SmartEye: Assisting Instant Photo Taking via Integrating User Preference with Deep View Proposal Network

Shuai Ma[1,2]*, Zijun Wei[3]*, Feng Tian[1,2]†, Xiangmin Fan[1,2], Jianming Zhang[4], Xiaohui Shen[5], Zhe Lin[4], Jin Huang[1,2], Radomír Měch[4], Dimitris Samaras[3], Hongan Wang[1,2]

[1] State Key Laboratory of Computer Science and Beijing Key Lab of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences.

[2] School of Computer Science and Technology, University of Chinese Academy of Sciences.

[3] Stony Brook University. [4] Adobe Research. [5] ByteDance AI Lab.

mashuai171@mails.ucas.ac.cn, {zijwei, samaras}@cs.stonybrook.edu, shenxiaohui@gmail.com, {tianfeng,xiangmin,huangjin,hongan}@iscas.ac.cn, {jianmzha,zlin,rmech}@adobe.com

*: Both authors have contributed equally to this work; †: Corresponding author

ABSTRACT

Instant photo taking and sharing has become one of the most popular forms of social networking. However, taking high-quality photos is difficult as it requires knowledge and skill in photography that most non-expert users lack. In this paper we present *SmartEye*, a novel mobile system to help users take photos with good compositions in-situ. The back-end of *SmartEye* integrates the View Proposal Network (VPN), a deep learning based model that outputs composition suggestions in real time, and a novel, interactively updated module (*P-Module*) that adjusts the VPN outputs to account for personalized composition preferences. We also design a novel interface with functions at the front-end to enable real-time and informative interactions for photo taking. We conduct two user studies to investigate *SmartEye* qualitatively and quantitatively. Results show that *SmartEye* effectively models and predicts personalized composition preferences, provides instant high-quality compositions in-situ, and outperforms the non-personalized systems significantly.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Interactive systems and tools*;

KEYWORDS

Photo composition, user preference modeling, interactive feedback, deep learning

ACM Reference Format:

Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, Hongan Wang. 2019. SmartEye: Assisting Instant Photo Taking via Integrating User Preference with Deep View Proposal Network. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4-9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA, Paper 471, 12 pages. https://doi.org/10.1145/3290605.3300701

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00

https://doi.org/10.1145/3290605.3300701

1 INTRODUCTION

With the proliferation of photo-taking smartphones, personal photography has become a significant part of contemporary life: people engage in capturing every memorable moment in their lives [37, 50] and sharing these photographs through social networks such as Instagram, Snapchat, and others. However, although smartphone cameras have expanded the accessibility of photography to general users by making photography simple, cheap, and ubiquitous, taking high-quality and visually aesthetic photos is still quite challenging as it requires expertise in photography such as deciding the distance to target objects, their poses and the overall scene compositions. Non-expert photographers may have difficulty adjusting their compositions to achieve visually pleasing photographs.

One way to help non-expert users to get high-quality photos is to apply intelligent photo composition algorithms to photos for offline post-processing. Researchers have proposed various methods for auto-composition [11, 20, 26, 30, 33, 45, 57]. These algorithms generally require users to take a photo first, save it and then feed it into a re-composition pipeline. Compared to the in-situ pipeline proposed in this

(a) (b)

Figure 1: (a) Given an input photo on the upper left, the View Proposal Network (VPN) suggests a set of diversified well-composed views (as shown on the upper right); the *P-Module* adjusts the suggestions (as shown on the bottom) based on the learned user preferences in real-time. (b) *SmartEye* learns user preferences interactively and progressively: as the user selects their favorite composition at the bottom of the screen, the *P-Module* gets updated. Thus *SmartEye* will suggest more personalized compositions the longer it is used.

work, there are several limitations associated with post-processing approaches: i), they take extra storage and time; ii), post-processing can only operate on already taken photos, hence potentially good compositions can be overlooked during capturing; iii), most of these methods ignore user-specific preferences in composition. Recent work [43] addresses such personalized preferences in an offline manner. However, it requires additional data collection and significant computational resources for offline model-training.

Hence, a successful in-situ composition suggestion system should address the following challenges: first, online methods should provide knowledgeable feedback in real-time to avoid the unpleasant experiences caused by lags in control [25]; second, interactions with the interface should be intuitive; third, since photo composition is well known to be highly subjective as individual users have very different visual preferences [6], the proposed model should take into account personalized composition preferences and model these preferences interactively and effectively.

To address these challenges, we propose *SmartEye*. The back-end of *SmartEye* is based on a View Proposal Network (VPN), a deep learning based model that outputs composition suggestions in real time. The VPN is general-purpose as it learns photo compositions from a large scale dataset with crowdsourced annotations. We propose to append an additional module (*P-Module*) to model personalized composition preferences. As opposed to the offline machine learning approaches [2, 5, 6, 22, 43], in our work, the *P-Module* is designed to model and predict the user's preferences interactively and progressively, following a standard interactive machine learning paradigm [16]. We then combine the VPN scores and the *P-Module* scores using a memory-based weighting algorithm [13]. As *SmartEye* captures each user's preferences, it offers user-specific composition recommendations (Fig. 1). The front-end of *SmartEye* is a novel interface which enables a set of useful functions (Fig. 2) named smart-viewfinder, smart-zoom and smart-score. With the interface, the users can interact with the back-end algorithm effectively and intuitively.

We have conducted two user studies to investigate the overall performance of *SmartEye* and its effectiveness in modeling and predicting personalized composition preferences. Quantitative results reveal that photos taken with *SmartEye* have significantly higher agreement with the compositions selected by the user, compared to the tested real-time in-situ photo composition systems. Qualitative results show that *SmartEye* effectively models personalized preferences and that users enjoy interacting with it.

To the best of our knowledge, this is the first work that incorporates personalized preference modeling into a mobile, in-situ, user assistance system for photo-capture. The contributions of this work are: 1) We propose *SmartEye*, a novel in-situ mobile system: the back-end integrates a deep composition suggestion model (VPN) and a user preference module (*P-Module*) to provide personalized compositions; the front-end presents a novel interface with multiple functions that allow users to interact with the system effectively and intuitively. 2) We show in qualitative and quantitative results from user studies that *SmartEye* outperforms the examined baselines significantly both in the qualities of the suggested compositions and the levels of user-satisfaction. The surveys on participants additionally provide valuable implications and recommendations for design systems on other subjective tasks.

2 RELATED WORK

Photo Composition Suggestion: Algorithms

There are two major types of photo composition algorithms: rule-based and learning-based. Rule-based models generally utilize empirical rules such as rule of thirds [28, 54], or symmetry [8]. However, photo composition is a complex task. Only applying pre-defined rules might lead to sub-optimal results. Learning-based algorithms learn composition from data of different types [4, 17, 54] using various models [38, 55].

Recently the learning based models using deep neural networks have achieved promising results. Many deep learning methods have been proposed to find views with good compositions for various applications, such as image cropping [8, 9, 17, 21, 26, 30, 31, 54, 55], image thumbnail generation [15] and view recommendation [7, 10, 48]. Despite the improved performance, these deep learning models are based on inefficient sliding window paradigm to evaluate candidates, thus very inefficient. Two recent works [15, 51] adopt fast object detection frameworks [12, 44] for image thumbnailing and image cropping for efficient inference. However, running at $\leq$ 10 FPS on the server side, these models are still not fast enough to enable real-time suggestion generation. A recent work [52] has proposed a View Proposal Network (VPN) which can process up to 75 FPS. We base *SmartEye* on the VPN. However, the VPN, like most machine learning based models, learns general composition knowledge that is agnostic to specific users from a large scale image composition dataset annotated on Amazon Mechanical Turk. As view suggestion is a very subjective task, such general knowledge may not be enough.

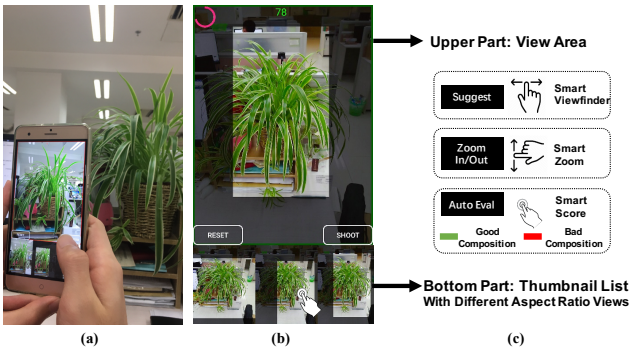

Figure 2: The interface of *SmartEye*. (a): *SmartEye* in action; (b) overview of the interface, (c) touch gestures to activate the Smart Viewfinder, Smart Zoom and Smart Score

Photo Composition Suggestion: Interface Design

Multiple interfaces have been proposed for good user experiences in photo-compositions or modeling user preferences. Element-level interactions [3] are used to capture user preference on the placement of a certain visual element. Besides interface interaction, modern devices, such as eye trackers, facilitate the capture of user intents regarding which elements are to be retained in the cropping [45, 56]. However, these methods either require an extra input device such as an eye tracker [45], or they consider photos with only a single isolated object [3].

Perhaps the most related work to ours is [53] that uses real-time on-screen guidance to help users take better photos with mobile devices. However, there are three prominent distinctions: First, the back-end algorithm of [53] only considers one single rule (i.e. rule-of-thirds). In comparison, our back-end algorithm (i.e. VPN) is data-driven and is trained with a large-scale human-annotated dataset (over 1 million view pair annotations [52]), which may cover a richer set of rules for composition. Second, we propose a user preference modeling module to allow adaptation and personalization, which is important in photo composition tasks that highly depend on subjective aesthetic judgment. Third, the interactions are different: [53] provides guidance on which direction users should move, our interface enables users to explore and compare multiple compositions suggested by the system. We also propose additional features such as Smart Zoom and Smart Score to help users better explore the scenes.

User Preference Modeling

Methods for learning aesthetic preference have been presented in various design domains [18, 39, 40, 42, 46], such as web pages [42] and color themes [39]. Assessment of photo preference has been extensively investigated in [14, 23, 32, 34]. Kim et al. exam the subject's preference in photo sharing activities [24]. Pallas et al. introduce a set of four elementary privacy preferences a photo subject can have [41].

In the area of recommendation, a lot of preference learning methods have been proposed. Massimo et al. propose a preference learning model that takes into account the sequential nature of item consumption [35]. Unger et al. present a method for inferring contextual user preferences by using an unsupervised deep learning technique applied to mobile sensor data [49]. Solomon proposes customizable recommending systems allowing users to directly manipulate the system's algorithm in order to help it match those preferences [47].

There are also works on photo reorganization that adapt the concept of modeling user preference. Liang et al. allow user interactions to drag and drop the photo hierarchy suggested by the system to model preferences [27]. Yuki Koyama et al. collect user's previous choices for color enhancement [19].

3 DESIGN OF SMARTEYE

In this section we first describe the back-end algorithms adopted by *SmartEye*. We then describe its front-end interface to enable users interacting with the back-end algorithms effectively and intuitively.

Back-end Algorithms

The back-end of *SmartEye* has two core modules: the View Proposal Net (VPN) model [52] that suggests initial compositions in real-time and the preference module (*P-Module*) that refines these compositions based on individual preferences learned from the user's previous selections.

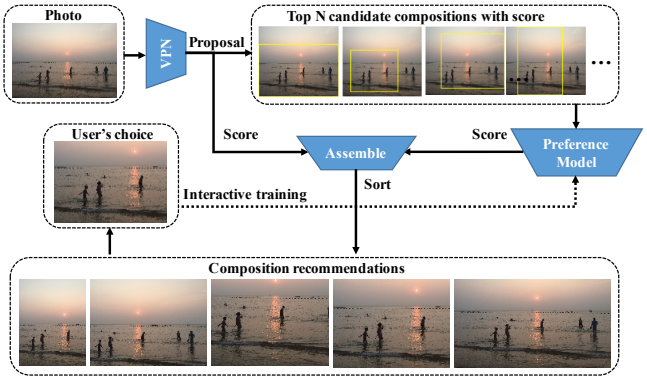

Figure 3: Overview of the back-end algorithm. The *P-Module* adjusts the suggestions provided by the VPN based on learned user preferences, while the user's selections are used in turn to update the *P-Module*.

Fig. 3 shows the pipeline of the back-end algorithms. Given a photo, the back-end algorithms work as follows:

(1) the VPN suggests initial compositions;
(2) the *P-Module* computes personalized scores of the compositions suggested by VPN ;
(3) the adjusted score for each composition is computed using Eq. 2. The compositions are sorted by their adjusted scores and displayed to the user;
(4) the user selects compositions from the displayed candidates. The selected compositions will serve as positive samples to update the *P-Module*.

View Proposal Network (VPN). The VPN [52] is a real-time Convolutional Neural Network based on anchor boxes (pre-defined bounding boxes), inspired by recent successful object detection frameworks [29, 44] in computer vision.

An illustration of the VPN is shown in Fig. 4. The VPN first pre-defines a set of anchor boxes that covers the most possible universal candidate compositions. Then it takes a photo as input and outputs a score list of the predefined anchor boxes. Based on the score list, the highest scored anchor boxes (compositions) will be suggested to the users.

More specifically, the back-bone network of VPN is based on the Single-Shot-Detector [29] truncated after the ninth convolutional module. On top of the SSD there is a convolutional layer, an average pooling layer and a fully connected layer that outputs an N-Dimensional vector S^{VPN} corresponding to the N anchor boxes. In our work we directly use the predefined 895 candidates from [52], in which the candidates are selected to cover the most common aspect ratios, sizes and regions over the photos. VPN is trained to predict scores for all the anchor boxes: for any anchor box x_i that is a good composition of the photo, the VPN will output a high score at S_i^{VPN} corresponding to that anchor box and vice versa.

The VPN can process photos up to 75 FPS [52]. We set the refresh rate to be 15 FPS for *SmartEye* after an iterative design process, considering factors such as network latency, computational cost in smartphones, and smoothness in user experience.

Details on data collection and training of the VPN are provided in [52].

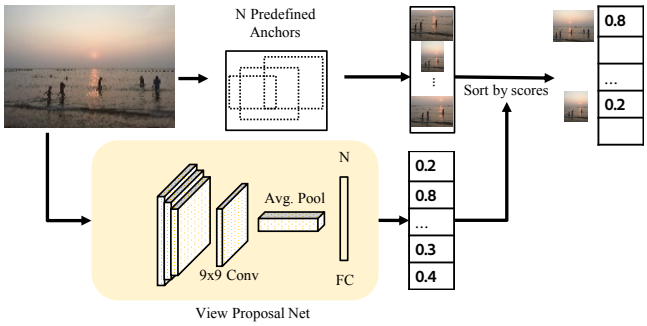

Figure 4: An illustration of the View Proposal Network (VPN). The VPN pre-defines a set of anchor boxes (bounding boxes) for all photos. It takes a photo as input and outputs a score list of the predefined anchor boxes in real-time.

Personalized Preference Model (P-Module). The VPN is trained with millions of image compositions, annotated via Amazon Mechanical Turk. Hence, the suggestions that the VPN outputs are general in that they are agnostic to specific user preferences. However, photo composition is a subjective task and varies among different users. General suggestions are not enough. One way to address this is to fine-tune the VPN for each individual with her selections. However, it is impractical to either collect large scale user preferences from each user or train an individual VPN for each user. Additionally, fine-tuning deep neural networks like the VPN with small training data is a challenging problem.

We expect the *P-Module* to be: i) able to accurately model user preferences and robust to noises; ii) efficient both in terms of model update and score prediction.

In this work we implement *P-Module* with logistic regression: given a composition $\mathbf{x}_i$ from a photo I, we extract its feature $\varphi(\mathbf{x}_i, I)$ and feed it into a logistic regression model to get a scalar user-preference score.

$$S_i^P = \frac{1}{1 + e^{\mathbf{w}\varphi(\mathbf{x}_i, I)+b}} \quad (1)$$

where $\mathbf{w}$ and b are learnable parameters for each user.

The simplicity of logistic regression makes the *P-Module* robust to noise, easy to be updated online interactively, and predict the scores efficiently.

$\varphi(\mathbf{x}_i, I)$ is a 32-Dimension feature vector describing the composition $\mathbf{x}_i$ that resides in photo I. $\varphi(\cdot,\cdot)$ is constructed based previous works [1] and feedback from the user-study (Sec. 4).

The components of the 32D feature can be classified into the following categories: 1) geometry based (10D): the geometrical center (coordinates of the center-point), size in pixels, relative size to the original photo, aspect ratio and offset to left/right/top/bottom boundaries of the original photo; 2) saliency based (10D): for each composition we employ [36], a widely used saliency detection algorithm to find its salient-region (Fig. 5). We then compute the geometrical properties of the salient-region inside the composition: coordinates of the center-point, lengths of major/minor radius [27], the offset to left/right/top/bottom boundary of the composition and relative size to the composition; 3) composition-rule based (8D): we designed features to cover the distances from the center of the salient-region inside of a composition to the four split-lines intersections and to the four golden-ratio lines; 4) photograph based (4D): we compute brightness, contrast, saturation, and color balance values for each of the composition.

While we demonstrate the effectiveness and the efficiency of the *P-Module* using the proposed features in Sec. 4, we make no claim that these features are canonical. We expect that better or worse behavior would be available using different sets of features. It remains an interesting open question to construct feature sets that offer very good performance for our task and can be computed efficiently.

Figure 5: Visualizations of the salient regions in photos detected by [36].

Integrating VPN with P-Module. We combine the VPN scores and the *P-Module* scores following the memory-based algorithms [13] that predict ratings for users based on their past ratings.

More concretely, we weighted combine the VPN score S_i^{VPN} and *P-Module* score S_i^P of $\mathbf{x}_i$ as following:

$$S_i = (1 - c_i)S_i^{VPN} + c_i S_i^P \quad (2)$$

where $c_i \in [0, 1]$ is a confidence score describing how similar $\mathbf{x}_i$ to previous user selected compositions:

$$c_i = \frac{1}{\left(1 + min_j d(\varrho(\mathbf{x}_i), \varrho(\mathbf{x}_j))\right)^\beta} \quad (3)$$

where $d(\cdot,\cdot)$ is the Euclidean distance between two compositions in feature space $\varrho(\cdot)$. ϱ is similar to φ except that the geometrical features were not used in ϱ since in this case we compare compositions independent of the photos they reside in. Fig. 6 shows that similar compositions have shorter distances in the feature space $\varphi(\cdot)$. β is a parameter controlling the change rate of c_i. We fix $\beta = 2$ throughout this work.

The confidence value is close to 1.0 if a previous composition is similar to the current, otherwise close to 0.

As finding the nearest neighbors for each new composition $\mathbf{x}_i$ only involves computing Euclidean distance between $\mathbf{x}_i$ and the pre-computed features of previously selected compositions, the integration of VPN and *P-Module* is in real-time.

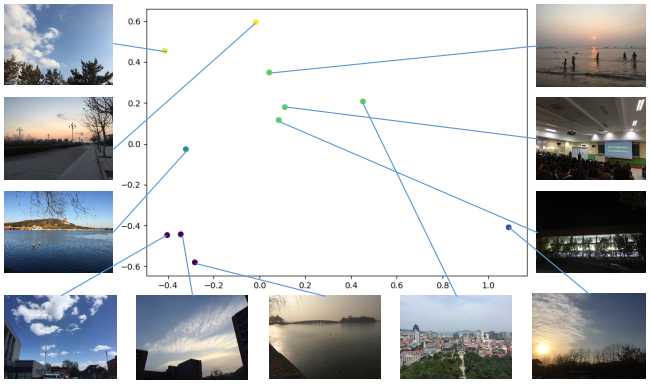

Figure 6: A visualization of a few photos after reducing their features $\varrho(\cdot)$ to 2D using Principle Component Analysis (PCA), illustrating that photos of similar content are closer in feature space.

Interactively Updating the P-module. The *P-Module* learns from user's previous selections on compositions and updates itself following a standard interactive machine learning pipeline [16].

More specifically, selecting a composition is a natural *labeling* procedure: the selected composition is regarded as a positive sample while the others are regarded as negative. Each time there are new samples added, the *P-Module* is retrained with all the positive samples and a set of randomly selected negative samples. As the *P-Module* is essentially a

logistic regression, the computation is efficient. The updating frequency of *P-Module* can be every photo or every day, depending on the user preferences.

Front-end Interfaces

Fig. 2 shows the user interface of *SmartEye*. The view area on the top is like a normal view in a photo-taking application but extends with the following functions: i) as the user adjusts the camera, it will show the composition suggested by the Smart Viewfinder (described below) in real time; ii) it can display an enlarged view of a composition in the thumbnail list in the bottom if the user wants to adjust based on that composition. iii) it functions as a preview window after user selects a composition candidate.

The thumbnail list on the bottom of the interface shows the top suggestions from various aspect ratios. The thumbnail list area can be swiped left or right to accommodate various number of composition candidates.

The viewfinder and recommended thumbnail list are designed to be separate from each other. *i.e.*, system recommendations of scenes in the view area are not constrained within the active viewfinder region even if the user selects a view from the thumbnail list. Switching between viewfinders can allow users to freely explore and compare different recommendations. At the same time, making them independent avoids the users being "trapped" by their previous selections.

SmartEye also provides several support functions:

Smart Viewfinder. The Smart Viewfinder provides real-time composition suggestions for the scenes captured by the camera, displays the top composition in the view area, and the other recommendations of different aspect ratios in the thumbnail list.

Smart Score. Smart Score displays a real-time dynamic score on top of the view area indicating the quality of the current composition shown in the view area. If the composition is of high quality, the color of the frame is green, otherwise red. This feature helps users to know the quality of the compositions and encourage them to adjust the cameras accordingly to get better scores.

Smart Zoom. Smart Zoom is an intelligent zoom function that allows the user to zoom automatically to a scale of the current photo for the best composition. It make the zoom in/out easier in that the users do not have to adjust the lens back and forth for the best scale. Smart Zoom regulates the compositions suggested by the back-end algorithm with additional scale information.

In addition, *SmartEye* provides several complementary functions to further improve user experience: (1) Confidence Score: in addition to Smart Score, the interface optionally displays confidence score (*i.e.*, c) indicating how much this recommendation is based on user preferences; (2) Customized Thumbnail List: users can customize the number of suggestions in the thumbnail list; (3) Free Cutting: in addition to selecting from the suggested compositions, *SmartEye* also supports *re-compose*, *i.e.*, it also allows the user to manually adjust the system's suggestions. The user-created crop is served as a positive training example as well.

SmartEye also provides some gestures to active the support functions. Figure 2 (c) shows the gestures activating the support functions: users can swipe left and right to switch to compositions of different aspect ratios, swipe up and down for Smart Zoom, and long press the screen to activate Smart Score.

SmartEye works in both portrait and landscape mode. The recommendations will reside on the rightmost side of the screen in landscape mode.

4 USER STUDY

We conducted two user studies to investigate: 1) the important components (features) for personalized preference modeling, 2) whether the *P-Module* helps to improve model's agreement to the user's selections; and 3) the user experience of taking photos with *SmartEye*.

Participants

The user study was conducted on 16 participants with various photograph skills. Among the participants, five (P9-P13) were members of college photography association, three (P14-P16) were graduate students majoring in arts, and the rest did not report any expertise in photography. On average, the participants had 4.13 years of experience in photograph.

Study 1: Effectiveness of *P-Module*

- *Task 1: Select good compositions from VPN suggestions*

We first of all randomly collected a dataset, *PhotoSetA*, that consists 50 photos to cover various types of daily photos people normally take. Then, we asked every participant to contribute 100 photos that they took to form *PhotoSetB* (including 1600 photos taken by 16 participants). The two datasets are diverse in terms of of photo contents, styles and aspect ratios.

We processed all photos in *PhotoSetA* and *PhotoSetB* with VPN, and each photo got 5 recommended compositions. For each participant, we gave him/her 150 photos (50 from *PhotoSetA* and 100 from *PhotoSetB* taken by himself) to score, and we collected 12000 (16 participants $\times$ 150 photos $\times$ 5 compositions) photos with subjective scores. We also asked each participants to fill a questionnaire and interviewed some of them. For some of the photos that participants had just composed, we asked them why they selected such composition and what factors were considered, and got a lot of valuable insights.

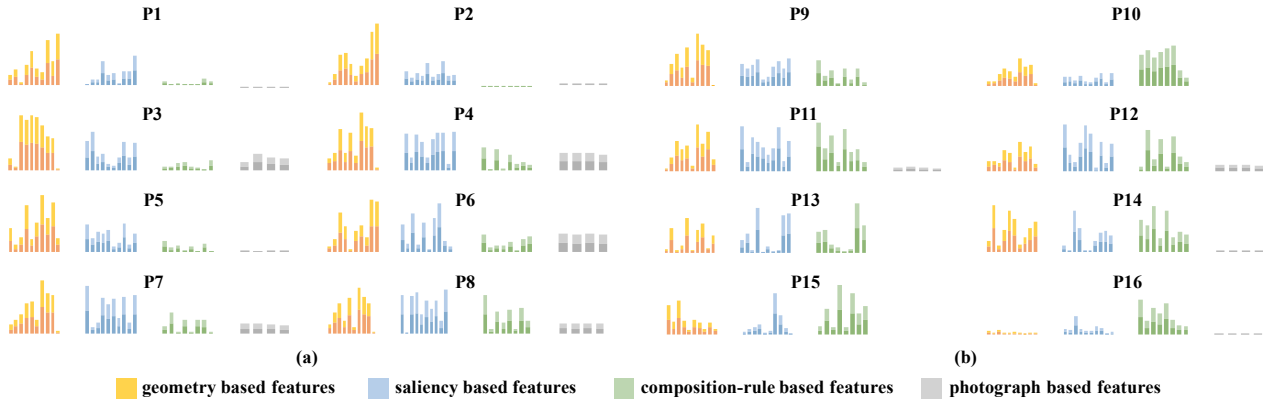

Figure 7: Feature correlation analysis of participant preferences. The upper part of each histogram column (light color) represents the Pearson correlation coefficient and the lower part (dark color) represents the Spearman correlation coefficient. Most features are significantly ($p < 0.05$) correlated with user preferences. Overall the features proposed correlate well to user preferences. Also note that the preferences of different participants correlate differently with the features proposed, showing the variations in user preferences.

We used the annotations from 16 participants to train 16 *P-Modules* for everyone for *Task 2*.

- *Task 2: Rate composition proposed by different models*

In this task, we want to investigate if the *P-Module* is helpful.

We demonstrate the quality of compositions suggested by *SmartEye* through a challenging user experiment: we additionally collected 50 photos of various styles and contents. For each photo, we picked the top 5 views generated by different models and let participants select the best (top 1).

The models we experimented are as follows: 1) the proposed *SmartEye* with the *P-Module*, denoted as *P-Module*. The *P-Module* is trained using the data collected from *Task 1*. 2) the *SmartEye* with VPN only, denote as VPN and 3) a baseline composition suggestion algorithm based saliency and detected faces following [17], denoted as Sal+Face. Sal+Face works as follows: given a photo, Sal+Face computes its saliency map and detects the faces, it then suggests the candidate compositions with top scores in terms of average saliency and face in them.

We mixed the outputs from different models and displayed them to 16 participants. We asked them to select their favorite composition in each photo.

Study 2: Usability of *SmartEye*

- *Task 3: Compose photos on mobile devices with different systems*

We deployed the following systems on Android devices: 1) the proposed *SmartEye* with the *P-Module*, 2) *SmartEye* with VPN only, 3) Sal+Face. We additionally include the Android native photo-taking application as a reference for non-composition-recommendation systems.

We instructed participants how to use the systems and encouraged them to play with all the functions before starting this task. We alternated the order participants using different systems to counterbalance order effects. Participants were asked to take at least 30 photos using each system. Then they were required to fill a post-task questionnaire. This post-task questionnaire contains questions about the tested approaches for auto composition, effects of preference modeling and the support functions available in *SmartEye*.

- *Task 4: Use SmartEye to take photos for a month*

In this task, we let each participant use *SmartEye* continuously for a month. Each participant is required to take at least 5 photos every day with *SmartEye*. The contents and styles of the photos taken were not limited. It means that users could use *SmartEye* arbitrarily, as long as they take 5 photos everyday.

At the end of the month, we accessed selected compositions and investigated how *P-Module* progresses over time scale.

5 RESULTS

Quantitative Results

Study 1. Based on the 16 participants' selections in *Task 1*, We computed the Spearman and Pearson correlation coefficients between the users score and each of the 32D features. The correlations were shown in Fig. 7. Note that the correlation varies for different participants. Almost every participant pays attention to geometry based features and saliency based features. It's also interesting that the skilled users seem to

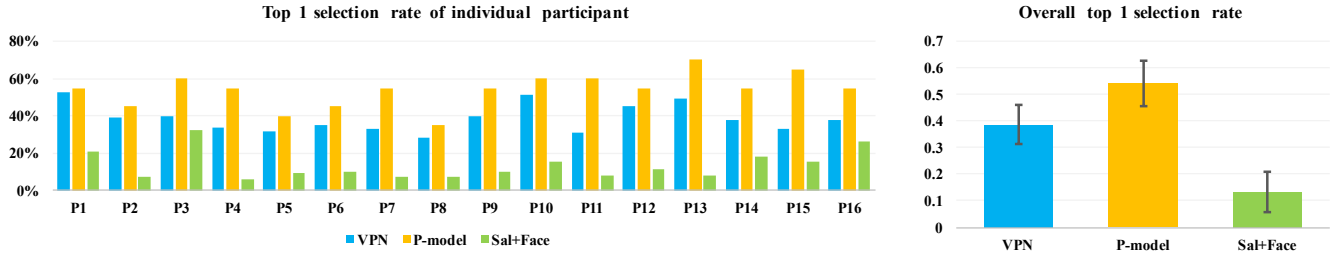

Figure 8: The comparison of VPN, *P-module* and Sal+Face in the top 1 suggestion selection rate for each user. The *P-Module* outperforms the VPN and Sal+Face algorithms by a significant margin. The gap between the *P-Module* and the VPN demonstrates the benefits of modeling user preferences.

focus more on saliency and composition-rule based features, while the others may rely more on geometry and photograph based features. The difference in feature correlations also reflects the difference in composition preferences among participants.

We additionally evaluate the rate that the top 1 composition a system suggests is the favourite composition from a user. We denote this measure the *top 1 selection rate*. Fig. 8 shows the comparison of VPN, *P-Module* and Sal+Face based on the the participants data collected in *Task 2*. We can see that *P-Module* outperforms VPN on each user's composition data and overall it outperforms the other baselines by a large margin. Based on pair-wise t-tests we found the result are significant: comparing VPN with *P-Module*, the T value is -7.229 and $p < .001$; comparing VPN with Sal+Face, the T value is 11.597 and $p < .001$; comparing *P-Module* with Sal+Face, the T value is 16.318 and $p < .001$. We also computed standard deviation values of three models and show in Fig. 8.

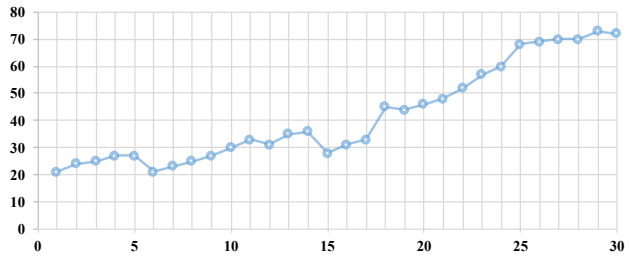

Figure 9: The progressive change of top 1 selection rate of *SmartEye* in 30 days.

Study 2. We access the participants selections for *Task 4* and show their average top 1 selection rate on a daily basis in Fig. 9. Note that overall the top 1 selection rate of *SmartEye* with *P-Module* increases progressively over the time scale. It shows that *SmartEye* is able to model user preferences and refines itself with increasing user selections used. Around

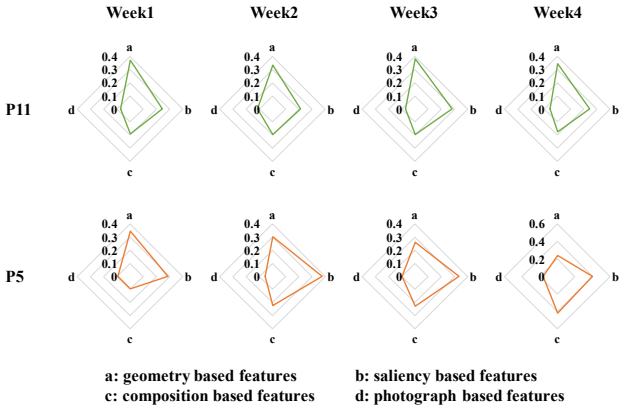

Figure 10: The progressive change of two participants' preferences in four weeks. It shows that some users may have changed their preferences over time.

day 6 and 15, there were slight dips of performance. We speculate that these dips may due to the inconsistency of user preference over time. Photo composition closely involves subjective judgment, local dips in certain time spans could happen. Investigating the longer-term effect can be an interesting future work.

Additionally, Fig 10 illustrates the change of the correlations between feature of different types and two users' (P11 and P5) preferences at each week over the 30 day time scale. It confirms with our *Task 1* that different users may rely on different factors for composition. It also shows that some users may have different preferences over time. In the way it demonstrates the necessity of modeling user preferences for subjective tasks.

Feedback from Interviews and Questionnaires

Study 1. We select participants feedback from *Task 1*:

1. Different people like different aspect ratios. P3 said, *"It (aspect ratio) is related to my usual photography habits. I prefer using a 16:9 aspect ratio on my mobile phone."* P9 mentioned,

"I prefer square photos or 4:3 aspect ratio, and photos can't be too narrow or too long."

2. Position of border of cropped region is a key factor. P7 said, *"I realized I always put object too close to the left border, so too many unnecessary objects on the right side of the photo need to be cut off."*

3. The range of shots of photos taken by different people is also different. P11 said, *"I found that I prefer to adjust the lens range too wide, which causes important objects being too inconspicuous in the photo."*

4. Users have different views on the size of the retained after cropping. P5 said, *"I don't like the photos to be cropped too much.".* While P11 said, *"I prefer cutting more off. I always shoot unnecessary things. I didn't realize I was in the habit until the algorithm suggested better composition to me."*

5. The position of main object influences users' choice. Almost everyone paid attention to the location of the main content. P11 mentioned, *"When I took a photo of a dog, the position of the dog in the composition was crucial. It couldn't be too close to the side, preferably in the middle."*

6. The proportion of salient regions is important. P3 said, *"The proportion of objects I care about in the photo is very important. I like to fill more than half of the photo with objects I care about."* P15 was just the opposite of him on this point. P9 mentioned, *" In addition to preserving the object I want to shoot, I want to preserve some background. But I can't tell exactly how much background I want to preserve."*

7. Salient regions should not be truncated. Several people have mentioned the factor of the integrity of objects. P2 said, *"There can't be an incomplete object in a photo, and if a person is cut into half, I will not choose this composition, despite that it performs well in other factors."*

8.Some photographers intend take into account the composition guidelines. P9 commented, *"the rule of thirds is one of the most common techniques I use when taking photo"* P10 also pays more attention to the rule of thirds. P11 and P13 mentioned, *"When I take photos, I want the objects to form the Golden Ratio, especially when I take landscape photos."*

9. The brightness and color of the photo should be considered. P6 said, *"In a landscape photo, I usually like the bright part, and those areas which are dark should be cut off."*

In addition to these factors, the participants also give many useful suggestions.

P13 said, *"When taking photos, the depth of the photo often plays an extraordinary effect because the sense of depth is very important."* Similarly, P14 who is also a skilled photographer said, *"Triangulation makes the photo more solid."*

We also found some interesting answers. P11 mentioned, *"I think composition depends on the content of the photo. When I shoot the sea, I often want to express a broad prospect, so I hope the photo to be horizontal, which is better to have 16:9 aspect ratio."* P8 said, *"When I'm shooting a building, I like to let the lines of the building appear very regularly horizontal or vertical in the photo, but I don't think so when I'm shooting other types of photographs."*

Study 2. Some of the interesting feedback in the user study 2 is provided in the following

Auto-Composition is very appealing. P8 said, *"(SmartEye) gave me a sense of continuity"* between shooting step and composition step. P5 said, *"The five recommended compositions are sometimes so satisfying that I can't even make a choice."* P6 agreed with him, *"(It is) just like a professional photographer trimming my photos."* P9 said that as she used the system, she more and more trusted *SmartEye* because it always gave her an expected composition recommendation. P11 said, *"When I saw that the first recommendation for composition was the original photo I took, I felt like a professional photographer praising my work. It was inspiring and made me more eager to use this system."* P13 commented that he *"used it (as a starting point) for further exploration."*

The system learned user preferences successfully. P5 commented that *"I felt that my preference and tendencies were gradually learned, which was amazing."* P7 said, *"SmartEye reflected my tendency of cropping the bottom part of the photo,"* and thus she realized that *"the system actually learned my preference.".* P16 said, *"What surprised me was that the system really recommended a composition that met my habit."* P4 commented that, compared to *SmartEye*, *"(Baseline) could not reflect my intent"* and *"only provided a safe composition."* P3 said that, typically when using Baseline, *"I rarely used the auto-composition"* because *"usually it did not fit my preference."* On the other hand, she agreed that she *"would like to use the auto-composition in SmartEye as it learned my preference."*

Visualization of confidence value is useful. P2 said, *"This function evoked the feeling of collaborating with another me."* P4 mentioned that, in *SmartEye*, *"there was interaction with the system"* through the support functions, thus *"executing the task (with SmartEye) was fun."* P7 commented, *"This (visualization) was helpful (in making decisions) when I had little confidence in myself."*

Real-time recommendation helps a lot. Participants all agreed that real-time recommendation was useful. P6 liked it because *"It made me enjoy taking photos. I didn't really need to press the SHOOT button to know what the composition will look like because it allowed me to adjust the camera lens in real-time."* P7 also liked it because *"it made me feel very comfortable and natural. The whole shooting process was dynamic. I didn't have to worry about taking an unsatisfactory photo because I knew how the system would optimize it before I pressed the SHOOT button."*

Smart Score is a very convenient feature. P5 said, *"It's my favorite feature, and when I moved my phone, I got the composition score of the current lens, which made me more confident*

in my photo." P6 commented, *"this feature made it easier for me to press the SHOOT button at will, and I usually moved the lens to explore composition with the highest score and then shot it."* P10's answer was fancy, *"this function was like that there was a teacher helping me take photos, and he would constantly give me the advise for current composition, prompting me to adjust the lens angle."* P16 also mentioned, *"I felt like I was getting better at taking photos with this feature. I was used to trying my best to get a higher score, which slowly helped me learn how to take a high-quality photo."*

Preliminary and Post-Task Questionnaires

Our preliminary and post questionnaires were based on a 5-point scale, where 5 corresponds to *strongly agree* and 1 corresponds to *strongly disagree*. Q1-Q8 in Fig. 11 verified the usefulness of automatic composition and personalized composition recommendation algorithm. Q9-Q20 show the results of the post-task questionnaire with respect to the auto composition approach with preference learning and the user support functions. Overall, participants gave positive scores.

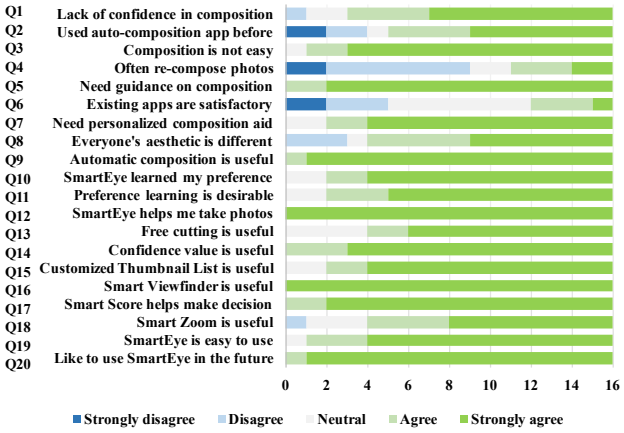

Figure 11: Results of the preliminary (Q1-Q8) and post-task (Q9-Q20) questionnaires. We used a 5-point Likert scale.

6 DISCUSSION

Implications

Several lessons were learned to further improve user experience for photo composition system with personalized preference modeling. We believe that these lessons are also applicable to other systems that attempt to incorporate personalized preferences for subjective tasks.

Modeling personalized preference is important for subjective tasks. According to the interview, we found that participants enjoy the benefit from the procedure that the system learns their habit and preference from history data.

It is preferable to show users how the system makes suggestions, rather than to make the system a "black box." In our interview, we found that the score from Smart Score as well as the confidence value received a lot of positive feedback; it helped users with the composition task, and also made the system more transparent and reliable. Showing "why the system thinks so" in more details is a possible future direction in this respect.

Future Work

Advance the general composition suggestion models. Our work is based on the VPN and improves the user experience by integrating the *P-Module*. Note that the VPN is not the perfect algorithm for general composition suggestions. With better suggestions models, *SmartEye* can further advance its performance.

Extending to multiple users using collaborative filtering. The *P-Module* in our work for preference learning is updated for a single user; thus the suggestion is based on only his/her own composition history. It would also be interesting to develop algorithms and interaction techniques to share the learning results among many users and utilize them in a collaborative manner [6, 19].

Suggesting moving directions. Suggesting moving directions seems on the surface a direct extension of our algorithm but we found it quite challenging in practice: first, since one image may have multiple good suggestions, it may hurt user experience when a composition the system suggested to move to is not the one the user intended; second, the system will have to track, smooth and record the history of movements in order to predict the next direction; third, it is more time critical that the directional suggestions feel smooth. Addressing it could also be interesting future work.

Explaining more about the decisions made by the model The VPN is a data driven model that learns directly from human data. Even though the data was collected intentionally for composition, it is difficult to fully disentangle composition from other aspects such as lighting and focus. By observing the outputs, we speculate that the VPN has already implicitly accounted for these aspects. But it is hard to explicitly show which aspect contributes to what extent, in the output for a data driven model. To explicitly model other aspects, we can append modules after the output of our model that are specific to these aspects. Many off-the-shelf models for these aspects have achieved decent performance.

7 CONCLUSION

We have investigated the concept of user preference modeling in photo composition, *i.e.*, the system progressively and interactively learns user preferences on photo compositions. Meanwhile, we have verified that in photo composition tasks, preferences are different among different users and

even each individual's preferences may change over time, which demonstrates the necessity of applying the *P-Module* to current systems. Furthermore, we have integrated the *P-Module* and the VPN into an interactive in-situ mobile system, *SmartEye*, with a novel interface and a set of useful functions such as the real-time Smart Viewfinder, Smart Score and Smart Zoom. Our user studies have demonstrated the effectiveness of *SmartEye*: we have shown that *SmartEye* is preferable to the tested baselines, the support functions were helpful, and participants were overall satisfied with *SmartEye*.

Acknowledgement

We thank the reviewers for their constructive feedback. This work was supported by National Key R&D Program of China (Grant No. 2016YFB1001405), National Natural Science Foundation of China (Grant No. 61872349), Key Research Program of Frontier Sciences, CAS (Grant No. QYZDY-SSW-JSC041), CAS Pioneer Hundred Talents Program, NSF-CNS-1718014, NSF-IIS-1763981, NSF-IIS-1566248, the Partner University Fund, the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe.

REFERENCES

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. 2003. Matching words and pictures. *Journal of machine learning research* 3, Feb (2003), 1107–1135.

[2] Floraine Berthouzoz, Wilmot Li, Mira Dontcheva, and Maneesh Agrawala. 2011. A Framework for content-adaptive photo manipulation macros:Application to face, landscape, and global manipulations. *Acm Transactions on Graphics* 30, 5 (2011), 1–14.

[3] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM International Conference on Multimedia*. 271–280.

[4] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2011. A holistic approach to aesthetic enhancement of photographs. *Acm Transactions on Multimedia Computing Communications & Applications* 7S, 1 (2011), 1–21.

[5] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 97–104.

[6] J. C. Caicedo and A. Kapoor. 2011. Collaborative personalization of image enhancement. In *Computer Vision and Pattern Recognition*. 249–256.

[7] Yuan Yang Chang and Hwann Tzong Chen. 2009. Finding good composition in panoramic scenes. In *IEEE International Conference on Computer Vision*. 2225–2231.

[8] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. 2016. Automatic Image Cropping: A Computational Complexity Study. In *Computer Vision and Pattern Recognition*. 507–515.

[9] Yi Ling Chen, Jan Klopp, Min Sun, Shao Yi Chien, and Kwan Liu Ma. 2017. Learning to Compose with Professional Photographs on the Web. (2017), 37–45.

[10] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. 2010. Learning to photograph. In *International Conference on Multimedia*. 291–300.

[11] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. 2007. Self-Adaptive Image Cropping for Small Displays. *IEEE Transactions on Consumer Electronics* 53, 4 (2007), 1622–1627.

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. (2016).

[13] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 271–280.

[14] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. *Lecture Notes in Computer Science* 3 (2006), 288–301.

[15] Seyed A. Esmaeili, Bharat Singh, and Larry S. Davis. 2017. Fast-At: Fast Automatic Thumbnail Generation Using Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4178–4186.

[16] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.

[17] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. 2014. Automatic Image Cropping using Visual Composition, Boundary Simplicity and Content Preservation Models. 18, 18 (2014), 1105–1108.

[18] Aaron Hertzmann, Aaron Hertzmann, and Aaron Hertzmann. 2015. DesignScape: Design with Interactive Layout Suggestions. In *ACM Conference on Human Factors in Computing Systems*. 1221–1224.

[19] Takeo Igarashi, Takeo Igarashi, and Takeo Igarashi. 2016. SelPh: Progressive Learning and Support of Manual Photo Color Enhancement. In *CHI Conference on Human Factors in Computing Systems*. 2520–2532.

[20] Laurent Itti, Christof Koch, and Ernst Niebur. 2002. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 20, 11 (2002), 1254–1259.

[21] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional Neural Networks for No-Reference Image Quality Assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1733–1740.

[22] Sing Bing Kang Sing Bing Kang, A Kapoor, and D Lischinski. 2010. Personalization of image enhancement. 23, 3 (2010), 1799–1806.

[23] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The Design of High-Level Features for Photo Quality Assessment. *Proc Cvpr* 1 (2006), 419–426.

[24] Auk Kim and Gahgene Gweon. 2014. Photo Sharing of the Subject, by the Owner, for the Viewer: Examining the Subject's Preference. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 975–978. https://doi.org/10.1145/2556288.2557247

[25] Christin Kohrs, Nicole Angenstein, and André Brechmann. 2016. Delays in human-computer interaction and their effects on brain activity. *PloS one* 11, 1 (2016), e0146250.

[26] 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.

[27] Yuan Liang, Xiting Wang, Song Hai Zhang, Shi Min Hu, and Shixia Liu. 2017. PhotoRecomposer: Interactive Photo Recomposition by Cropping. *IEEE Transactions on Visualization & Computer Graphics* PP, 99 (2017), 1–1.

[28] Ligang Liu, Renjie Chen, Wolf Lior, and CohenâĂŘOr Daniel. 2010. Optimizing Photo Composition. In *Comput. Graph. Forum*. 469–478.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*. 21–37.

[30] 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*. ACM, 457–466.

[31] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z. Wang. 2016. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *IEEE International Conference on Computer Vision*. 990–998.

[32] Yiwen Luo and Xiaoou Tang. 2008. Photo and Video Quality Evaluation: Focusing on the Subject. In *European Conference on Computer Vision*. 386–399.

[33] M. Ma and J. K. Guo. 2004. Automatic image cropping for mobile device with built-in camera. In *Consumer Communications and Networking Conference, 2004. CCNC 2004. First IEEE.* 710–711.

[34] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *International Conference on Computer Vision*. 1784–1791.

[35] David Massimo, Mehdi Elahi, and Francesco Ricci. 2017. Learning User Preferences by Observing User-Items Interactions in an IoT Augmented Space. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 35–40. https://doi.org/10.1145/3099023.3099070

[36] N. J. Mitra. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 37, 3 (2015), 569.

[37] Susan Murray. 2008. Digital images, photo-sharing, and our shifting notions of everyday aesthetics. *Journal of Visual Culture* 7, 2 (2008), 147–163.

[38] Bingbing Ni, Mengdi Xu, Bin Cheng, Meng Wang, Shuicheng Yan, and Qi Tian. 2013. Learning to Photograph: A Compositional Perspective. *IEEE Transactions on Multimedia* 15, 5 (2013), 1138–1151.

[39] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2011. Color compatibility from large datasets. *ACM Transactions on Graphics (TOG)* 30, 4 (2011), 1–12.

[40] Peter Odonovan, Aseem Agarwala, and Aaron Hertzmann. 2014. *Learning Layouts for Single-PageGraphic Designs*. IEEE Educational Activities Department. 1200–1213 pages.

[41] Frank Pallas, Max-Robert Ulbricht, Lorena Jaume-Palasí, and Ulrike Höppner. 2014. Offlinetags: A Novel Privacy Approach to Online Photo Sharing. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 2179–2184. https://doi.org/10.1145/2559206.2581195

[42] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z. Gajos. 2013. Predicting Users' First Impressions of Website Aesthetics with a Quantification of Perceived Visual Complexity and Colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2049–2058. https://doi.org/10.1145/2470654.2481281

[43] Jian Ren, Xiaohui Shen, Zhe L Lin, Radomir Mech, and David J Foran. 2017. Personalized Image Aesthetics.. In *ICCV*. 638–647.

[44] S. Ren, K. He, R Girshick, and J. Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39, 6 (2017), 1137–1149.

[45] Anthony Santella, Maneesh Agrawala, Doug Decarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April*. 771–780.

[46] Adrian Secord, Jingwan Lu, Adam Finkelstein, Manish Singh, and Andrew Nealen. 2011. Perceptual models of viewpoint preference. *Acm Transactions on Graphics* 30, 5 (2011), 1–12.

[47] Jacob Solomon. 2016. Heterogeneity in Customization of Recommender Systems By Users with Homogenous Preferences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4166–4170. https://doi.org/10.1145/2858036.2858513

[48] Hsiao Hang Su, Tse Wei Chen, Chieh Chi Kao, Winston H. Hsu, and Shao Yi Chien. 2012. Preference-Aware View Recommendation System for Scenic Photos Based on Bag-of-Aesthetics-Preserving Features. *IEEE Transactions on Multimedia* 14, 3 (2012), 833–843.

[49] Moshe Unger, Bracha Shapira, Lior Rokach, and Ariel Bar. 2017. Inferring Contextual Preferences Using Deep Auto-Encoding. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 221–229. https://doi.org/10.1145/3079628.3079666

[50] Nancy A Van House. 2011. Personal photography, digital technologies and the uses of the visual. *Visual Studies* 26, 2 (2011), 125–134.

[51] Wenguan Wang and Jianbing Shen. 2017. Deep Cropping via Attention Box Prediction and Aesthetics Assessment. (2017).

[52] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. 2018. Good View Hunting: Learning Photo Composition from Dense View Pairs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

[53] Yan Xu, Joshua Ratcliff, James Scovell, Gheric Speiginer, and Ronald Azuma. 2015. Real-time Guidance Camera Interface to Enhance Photo Aesthetic Quality. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1183–1186. https://doi.org/10.1145/2702123.2702418

[54] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. 2013. Learning the Change for Automatic Image Cropping. In *Computer Vision and Pattern Recognition*. 971–978.

[55] Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, and Chun Chen. 2013. Probabilistic Graphlet Transfer for Photo Cropping. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 22, 2 (2013), 802–815.

[56] Luming Zhang, Yingjie Xia, Kuang Mao, and He Ma. 2015. An Effective Video Summarization Framework Toward Handheld Devices. *Industrial Electronics IEEE Transactions on* 62, 2 (2015), 1309–1316.

[57] Mingju Zhang, Lei Zhang, Yanfeng Sun, and Lin Feng. 2005. Auto cropping for digital photographs. In *IEEE International Conference on Multimedia and Expo*. 4 pp.