What's the Fairest of Them All? Aesthetic Assessment of Visuals

Marc Willhaus
MTC, ETH Zürich
marc.willhaus@inf.ethz.ch

Daniel Vera
MTC, ETH Zürich
daniel.veranieto@inf.ethz.ch

Clara Fernandez-Labrador
MTC, ETH Zürich
clabrador@inf.ethz.ch

Severin Klingler
MTC, ETH Zürich
severin.klingler@inf.ethz.ch

ABSTRACT

Attractive images and videos are the visual backbone of journalism and social media. From trailers to teaser images to image galleries, appealing visuals have only grown in importance over the past years. Especially online, eye-catching visual content can significantly impact user engagement. However, selecting the best shots from a long video or selecting the perfect image from a vast image collection is a challenging and time-consuming task. This paper presents a system to automatically assess image and video content from the perspective of aesthetics. While this is a highly subjective task, we find that it is possible to combine expert knowledge with data-driven information to perform such an assessment. In order to do so, we identify relevant aesthetic features together with experts from the media industry and implement machine learning algorithms to infer them from the visual content. We combine the features under a single aesthetics retrieval system that allows users to sort uploaded visuals according to an aesthetic score and interact with additional photographic, cinematic, and person-specific features. The system is built into a containerized application to guarantee reproducibility. A demo video of our tool is available [1].

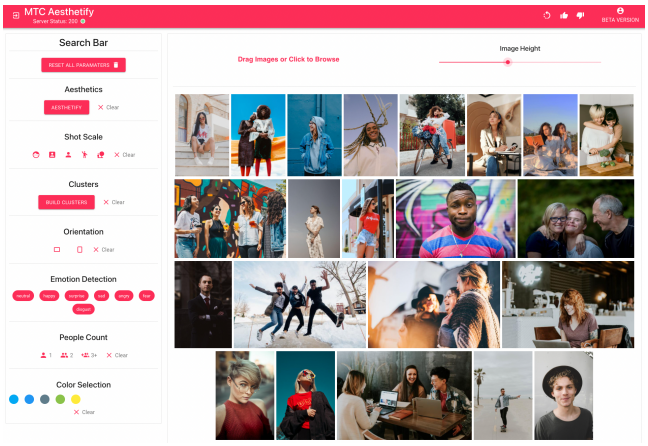

Figure 1: Our aesthetics tool supports multiple aesthetic features providing an easy and fast way to find aesthetically pleasing images and video shots. Images are from Unsplash.

1 INTRODUCTION

Choosing appealing images and videos from a larger pool of images or long videos is an integral part of many workflows: from journalists, who have to select the perfect teaser image for their article to video editors who create movie trailers for different distribution channels. However, identifying the best shot or frame that will capture an audience's attention can be tedious and ineffective.

Image and video retrieval systems developed to support these tasks typically focus on retrieving visuals based on available metadata or based on a semantic analysis of the content. Some examples are Flickr[2], Google Photos[3], Vitrivr [22], Pexels[4] or Unsplash[5]. While content-based retrieval systems have received great attention, systems that consider aesthetics are underexplored due to the subjective nature of the problem. In this work, we are interested in developing a retrieval system to find images or videos that not only are relevant in terms of its content but also are aesthetically pleasing and make good use of photographic and cinematic principles. Many different factors are considered by experts when selecting aesthetic images and video shots: from the image composition, the color, the camera position, objects in the scene and facial expressions to whether the frame is representative of the story, e.g., showing the main character or being relevant for the genre. In the last decade, complex models such as convolutional neural networks have been explored to address this problem known as image aesthetic assessment (IAA). These models can be classified into two types. Namely binary classification methods [16, 19], which classify images as "good" or "bad" using a specific threshold, and regression methods [3, 7, 11, 17, 18, 25], which predict a continuous aesthetic score per image based on human ratings. However, the predicted scores do not provide explicit information about the aesthetic principles that characterize the images. Hence, we believe an aesthetic score is not sufficient for the aesthetic assessment task.

We conducted qualitative interviews with various experts across different media companies. The main purposes were to learn more about the video and image selection process and to identify what features these experts consider relevant for visual aesthetics. Specifically, we interviewed people from TX Group[6] and SRF[7] (Swiss Radio and TV, Public Service Broadcaster), two big media houses in Switzerland. The interviews concerned four main questions. Namely the use case on which they work, data sources, criteria to select aesthetically pleasant visuals and what would help to improve their daily workflow. The use cases reflect the commercial and societal impact that our tool could provide. To name a few: find the teaser image for a news article, select the best image of a social

[1] https://mediatechnologycenter.github.io/mtc-aesthetics-website
[2] https://www.flickr.com/
[3] https://www.google.com/photos/
[4] https://www.pexels.com/
[5] https://unsplash.com/
[6] https://tx.group/
[7] https://www.srf.ch/

media post and find a good thumbnail or the best shots for trailer production for movies, documentaries or TV shows. Additionally, we believe the same tool could be useful in marketing departments, to help users decide on the visuals they upload to social networks or to create professional presentations. The data source may be very diverse, from images coming from professional photographers to archive black and white images or videos. This diversity poses several challenges, i.e. how to discriminate between aesthetically good and bad images, but also how to select the best from a group of already highly aesthetic images or from a group of poor quality images. We collected relevant insights from their criteria to select appealing visuals that inspired us to select the top aesthetic features to expose in our retrieval system. These features are meant to complement the aesthetic scores. See Section 2 for an in-depth review. The technical challenge and impact comes from unifying all the relevant features in a single and easy-to-use application, which would improve the image or video search in terms of getting a better sorting, save time, intuitive aesthetic search and more variance in the results. EyeEm[8] is one of the few retrieval systems that consider aesthetics and hence the closest method for comparison. They propose a supervised algorithm to predict aesthetic scores that receives triplets of images, two well-crafted and one mediocre photographs, and learns commonalities between the first two and differences with the latter. However, solving a binary classification problem, EyeEm cannot provide any sorting based on aesthetics but only discard "bad" images. EyeEm also exposes additional aesthetic features but they are not clearly defined. Additionally, it lacks the capabilities to perform this aesthetic analysis for video content. Therefore, to the best of our knowledge, we propose the first full-stack multimedia system that tackles the image and video retrieval problem from the perspective of aesthetics in a unified framework.

To summarize, our main contributions are the following. We select relevant features for aesthetic assessment of image and video content based on in-depth interviews with experts from the media industry. We develop machine learning algorithms to label visuals with the selected aesthetic features. We create a full-stack multimedia retrieval system that integrates and exposes them in an intuitive way. Also different from the aforementioned frameworks, instead of working with stock images our tool allows the user to upload personalized content which is analyzed in a matter of minutes. Our tool comes with a ready-to-use Docker image and user interface to help editors, journalists and average-users to sort the uploaded content according to a continuous aesthetic score and interact with carefully selected features to find the best image and video shots in a quicker and intuitive way.

2 AESTHETIC FEATURES

Since the aesthetics of an image or video is an ill-defined concept, we propose a selection of features relevant for the aesthetic assessment of images and video content. This selection is the result from interviews with professionals from several media companies and in-depth discussions with experts from TX Group and SRF. We present a taxonomy of the selected aesthetic features with four main categories, namely *'Photographic', 'Cinematic', 'Technical'* and *'Person-specific'*. We include specific features tailored to images and shots showing people as they are an important and frequent subject of media coverage for movies, TV shows, news and in social networks. Besides the aforementioned classification, we make a distinction of such features into two different levels of abstraction: low-level and high-level features. Below, we detail the principles underlying such a categorisation and go through the ML algorithms used to automatically extract the features from image and video data. Our tool therefore combines experts knowledge (taxonomy) and data driven information (inferred by various ML algorithms). We give high-level descriptions of the implementations only, due to the scope of this paper.

2.1 Low-level Features.

We consider mainly *'Technical'* features such as sharpness, luminance, motion blur, uniformity, saturation, contrast or complexity. Such features are hard to expose individually in a meaningful way however, they are all together highly relevant to determine an overall aesthetic score or signature of an image or video shot. Hence, we propose an aesthetic predictor based on vision transformers (ViT) [8] that learns the aforementioned features implicitly to regress a continuous aesthetic score per image or shot keyframe. The model is trained on the public AVA dataset [19] in a supervised manner using the mean squared error during optimization. We give more details about the aesthetics predictor in Section 3. The aesthetic score represents the most relevant feature of our framework, as it reflects the subtleties photographers exploit and is used to sort the images and video shots according to its appealing value. This allows the user to select better visuals faster. We also exploit the low-level features to create clusters of images. The interviewees raised a common concern, which is the lack of variance in the results when they use content-based image retrieval platforms. By clustering the images according to the similarity of their low-level features, we can reduce the, sometimes overwhelming, quantity of results to only a few. For every batch of data uploaded by the user, we use the Elbow method [23] to determine the quantity of clusters. If there is enough variation in the input data, clustering is not necessary.

2.2 High-level features.

We believe that an aesthetic score is not sufficient for the aesthetic assessment task. For that reason, we define additional features that should be accessible and can be easily exposed to the user. We list them below. Please note that some are only meant for videos (v).

Shot scale: 'Cinematic' feature to distinguish five scale types, namely extreme close-up, close-up, medium, full and long. We developed a Subject Guidance Network (SGNet) inspired by [21] to perform such classification. The key idea is to use a subject map to determine the portion occupied by the subject with respect to the frame. The model is trained on the public MovieNet dataset [12] and optimized using the cross-entropy loss. Filtering for the shot scale allows users to find visuals that emphasize the location (long), event (medium/ full) or the identity of a subject (extreme close-up/close-up).

Shot angle: 'Cinematic' feature that differentiates eye-level, low and high frames with a standard convolutional neural network

[8] https://www.eyeem.com/

(CNN). Specifically, we use ResNet-50 [10] optimized using the cross-entropy loss. The user can use the angle feature to select an image with neutral perspective or to emphasize power dynamics (low/ high).

Shot boundaries (v): 'Cinematic' feature. A shot represents a series of frames that runs for an uninterrupted period of time and is the minimal visual unit of a video. We automatically detect changes in videos and split them into shots [12] for our video application. The shots are exposed to the user and the rest of aesthetic features are computed on the shots independently.

Shot keyframe (v): 'Cinematic' feature. Refers to a compact and non-redundant representation of the shot. We select a keyframe per shot based on the aesthetic scores and on a stillness metric computed by convolving the input image with the Laplacian operator and taking the variance. The keyframe is used as teaser image of the shot in the tool.

Shot movement (v): 'Cinematic' feature whereby we distinguish four movement types, namely static, motion, push and pull. We developed a ML model inspired by [21] to perform such classification. The key idea is to separate the background from the subject and rely on the background changes which are closely related to the camera motion. The model is trained on the public MovieNet dataset [12] and optimized using the cross-entropy loss. The shot movement feature allows the user to select a shot based on whether they want to track a moving object (motion), narrate an event (static), emphasize the main subject (push) or gradually reveal the surroundings (pull).

People Count: 'People specific' feature to determine the number of protagonists in the image. We run a face detection model [2] and use a threshold of confidence selected experimentally to perform the face counting. The model is trained on the public FER2013 dataset [9].

Emotion Detection: 'People specific' feature. We use the aforementioned face detector algorithm to run an emotion classifier [2] on the cropped faces. The classifier predicts probabilities for seven emotions, namely anger, fear, sadness, disgust, happiness, contempt and surprise. The model is trained on the public FER2013 dataset [9] and optimized using the cross-entropy loss. The conducted interviews revealed that the overall mood of an image is very important when selecting key visuals. In images showing people, the face expressions have large impact on the perceived mood of the image, which is why we expose it here.

Color Selection. 'Photographic' feature that extracts the main colors present in an image or video shot by clustering pixel intensities. We define 20 base colors distributed over the color spectrum to which we assign the main colors of an image by computing the euclidean distance between RGB values. The color selection can be used to discover the color palette of a filmmaker or to choose an image which fits well to the color scheme of a journal or a website.

3 AESTHETICS PREDICTOR

We propose a model to aesthetically rank images based on Vision Transformer (ViT) [8]. We use the weights of ViT pre-trained on ImageNet [6] and JFT300 [24]. Then, we add a fully-connected layer which is randomly initialized to predict the final aesthetic score. The model is trained on the public AVA dataset [19] in a supervised manner using the mean squared error during optimization. The AVA database contains images with diverse resolutions, with an average resolution of 629 × 497 pixels. We proportionally rescale input images to have maximum input dimension (width or height) of 700 pixels, keeping their original aspect ratio. To use images with different resolutions, [13] proposed a hash-based 2D spatial embedding and a scale embedding to support positional encoding. Instead, we obtain better performance by resizing the input positional embedding accordingly with the image resolution performing bilinear interpolation during training. Further work is necessary to solve the inefficiency caused by the use of batch size of 1. We train the model until convergence, which happened at epoch 2. The learning rate is set experimentally to 10^{-6}. We report in Table 1 Spearman's Rank Correlation Coefficient (SRCC), Linear Correlation Coefficient (LCC) and Accuracy on AVA dataset [19], comparing our results to previous works (we use their self-reported results). The accuracy is computed defining as high quality images those with an score above 5, and poor quality otherwise.

Model	Accuracy (↑)	SRCC (↑)	LCC (↑)
Murray et al.[19]	66.70%	-	-
Lu et al.[15]	74.46%	-	-
Ma et al. [18]	81.7%	-	-
Kong et al. [14]	77.33%	0.558	-
Talebi et al. [25]	81.51%	0.612	0.636
Chen et al. [5]	**83.20**%	0.649	0.671
Xu et al. [28]	80.9%	0.724	0.725
Ke et al. [13]	81.15%	0.726	0.738
Celona et al. [4]	80.75%	0.731	0.732
Hosu et al. [11]	81.72%	0.756	0.757
Ours	81.43%	**0.762**	**0.759**

Table 1: Models analysis.

We obtain a slightly better performance than Hosu et al.[11], while following a simpler approach. We also agree with their reasoning of the suitability of correlation metrics rather than accuracy, since image labels are defined arbitrarily and it is usually reported on the original test set, which is imbalanced. Additionally, correlation metrics are representative of the entire range of quality scores.

4 AESTHETICS TOOL

We integrate the carefully selected aesthetic features (section 2) in a single framework and provide an interface for the user. The tool has been already tested by our media industry partners, TX group and SRF. In a glance, the user can upload a batch of images or a video and quickly receive aesthetic feedback by getting the content sorted by an aesthetic score. The user can also interact with additional features such as number of people, emotions, scale of shot, and color scheme to refine the search. In this section we go through the main components of the tool (4.1), we explain the

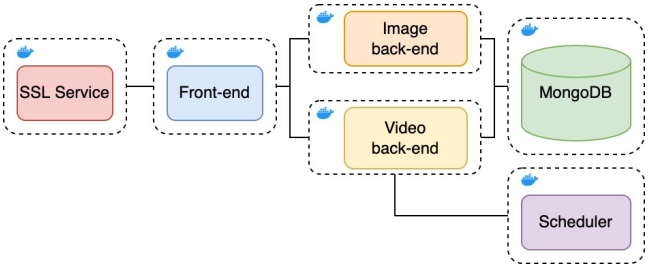

Figure 2: Aesthetics Tool Architecture.

workflow of the image and video application (4.2) and describe the containerized setup (4.3).

4.1 Tool Components

An overview of the aesthetics tool architecture is shown in Figure 2. The details of each component are detailed below.

Back-end: We orchestrate in the back-end the different ML algorithms needed to infer the aesthetic features, which work sequentially. All back-end calls are proxied to the back-end which is running in Python. Python is chosen since all the ML models are implemented in Python and use either the PyTorch framework [20] (low-level and *'Cinematic'* features) or the TensorFlow [1] framework (*'Person-specific'* features). Flask is used to implement the Application Programming Interface (API) endpoints. As a Web Server Gateway Interface (WSGI) the Gunicorn library is used due to its simplicity to implement and compability to Flask. Firebase Authentication serves as a authentication solution which comes out of the box with an interface for user management.

Front-end: Interface with the user. Figure 1 shows the interface of the aesthetics tool. The front-end is built using React JS (JavaScript), an open-source JavaScript library for user interface components. The interface is designed so that it is easily usable but also reflects the back-end structure. Nginx serves as webserver and as a proxy. All requests directed to the */api* routes are proxied to the Gunicorn Python Server. The Firebase Authentication is used to secure the front-end from loading if the authentication is not given. Additionally, a token is fetched while logging in from Firebase and is sent to the back-end API routes within the header of the requests

Database: MongoDB is used as a persistent storage for all related information of the application. Within the database four data collections are created, two for the image and two for the video application. The first collection stores the metadata of the images, including the inferred features. The second collection stores the low-level features (ViT embeddings), so that the clustering can be performed. The third collection stores the video metadata. The fourth collection stores information about the video shots and the inferred features. For the image application, the features can be queried directly from the image collection, whereas for the video case, a database entry is created to save the shot metadata and the queries from the front-end are run against the shots collection. If the user wishes to see an overview of all videos, queries are run against the video collection.

Scheduler: We develop the image and video applications to be as similar as possible, although some specifics are required. Due to the workload of processing videos, a queuing system is implemented for the video application. Like so the upload and the inference are separated and multiple objects can be uploaded at the same time. The scheduler is an implemented Python script and works similar to a Cron job. Every 10 seconds the inference endpoint on the video API is called to check whether a new element in the video database is available. If a new object is available, it is transferred to the inference handler. The inference handler splits the video into its different shots and the ML models perform inference on each shot.

4.2 Tool walkthrough

We provide Figure 1 to show the interface and guide the reader through the tool. Right after logging in, the user can select between the image and the video applications. In both applications the header shows general information, such as the server status, and navigation options, such as log out and switching the application from image to video. The images or videos are uploaded with a drag and drop option or are selected from the user's computer. Then, the uploaded content is directly passed to the inference endpoint where the ML models infer the aesthetic features. The images or video are stored on the file-system and the inferred features are saved as metadata into the database. This method provides the tool with a highly responsive solution. Once the inference is done, the images or video shots are exposed. If an image or shot is clicked, a preview is shown with its metadata, i.e. aesthetic features. The user can click on "like" or dislike", which are represented as thumbs (up and down respectively). The user may want to download the liked images or shots but also the disliked ones to send them for further review and avoid similar images or shots in the future. On the left side a query-bar interacts directly with the database, where all the aesthetic features are available. All selected options in the query-bar are passed to the API as parameters and it returns the queried images or shots back from the database. As an example, if the user selects the *aesthetify* option, the images or video shots get sorted according to the predicted aesthetic scores, showing the best content on the top. The *build clusters* option is useful when the uploaded data contain repetitive scenes. It offers a quick overview of the uploaded content showing only a representative (most aesthetic) image per cluster. By clicking on any of the returned images, the user can still explore the rest of the images in the cluster. The remaining aesthetic features are exposed with their respective classification scheme, e.g. for *emotion detection*, the seven emotion labels are available. To ease the analysis of long videos, we also introduce the *timeframe* sliding bar. The user can reset all the parameters at any moment to start a new search. We support multiple image and video formats, including raw formats which is recommended by expert photographers to not lose quality.

4.3 Docker Setup.

To simplify the setup process and rebuilding the environment, we built the entire tool into a containerized application. Additionally a docker-compose SSL docker service is used [26] which provides automated SSL certificate generation. Most of our models depend on a Graphical Processing Unit (GPU), requiring specific combinations

of CUDA versions and Tensorflow or Pytorch versions. We propose a containerized solution which is favorable in this cases. Within the container an API endpoint can be run and the requests are sent from the front-end to the back-end container. We create one container for the image API and one container for the video API. With one container for a dedicated API data traffic and complexity is minimized. Within a distributed system, e.g. on the cloud, we suggest to run every ML model within its own docker container and a separate container for the API endpoint. If the inference is not time relevant, a CPU deployment would also be possible. After testing multiple options, the usage of Miniconda in combination with a specific CUDA version, Tensorflow or Pytorch image lead to the best reproducible results [27].

5 CONCLUSION

This paper investigates the aesthetic assessment of images and video content. While visual aesthetics is an ill-defined concept, we find that combining experts knowledge and data driven information, we can create a framework to perform such assessment. Specifically, we define a taxonomy of aesthetic features together with experts from the media industry, we develop ML algorithms to infer each of them and combine everything under a single system with an easy-to-use and attractive interface that serves as an aesthetic retrieval system. Future work could focus on optimizing the inference time of the ML models and adding new features such as shot focus (shallow, out, deep) and composition features (symmetry, rule of thirds, repetition).

ACKNOWLEDGMENTS

This project is supported by Ringier, TX Group, NZZ, SRG, VSM, viscom, and the ETH Zurich Foundation.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. {TensorFlow}: A System for {Large-Scale} Machine Learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 265–283.

[2] Octavio Arriaga, Matias Valdenegro-Toro, Mohandass Muthuraja, Sushma Devaramani, and Frank Kirchner. 2020. Perception for Autonomous Systems (PAZ). arXiv:2010.14541 [cs.CV]

[3] Tunç Ozan Aydın, Aljoscha Smolic, and Markus Gross. 2014. Automated aesthetic analysis of photographic images. *IEEE transactions on visualization and computer graphics* 21, 1 (2014), 31–42.

[4] Luigi Celona, Marco Leonardi, Paolo Napoletano, and Alessandro Rozza. 2021. Composition and Style Attributes Guided Image Aesthetic Assessment. *arXiv preprint arXiv:2111.04647* (2021).

[5] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. 2020. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14114–14123.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (2017), 80–106.

[8] 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).

[9] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*. Springer, 117–124.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 2019. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9375–9383.

[12] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*. Springer, 709–727.

[13] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. https://doi.org/10.48550/ARXIV.2108.05997

[14] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*. Springer, 662–679.

[15] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. 2014. RAPID: Rating Pictorial Aesthetics Using Deep Learning. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. Association for Computing Machinery, New York, NY, USA, 457–466. https://doi.org/10.1145/2647868.2654927

[16] Xin Lu, Zhe Lin, Xiaohui Shen, Radomír Mech, and James Z. Wang. 2015. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 990–998. https://doi.org/10.1109/ICCV.2015.119

[17] Wei Luo, Xiaogang Wang, and Xiaoou Tang. 2011. Content-based photo quality assessment. In *2011 International Conference on Computer Vision*. IEEE, 2206–2213.

[18] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4535–4544.

[19] 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*. IEEE, 2408–2415.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[21] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A unified framework for shot type classification based on subject centric lens. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 17–34.

[22] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. 2016. Vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) *(MM '16)*. Association for Computing Machinery, New York, NY, USA, 1183–1186. https://doi.org/10.1145/2964284.2973797

[23] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*. 166–171. https://doi.org/10.1109/ICDCSW.2011.20

[24] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 843–852.

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

[26] Marc Willhaus and Thomas Steinmann. 2022. Docker Compose SSL Service. https://github.com/mediatechnologycenter/neural-network-dockerfile.

[27] Marc Willhaus and Thomas Steinmann. 2022. Neural Network Dockerfile. https://github.com/mediatechnologycenter/neural-network-dockerfile.

[28] Munan Xu, Jia-Xing Zhong, Yurui Ren, Shan Liu, and Ge Li. 2020. Context-aware attention network for predicting image aesthetic subjectivity. In *Proceedings of the 28th ACM International Conference on Multimedia*. 798–806.