"[...] Reddit also allows third parties to access public Reddit content via the Reddit API and other similar technologies. [...]"*. The use of Reddit as a source of data for a large variety of scientific research has had an important impact in several fields as described in The Pushshift Dataset work [2]. We acknowledge that there is not explicit consent of the users to use their data for scientific purposes. However, we considered this to be covered by Reddit's Privacy Policy. Hence, instead of collecting and storing the metadata and data produced by users, we provide the identifiers necessary to access the data and the tools to construct the dataset.
- **Explicit content**. Images may contain explicit content of people. The first of the community rules state *1. Post only photos you took. Do not post a photo unless you took it! [...]* Thus, it is assumed this rule implies that the user posting a new image is the owner of the photography and hence has the right to distribute it. The sensitive content is labeled as "NSFW" in the dataset.
- **Bias against people of a certain gender, race, sexuality, or who have other protected characteristics**. This is a multi-factor issue that must be addressed from different perspectives and is beyond the scope of the first analyses presented in this paper to show the usability of this new data source. For instance, questions such as the impact of gender, race or sexuality on the perceived aesthetics of an image or how these images are critiqued are completely out of the scope of this work. However, we must note and acknowledge the work of the team of moderators of the `r/photocritique` community. Not only they approve each of the posts published in the community, but it is clearly stated that inappropriate or disrespectful posts are banned. As stated in the rules of the community: *Lewd comments or those deemed by the moderation team to be grossly inappropriate will result in a permanent ban. You have been warned.*. And as stated in the critiques guidelines: *We do not allow [...] inappropriate/sexist/racist comments.*.
- **Filtering of offensive content**. Due to the scale of the dataset, it has not been feasible to double check every post complies with the community rules. However, we have included a preliminary analysis of the presence of offensive content in the dataset (See Appendix A.5), in which we found that the predicted offensive content in the comments of the dataset is under 4%.

[15] `https://scicomp.ethz.ch/wiki/Euler`
[16] `https://wandb.ai/`
[17] `https://nips.cc/public/EthicsGuidelines`
[18] `https://www.reddit.com/policies/privacy-policy`

E License

We comply with Reddit User Agreement[19], Reddit API terms of use [20] and PushShift database Creative Commons License [21]. In particular, we refer to the Section 2.d of Reddit API Terms of Use, which states: "User Content. Reddit user photos, text and videos ("User Content") are owned by the users and not by Reddit. Subject to the terms and conditions of these Terms, Reddit grants You a non-exclusive, non-transferable, non-sublicensable, and revocable license to copy and display the User Content using the Reddit API through your application, website, or service to end users. You may not modify the User Content except to format it for such display. You will comply with any requirements or restrictions imposed on usage of User Content by their respective owners, which may include "all rights reserved" notices, Creative Commons licenses or other terms and conditions that may be agreed upon between you and the owners." We do not provide access to any data directly, but a list of IDs associated with a post on Reddit. This information is then used to retrieve the images, comments and metadata using the provided tools after obtaining a license key for the official Reddit API. Moreover, we do not modify the original content by no means, while we provide the necessary tools to process the data and run the same experiments we carried out.

We release the dataset under the Creative Commons Attribution 4.0 International license.

F Datasheet for RPCD

In this section we detail the datasheet presented in [9] for documenting the proposed dataset. Note that, while we do not provide any data other that the IDs associated to Reddit posts, we answer the questionnaire considering the constructed dataset resulted from using our code.

F.1 Motivation

- **For what purpose was the dataset created?**

  RPCD was created to drive the research progress in both image aesthetic assessment and aesthetic image captioning. The proposed dataset addresses the need for images acquired with modern acquisition devices and photo critiques that give a better understanding of how the aesthetic evaluation is carried out.

- **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  This dataset was created by the authors on behalf of their respective institutions, ETH Media Technology Center and University of Milano-Biococca.

- **Who funded the creation of the dataset?**

  The creation of this dataset was carried out as part of the Aesthetic Assessment of Image and Video Content project[22]. The project is supported by Ringier, TX Group, NZZ, SRG, VSM, viscom, and the ETH Zurich Foundation on the ETH MTC side.

F.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

  Each instance is represented as a tuple containing one image and several photo critiques, where the images are JPEG files and the photo critiques are in textual form.

- **How many instances are there in total (of each type, if appropriate)?**

  RPCD consists of 73,965 data instances. Specifically, there are 73,965 images and 219,790 photo critiques.

[19] https://www.redditinc.com/policies/user-agreement/

[20] https://docs.google.com/a/reddit.com/forms/d/e/1FAIpQLSezNdDNK1-P8mspSbmtC2r86Ee9ZRbC66u929cG2GX0T9UMyw/viewform

[21] https://zenodo.org/record/3608135#.Yp3XEXZBw2w

[22] https://mtc.ethz.ch/research/image-video-processing/aesthetics-assessment.html

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
  The dataset contains all samples (posts) available at the moment of collection, from the origin of the forum until the moment of collection. Additionally, included posts had to meet the following criteria:
  - The post has at least an image which could be retrieved.
  - The post has at least one comment critiquing the image
  - The post is not a discussion thread, a type of post to encourage general discussion in the forum.
- **What data does each instance consist of?**
  Each data instance consists of an image and one or more textual photo critiques.
- **Is there a label or target associated with each instance? If so, please provide a description.**
  There is no label associated with each sample. However, in this work we propose a method to compute said label, which is calculated using the processing scripts.
- **Is any information missing from individual instances?**
  Some of the samples in the dataset might be missing at the moment of future retrievals due to the users removing the data from Reddit.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
  Every image and comment in the dataset is associated with the user who created the post. Moreover, we build the tree of comments of the different users criticizing an image. However, the data is downloaded by using only the post IDs.
- **Are there recommended data splits (e.g., training, development/validation, testing)?**
  We provide the data splits we used in our experiments in the repository and they are used to retrieve the posts we used, although we encourage the use of other splits. The splits were randomly generated to divide the dataset in 70% train, 10% validation and 20% test splits.
- **Are there any errors, sources of noise, or redundancies in the dataset?**
  The source of data itself could be considered a source of noise. Additionally, we have not evaluated the case in which an image is posted by an user several times in different posts, although we consider this event to be non-existent or insignificant.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
  The dataset links to resources available on Reddit and Pushshift. In particular, posts and their metadata (including the URLs to images) are retrieved from Pushshift, while the comments are retrieved directly from Reddit. There is no guarantee that the dataset will remain constant, as this depends on the users exercising their rights to remove their content from the dataset sources. For this same reason, there are not any archival versions of the complete dataset available online. In order to retrieve the dataset in the future, Reddit API credentials are needed. Please, refer to the instructions about how to obtain the credentials[23].
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
  The dataset does not contain any confidential data as both images and comments are publicly available in Reddit.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
  There are data samples depicting explicit nudity with aesthetics purposes, and we acknowledge that this may be problematic for some people. According to the subreddit rules, this content must be marked: *"Not Suitable for Work (NSFW) must be marked. [...] Please keep NSFW posts respectful. Nothing that would be considered pornography."* For this reason, the dataset processing script creates a NSFW column in the dataframe to easily filter this content.

[23] `https://www.reddit.com/wiki/api/`

- **Does the dataset relate to people?**
  Yes, some of the images contain people or the main subject is a person.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?**
  No.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**
  All posts and comments are linked to users, which may be identifiable depending on the data made available by the user. Additionally, posts and comments may contain information linking to other social media which could serve to identify a certain user.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
  The retrieved data might contain sensitive data publicly disclosed by the users. However, we do not expect this to be common at all, and we would be surprised that some kinds of sensitive information are present in the community (financial, health, biometric, genetic or governmental data).

F.3 Collection process

- **How was the data associated with each instance acquired?**
  The data was directly observable (posts in Reddit stored in Pushshift's and Reddit's servers).
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**
  Software API to access both Reddit and Pushshift.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** NA.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
  Nobody was involved in the data collection process since all data was already available and observable.
- **Over what timeframe was the data collected?**
  Tha data was collected in February 2022, and comprises posts and comments in the span from May 2009 (first posts in the subreddit) to February 2022 (collection date).
- **Were any ethical review processes conducted (e.g., by an institutional review board)?**
  No ethical review process was conducted previous to the ethical review of this conference.
- **Does the dataset relate to people?**
  Yes, but not exclusively.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  Third party sources (Reddit and Pushshift).
- **Were the individuals in question notified about the data collection?**
  No.
- **Did the individuals in question consent to the collection and use of their data?**
  According to Reddit's Privacy Policy[24], which is accepted by every user upon registration, *"Reddit also allows third parties to access public Reddit content via the Reddit API and other similar technologies."* . Moreover, we note that no data from the users is made directly available in the dataset. It only contains the IDs of the posts and the tools to retrieve them from Reddit and Pushshift.

[24] `https://www.reddit.com/policies/privacy-policy`

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
  Users may remove their data from Reddit and Pushshift using their respective privacy enforcing mechanisms. Thus, they would be removing their data from the dataset.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**
  As we note above, no data from the users is made directly available in the dataset. The dataset only contains the IDs of the posts and the tools to retrieve them from Reddit and Pushshift.

F.4 Processing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
  We provide scrips to automatically process the downloaded raw posts. Only first level comments are kept, posts with no comments or whose image is no longer available are filtered.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
  The raw posts need to be downloaded for further processing.
- **Is the software used to preprocess/clean/label the instances available?**
  Yes. The software for downloading and preparing the dataset is available on our GitHub repository [25].

F.5 Uses

- **Has the dataset been used for any tasks already?**
  RPCD is introduced and used in the paper `Understanding Aesthetics with Language: A Photo Critique Dataset for Aesthetic Assessment`.
- **Is there a repository that links to any or all papers or systems that use the dataset?**
  Papers using RPCD will be listed on the PapersWithCode web page[26].
- **What (other) tasks could the dataset be used for?**
  RPCD can be used for modelling works in the areas of knowledge retrieval and multimodal reasoning.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  No, there are no known risks to the best of our knowledge.
- **Are there tasks for which the dataset should not be used?**
  RPCD should not be used for automatically judging a photographer's skills based on the photo critiques. The latter, in fact, are to be understood as highly subjective judgments that depend on the emotions and background of the commentators and could go beyond the mere technical evaluation of the shot.

F.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
  Yes, the dataset is made publicly accessible.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**
  See our GitHub repository [25] for downloading instructions. RPCD has the following DOI: `10.5281/zenodo.6985507`.

[25] `https://github.com/mediatechnologycenter/aestheval`
[26] `https://paperswithcode.com/dataset/rpcd`

- **When will the dataset be distributed?**
  RPCD will be released to the public in August 2022.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
  We release the dataset under the Creative Commons Attribution 4.0 International license[27].
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
  No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
  No.

F.7 Maintenance

- **Who is supporting/hosting/maintaining the dataset?**
  RPCD is supported and maintained by ETH MTC and University of Milano-Bicocca. The post IDs are available on Zenodo, the posts are on Reddit and Pushshift, and the code for automatically retrieving the posts is on GitHub.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  By emailing to {daniel.veranieto,clabrador}@inf.ethz.ch or luigi.celona@unimib.it. By opening an issue on our GitHub repository [25].
- **Is there an erratum?**
  All changes to the dataset will be announced on our Zenodo repository [28].
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')?**
  All updates (if necessary) will be posted on our Zenodo repository [28].
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
  The data related to users is stored on Reddit and Pushshift servers, and their data retention policies apply.
- **Will older versions of the dataset continue to be supported/hosted/maintained?**
  All changes to the dataset will be announced on our Zenodo repository [28]. Outdated versions will be kept around for consistency.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
  Any extension/augmentation by an external party is allowed under the release license. The dataset could be easily extended with other communities and other time periods using the available scripts. In order to add the extended version to the existing repositories, please contact the authors.

[27] `https://creativecommons.org/licenses/by/4.0/`
[28] `https://doi.org/10.5281/zenodo.6985507`