Increasing Image Memorability with Neural Style Transfer

Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, Nicu Sebe

Increasing Image Memorability with Neural Style Transfer

ALIAKSANDR SIAROHIN, University of Trento, Italy
GLORIA ZEN, University of Trento, Italy
CVETA MAJTANOVIC, University of Trento, Italy and University of Novi Sad, Serbia
XAVIER ALAMEDA-PINEDA, Inria, France
ELISA RICCI, University of Trento and Fondazione Bruno Kessler (FBK), Italy
NICU SEBE, University of Trento, Italy

Recent works in computer vision and multimedia have shown that image memorability can be automatically inferred exploiting powerful deep learning models. This paper advances the state of the art in this area by addressing a novel and more challenging issue: "*Given an arbitrary input image, can we make it more memorable?*". To tackle this problem we introduce an approach based on an *editing-by-applying-filters* paradigm: given an input image, we propose to automatically retrieve a set of "style seeds", *i.e.* a set of style images which, applied to the input image through a neural style transfer algorithm, provide the highest increase in memorability. We show the effectiveness of the proposed approach with experiments on the publicly available LaMem dataset, performing both a quantitative evaluation and a user study. To demonstrate the flexibility of the proposed framework, we also analyze the impact of different implementation choices, such as using different state of the art neural style transfer methods. Finally, we show several qualitative results to provide additional insights on the link between image style and memorability.

CCS Concepts: • **Computing methodologies → Image manipulation**;

Additional Key Words and Phrases: Deep Learning, Style transfer, Memorability

ACM Reference Format:
Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2010. Increasing Image Memorability with Neural Style Transfer. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 4, Article 39 (March 2010), 22 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

We live in an instant society, a society where everything has to be quick and fast and in which the enormous amount of visual stimuli we receive daily has greatly overwhelmed the human capacity to store information. Indeed, though it was shown that the human brain retains knowledge derived from images and videos at a greater capacity with respect to other types of signals [9], it is practically impossible to absorb all visual information which we are daily exposed to. As a consequence, marketers and graphic designers are continuously struggling for capturing consumers' visual attention, and even more, to leave a trace in his/her memory. This has lead to the development of several digital tools which facilitate the modification of visual contents for capturing users gaze and for improving images memorability, shareability and likability. While research studies in

Authors' addresses: Aliaksandr Siarohin, University of Trento, Italy; Gloria Zen, University of Trento, Italy; Cveta Majtanovic, University of Trento, Italy and University of Novi Sad, Serbia; Xavier Alameda-Pineda, Inria, France; Elisa Ricci, University of Trento and Fondazione Bruno Kessler (FBK), Italy; Nicu Sebe, University of Trento, Italy.

1551-6857/2010/3-ART39 $15.00
https://doi.org/0000001.0000001

Fig. 1. Illustration of the main idea behind the proposed framework (best viewed in colors). Given a generic natural image (left), our approach automatically finds the best style filters, *i.e.* the "seeds" (center) that increase the most the memorability of the input image. Style filters are sorted by the corresponding predicted memorability increases. Memorability scores in the range [0,1] are reported (bottom right corner of each image).

cognitive neuroscience are used by marketers and designers to derive general rules for reshaping visual contents, it would be desirable to reduce and/or complement human intervention and devise technological solutions to automatically modify images and videos according to specific properties or attributes. The specific case of memorability is interesting in practice, since it has not only a commercial impact in marketing and fashion as well as in the way visual content influences people, but has also a profound impact in education and more generally in knowledge dissemination.

Following this line of thought, this work introduces an approach to automatically modify images such as to increase their memorability, *i.e.* their chance to be remembered. Previous research studies have shown that memorability is an intrinsic image property [17], *i.e.* different viewers exhibit similar memory performance in remembering/forgetting specific images [5, 40]. Other works have demonstrated that image memorability can not only be measured with specific user studies but also automatically predicted by employing advanced deep learning models with an accuracy close to human performance [21].

This work makes a step forward along this research direction and we propose an approach to automatically increase the memorability of an arbitrary input image by changing its style, *i.e.* its low level features, while preserving its high level content. More in detail, inspired by the *editing-by-applying-a-filter* paradigm implemented in popular tools such as Instangram or Prisma, we cast the problem of increasing image memorability as a problem of retrieving the most "memoralizable" style for the given image.

The main idea behind the proposed method is illustrated in Figure 1. Given a generic image, our approach provides as output a list of suggested style images, *i.e.* the *style seeds*, sorted by the estimated increase in memorability that the input image would achieve after stylization. The suggested seeds are then used, together with the input image, to compute a set of stylized and highly memorable images which are provided to the user. Our framework, which we refer as S^3 or S-cube, relies on three different modules (see Figure 2): the *Synthesizer*, which takes as input a generic image and a given style and outputs the corresponding stylized image; the *Scorer*, which assigns a memorability score to an input image and the *Selector*, which takes as input a generic image and suggests the most "memoralizable" styles for that image. In details, at training time, given a set of input images and different style seeds, the Synthesizer is used to generate several stylized images using neural style transfer techniques. Subsequently, the Scorer is applied to each triplet of original image-seed-stylized image, in order to compute a score reflecting the increase/decrease in terms of image memorability. The obtained scores are used to train the Selector, a deep neural network which, given an arbitrary input image is able to retrieve the top N seeds, *i.e.* the N seeds which produce the largest increase in memorability. At test time (Figure 1), the Selector is used to retrieve

the most memorabilizing seeds and provide them to the Synthesizer in order to generate highly memorable images. We evaluate the proposed approach conducting experiments on the publicly available LaMem dataset, showing that our method is effective in increasing image memorability.

This paper extends our previous work [37]. In particular, with respect to [37] we demonstrate the flexibility of the proposed framework showing that (i) different network architectures can be used for implementing both the Synthesizer and the Selector and that (ii) the optimal values of the hyper-parameters used in the Synthesizer can be automatically estimated, further increasing the memorability gap. Moreover, we significantly extend the experimental evaluation by performing a user study and adding a qualitative analysis to investigate the correlation between memorability and other perceptual properties such as affective quality.

To summarize, the contribution of our work is threefold:

- We introduce a novel framework to increase the memorability of images. We cast this task as a problem of selecting the most memoralizable style "seeds" for a given image. In this way we keep the users in the loop allowing them to choose among a small set of top styles. The effectiveness of our approach in recommending the best styles is demonstrated with several qualitative and quantitative results and with a user study.
- We demonstrate the flexibility of our framework with respect to different implementation choices. Specifically, we show that our method is agnostic to the choice of the style transfer method in the Synthesizer or to the neural architecture implementing the Selector.
- We propose an extended version of the method described in [37] where the optimal value of the style transfer hyper-parameter regulating the trade-off between style and content can be automatically estimated for a given image. In this way, our approach is able to compute not only the best styles but also to determine the optimal degree of stylization.

2 RELATED WORK

There are mainly three research lines related to our work: (i) studies addressing automatic image manipulation for altering perceptual attributes, (ii) works on neural style transfer and (iii) studies which analyze visual memorability. In this section we provide an overview of the most recent papers within these lines.

Image Manipulation. Recent works have shown that it is possible to manipulate images with the ultimate goal of altering perceptual properties such as aesthetic value [45], evoked emotions [31] and memorability [20, 21]. For instance, Wang *et al.* [45] proposed a deep learning architecture which increases the aesthetic value of a given picture by suggesting the best image cropping. The crop suggestion is based both on saliency and on aesthetic quality criteria. Peng *et al.* [31] attempted to modify the emotions evoked by an image by adjusting its color tone and its texture features. Similarly, in [14] a methodology to exploit color palette in order to modify arbitrary images and evoke specific emotion was presented. Ali *et al.* [2] described an image transformation approach for modifying a source image in a way such that it can induce an emotional affect on the viewer. Kim *et al.* [23] proposed a methodology based on semantic segmentation to modify the valance-arousal score of images of natural scenes. Khosla *et al.* [21] showed that by removing visual details from an image through a cartoonization process its memorability score can be modified. However, they did not provide a methodology to systematically increase the memorability of pictures. The same group [20] also demonstrated that it is possible to increase the memorability of faces, while maintaining the identity of the person and properties like age, attractiveness and facial expression. Up to our knowledge, ours is the first attempt to automatically increase the memorability of generic images (not only faces). We cast the problem of increasing image memorability as a problem of

selecting the most "memorabilizable" style filters, thus addressing the task of image manipulation under the popular "editing-by-applying-a-filter" paradigm.

Neural Style Transfer. The pioneering work in [8] on neural style transfer has been followed up by many other studies mostly addressing its limitations in terms of computational cost. In particular, Ulyanov *et al.* [44] dramatically reduced the time required for stylization, despite introducing the constraint that only a prefixed number of styles can be adopted. Huang *et al.* [15] proposed a style transfer method which is fast and works with arbitrary styles. Other research studies proposed modifications of the original style transfer framework [8] in order to adapt it to different applications, such as photorealistic rendering [27] or semantically composite transfer [6].

More recent works [46] addressed the problem of improving the quality of the stylized images, considering second order statistics for style representation. Other works [33] addressed the problem of style transfer in videos, proposing methods which generate multiple frames with a specific style while ensuring temporal coherence. Sheng *et al.* [36] proposed an efficient technique for zero-shot style transfer, *i.e.* for transferring arbitrary styles into content images. A Multi-style Generative Network was presented in [46] with the purpose of retaining the functionality of earlier optimization-based approaches, while ensuring real time processing.

However, to the best of our knowledge no previous works on deep stylization focused on modifying images in order to enhance specific high level attributes such as memorability.

Memorability and Related Perceptual Attributes. Visual memorability has been investigated by several works both from a psychological [9, 41] and a computational perspective [16, 39]. While works in psychology mostly focused on the human memory capacity, recent works in computer vision studied memorability as an intrinsic property of images. Several works considered the question *What makes an image memorable?*, finding that image properties like distinctiveness [5, 40] and arousal [25, 29, 32] have a positive influence on memorability. Khosla *et al.* [21] demonstrated a positive correlation between memorability and popularity, suggesting that memorable images have a higher chance of becoming popular. Subsequent works from the same research group showed that it is possible to predict the number of views that an image will receive on social media even before it is uploaded [19]. Other studies lead to the conclusion that high level concepts like the presence of faces are what contribute mostly to memorability [18]. Similarly, a recent study demonstrated that users are more likely to watch videos which have highly memorable and interesting video summaries [30]. How to automatically modify generic images in order to make them memorable has been addressed in [35]. However, this work does not exploit the flexibility of neural style transfer techniques to solve this task.

The link between memorability and style-related cues like colors has been previously explored. For instance, studies found that harmonious colors appear to be more memorable [42]. Other works showed a negative correlation between memorability and other image properties like aesthetic value [18, 22] and interestingness [12]. In particular, visual interestingness was found to positively correlate with the perceived image naturalness [13, 39] and complexity [1, 3, 7, 11, 39]. Additionally, works in psychology investigated the link between visual complexity and human memory. For example, high complexity was found to reduce the ability to focus on important visual information in [16]. In this work, we advance the state of the art on computational models for studying memorability by analyzing the role of the image style and by further investigating the relation between memorability and other perceptual attributes such as interestingness. To the best of our knowledge, no works so far have investigated the effect of transferring a style to an image in order to enhance its memorability.

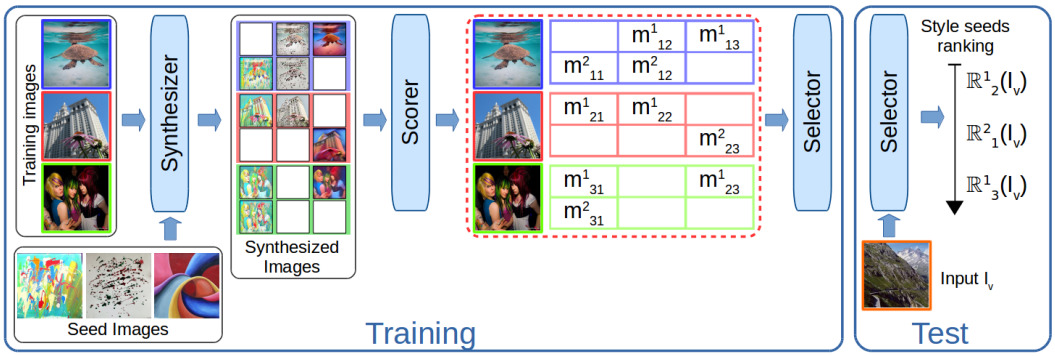

Fig. 2. Overview of our method. At training time, the Synthesizer Sy and the Scorer Sc serve to generate the training data (highlighted with a red dotted frame) for the seed Selector Sl. At test time, the seed Selector provides for each new image a sorted list of style seeds, based on the predicted memorability increase $\mathrm{Sl}_s^{\alpha}(\mathbf{I}_v)$.

3 METHOD

In this section we introduce the proposed framework to automatically increase the memorability of an input image. Our method is designed in a way such that the process of generating highly memorable stylized images is performed in an efficient way. In a nutshell, our approach exploits neural style transfer methods to create stylized images. However, in our framework the process of stylization is driven by a specific module which ensures that the generated images have increased memorability and, implicitly, that most of the high level content of the original images is preserved.

3.1 Overview

The proposed S-cube approach co-articulates three main components, namely the seed Selector, the Scorer and the Synthesizer. In order to give a general idea of our methodological framework, we illustrate the pipeline associated to S-cube in Figure 2. The Selector is the core of our approach: for a generic input image $\mathbf{I}$ and given a set of style image seeds $\mathcal{S}$, the Selector retrieves the subset of $\mathcal{S}$ that will be able to produce the largest increase of memorability. In details, the seed Selector predicts the expected increase/decrease of memorability that each seed $\mathrm{S} \in \mathcal{S}$ will produce in the input image $\mathbf{I}$, and consequently it ranks the seeds according to the expected increase of memorability. At training time, the Synthesizer and the Scorer are used to generate stylized images from many input image-seed pairs and to score these pairs, respectively. Each input image is then associated to the relative increase/decrease of memorability score obtained with each of the seeds. With this information, we can learn to predict the increase/decrease of memorability for a new image, and therefore rank the seeds according to the expected increase. Indeed, at test time, the Selector is able to retrieve the most memoralizable seeds and give them to the Synthesizer. In this paper we also present an extended version of the original method in [37], where we learn to predict not only the most memoralizable seeds but for each seed we also automatically compute the optimal degree of stylization. In the following, we first formalize the S-cube framework and then describe each of the three components in detail.

3.2 The S-cube approach

Let us denote the Scorer, the Synthesizer and the seed Selector models by Sc, Sy and Sl, respectively. During the **training phase** the three models are learned. The Scoring model Sc returns the memorability value $\mathrm{Sc}(\mathbf{I})$ of a generic image $\mathbf{I}$, and it is learned by means of a training set of images annotated with memorability: $\mathcal{M} = \{\mathbf{I}_i^{\mathcal{M}}, m_i\}_{i=1}^{I}$. In addition to this training set, we also consider

a generating set of natural images $\mathcal{G} = \{\mathbf{I}_g^{\mathcal{G}}\}_{g=1}^{G}$ and a set of style seed images $\mathcal{S} = \{\mathbf{S}_s\}_{s=1}^{S}$. The Synthesizer produces an image from an image-seed pair,

$$\mathbf{I}_{gs} = \mathsf{Sy}\left(\mathbf{I}_g^{\mathcal{G}}, \mathbf{S}_s\right). \tag{1}$$

The Scoring Sc and the Synthesizer Sy are the required steps to train the seed Selector Sl. Indeed, for each image $\mathbf{I}_g^{\mathcal{G}} \in \mathcal{G}$ and for each style seed $\mathbf{S}_s \in \mathcal{S}$, the synthesis procedure generates $\mathbf{I}_{gs}$. The Scoring model is used to compute the memorability score gap between the synthesized and the original images:

$$m_{gs}^{\mathsf{Sc}} = \mathsf{Sc}(\mathbf{I}_{gs}) - \mathsf{Sc}(\mathbf{I}_g^{\mathcal{G}}). \tag{2}$$

The seed-wise concatenation of these scores, denoted by $\mathbf{m}_g^{\mathsf{Sc}} = (m_{gs}^{\mathsf{Sc}})_{s=1}^{S}$, is used to learn the seed Selector. Specifically, a training set of natural images labeled with the seed-wise concatenation of memorability gaps $\mathcal{R} = \{\mathbf{I}_g^{\mathcal{G}}, \mathbf{m}_g^{\mathsf{Sc}}\}_{g=1}^{G}$ is constructed. Once the Selector is trained on $\mathcal{R}$, it is able to estimate the vector of memorability gaps for a test image, which is much faster than running the Synthesizer and the Scorer S times (one per seed). This is one of the main contributions of the present paper. Moreover, the memorability gap vector provides a ranking of the seeds in terms of their ability to memorabilize images (*i.e.,* the best seed corresponds to the largest memorability increase).

During the **test phase** and given a novel image $\mathbf{I}_v$, the seed Selector is applied to predict the vector of memorability gap scores associated to all style seeds, *i.e.* $\mathbf{m}_v = \mathsf{Sl}(\mathbf{I}_v)$. A ranking of seeds is then derived from the vector $\mathbf{m}_v$. Based on this ranking the Synthesizer is applied to the test image $\mathbf{I}_v$ considering only the top Q style seeds $\mathbf{S}_s$ and produces a set of stylized images $\{\mathbf{I}_{qs}\}_{q=1}^{Q}$.

In the following we describe the three main building blocks of our approach, providing details of our implementation. We also present the extended version of the proposed method for the automatic selection of the degree of stylization.

3.2.1 The Scorer. The scoring model Sc returns an estimate of the memorability associated to an input image $\mathbf{I}$. In our work, we use the memorability predictor based on LaMem dataset in [21], which is the state of the art to automatically compute image memorability. In details, following [21] we consider the AlexNet CNN model as pre-trained in [47] (named Hybrid-CNN), *i.e.* on the ImageNet and Places datasets. Then, we randomly split the LaMem training set into two disjoint subsets (of 22,500 images each), $\mathcal{M}$ and $\mathcal{E}$. Starting from the pre-trained mode, we minimize the Euclidean distance of the scores on each subset of LaMem, thus learning two independent scoring models Sc and Ev. While, as discussed above, Sc is used during the training phase of our approach, the model Ev is adopted for evaluation. For training, we run 70k iterations of stochastic gradient descent with momentum 0.9, learning rate 10^{-3} and batch size 256.

3.2.2 The Synthesizer. The Synthesizer takes as input a generic image $\mathbf{I}_g$ and a style seed image $\mathbf{S}_s$ and produces a stylized image $\mathbf{I}_{gs} = \mathsf{Sy}(\mathbf{I}_g, \mathsf{Sy}_s)$. In this work we consider two different neural style transfer methods to implement the Synthesizer, namely the approach in [43] and the most recent method in [15].

The strategy proposed in [43] consists on training a different feed-forward network for every seed. In this work as seeds we use 100 abstract paintings from the DeviantArt database [34], and therefore we train $S = 100$ networks for 10k iterations with learning rate 10^{-2}. The most important hyperparameter of the style transfer method in [43] is the coefficient α, which regulates the trade-off between preserving the original image content and generating something closer to the style image (see Figure 3). In our experiments we evaluated the effect of α on the creation of highly memorable images (see Section 4.2).

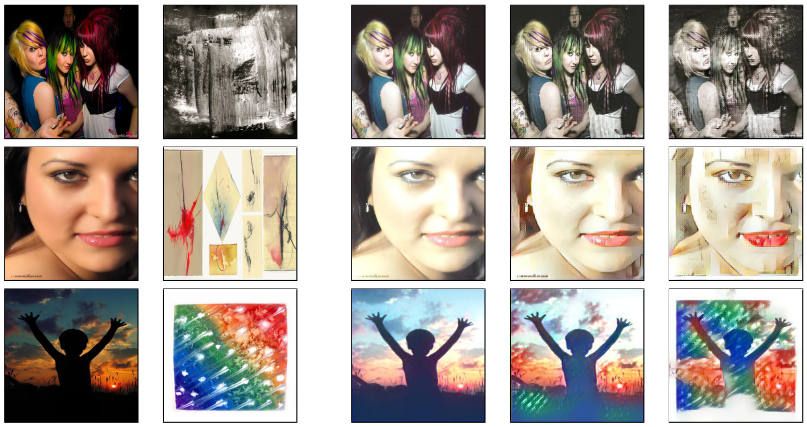

Fig. 3. Sample stylized images. (Left) Original images and applied style seeds. (Right) Synthesized images obtained with the method in [43] at varying α.

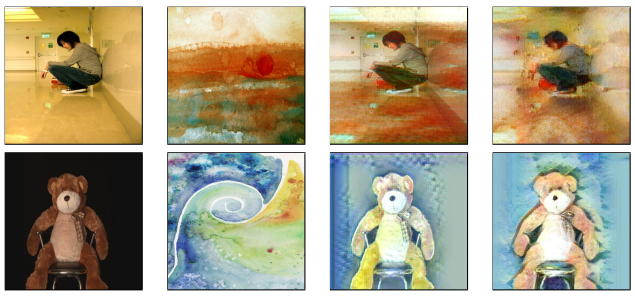

Fig. 4. Sample stylized images. Original images $\mathbf{I}_v$, applied style seeds $\mathbf{S}_s$ and stylized images obtained using as Synthesizer the methods in [43] and [15].

It is worth noticing that the methodology proposed in this article is independent of the synthesis procedure. Indeed, we also tried another method recently proposed by Huang *et al.* [15], which achieves very good stylization performance while keeping the computational complexity low. This is especially important in our framework since the Synthesizer is also used to generate the training set for learning $\texttt{Sl}$. Similarly to [43], in [15] a parameter α can be set to regulate the degree of stylization. Figure 4 depicts some sample stylization results obtained using the two style transfer methods considered in our work [15, 43].

3.2.3 The Seed Selector. The core part of our approach is the Selector. Given a training set of natural images labeled with the vector of memorability gaps: $\mathcal{R} = \{\mathbf{I}_g^{\mathcal{G}}, \mathbf{m}_g^{\texttt{Sc}}\}_{g=1}^{G}$, the seed Selector $\texttt{Sl}$ is trained minimizing the following objective:

$$\mathcal{L}_{\texttt{Sl}} = \sum_{g=1}^{G} \mathcal{L}\left(\texttt{Sl}(\mathbf{I}_g^{\mathcal{G}}), \mathbf{m}_g^{\texttt{Sc}}\right). \tag{3}$$

where $\mathcal{L}$ is a loss function which measures the discrepancy between the learned vector $\texttt{Sl}(\mathbf{I}_g^{\mathcal{G}})$ and the memorability gap scores $\mathbf{m}_g^{\texttt{Sc}}$. By training the seed Selector with memorability gaps, we are learning *by how much each of the seeds increases or decreases the memorability of a given image.*

This has several advantages. First, we can very easily rank the seeds by the expected increase in memorability they will produce if used together with the input image and the synthesis procedure. Second, if several seeds have similar expected memorability increase, they can be proposed to the user for further selection. Third, if all seeds are expected to decrease the memorability, the optimal choice of not modifying the image can easily be made. Fourth, once Sl is trained, all this information comes at the price of evaluating Sl for a new image, which is cheaper than running Sy and Sc S times.

Even if this strategy has many advantages at testing time, the most prominent drawback is that, to create the training set $\mathcal{R}$, one should ideally call the synthesis procedure for all possible image-seed pairs. This clearly reduces the scalability and the flexibility of the proposed approach. The scalability because training the model on a large image dataset means generating a much larger dataset (*i.e.*, S times larger). The flexibility because if one wishes to add a new seed to the set $\mathcal{S}$, then all image-seed pairs for the new seed need to be synthesized and this takes time. Therefore, it would be desirable to find a way to overcome these limitations while keeping the advantages described in the previous paragraph.

The solution to these issues comes with a model able to learn from a partially synthesized set, in which not all image-seed pairs are generated and scored. This means that the memorability gap vector $\mathbf{m}_g^{\mathbb{M}}$ has missing entries. In this way we only require to generate *enough* image-seed pairs. To this aim, we propose to use a decomposable loss function $\mathcal{L}$. Formally, we define a binary variable ω_{gs} set to 1 if the gs-th image-seed pair is available and to 0 otherwise and rewrite the objective function in (3) as:

$$\mathcal{L}_{\mathtt{Sl}} = \sum_{g=1}^{G} \sum_{s=1}^{S} \omega_{gs} \ell \left(\mathtt{Sl}_s(\mathbf{I}_g^{\mathcal{G}}), m_{gs}^{\mathtt{Sc}} \right). \tag{4}$$

where $\mathtt{Sl}_s$ is the s-th component of Sl and ℓ is the square loss. We implemented this model using an AlexNet architecture [24], where the prediction errors for the missing entries of $\mathbf{m}_g^{\mathtt{Sc}}$ are not back-propagated. Specifically, we considered the pre-trained Hybrid CNN model [47] and fine-tune only the layers fc6, fc7, conv5, conv4 using a learning rate equal to 10^{-3}, momentum equal to 0.9 and batch size 64. The choice of Hybrid-CNN is considered more appropriate when dealing with generic images since the network is pre-trained both on images of places and objects.

3.2.4 Learning the Degree of Stylization. In the S-cube method described above, the parameter α used by the Synthesizer and regulating the trade-off between content and style is assumed to be fixed and is defined *a priori*. In this section we introduce an extended version of S-cube, where the seed Selector is also able to predict the optimal α value, *i.e.* the optimal degree of stylization, for a given image-seed pair. To this aim we modify the implementation of the seed Selector and train it using an augmented training set generated with the Synthesizer. Specifically, by considering explicitly α we rewrite Eqn. 1 denoting the stylized image as $\mathbf{I}_{gs}^{\alpha} = \mathtt{Sy}\left(\mathbf{I}_g^{\mathcal{G}}, \mathtt{S}_s, \alpha\right)$. The memorability score gap between the synthesized and the original image in Eqn. 2 is denoted as $m_{gs\alpha}^{\mathtt{Sc}} = \mathtt{Sc}(\mathbf{I}_{gs}^{\alpha}) - \mathtt{Sc}(\mathbf{I}_g^{\mathcal{G}})$ and $\mathbf{m}_{g\alpha}^{\mathtt{Sc}}$ indicates the vector of the score gaps for all styles and a given value of α.

Assuming that the set $\mathcal{A}$ of the possible values of α is finite, for implementing the Selector we propose to learn different deep network models, each corresponding to a specific degree of stylization α and computing for each image $\mathbf{I}_g$ the associated memorability gap vector $\mathbf{m}_{g\alpha}^{\mathtt{Sc}}$. However, if the cardinality of the set $\mathcal{A}$ is large, learning different models may be inefficient both in terms of memory and computational cost. To address this issue, we resort to a multi-task learning framework. Specifically, we consider the pretrained Hybrid CNN model [47] and we implement the different network models imposing that they share the same parameters in the convolutional block and

differ only in the fully connected layers. Our assumption is that the learned models are related as their training sets have been generated with the same style seeds and images. In our experiments we considered three different values of α, *i.e* $\alpha \in \mathcal{A} = \{0.5, 2, 10\}$. This means that, for instance, for generating the augmented training set for the Selector with the method in [43] we use $S = 300$ networks, one for each pair of α value and style $\mathbf{S}_s$.

4 EXPERIMENTAL VALIDATION

In this section we report the results of our experimental validation. First, we describe the datasets used in our evaluation and the adopted experimental setup (Section 4.1). Then, we report the performance of *S*-cube in increasing image memorability by selecting the most memoralizable seed styles (Section 4.2). We also present the results of a user study, where we collected the actual memorability scores for a set of original and stylized image pairs (Section 4.3). Finally, in Section 4.4 we present some qualitative results.

4.1 Datasets and Experimental Setup

4.1.1 Datasets. We considered two datasets in our evaluation: LaMem [21] and DeviantArt [34].

The *LaMem* dataset [21] is currently the largest dataset used to study visual memorability. It is a collection of 58,741 images gathered from a number of previously existing datasets, including the Affective Images dataset [28] which consists of Art and Abstract paintings. The memorability scores were collected for all the pictures in the dataset using an optimized protocol of the memorability game. The corpus was released to overcome the limitations of previous works on memorability which used small datasets and very specific image domains. The large appearance variations of the images makes LaMem especially suitable for analyzing the performance of our method.

The *DeviantArt* dataset [34] consists of a set of 500 abstract art paintings collected from deviantArt (dA), an online social network site devoted to user-generated art. Since the main idea behind our method is to avoid substantial modifications of the high-level image content we selected the style seeds from abstract paintings. Indeed, abstract art relies on texture and color combinations, thus it is especially appropriate when targeting the automatic modification of low-level features.

4.1.2 Experimental Setup. In our experiments using the LaMem dataset we considered the same training (45,000 images), test (10,000 images) and validation (3,741 images) data adopted in [21]. We split the LaMem training set into two subsets of 22,500 images each (see also Section 3.2.1), $\mathcal{M}$ and $\mathcal{E}$, which are used to train two predictors Sc and Ev, respectively. The model Sc is the Scorer employed in our framework, while Ev (which we will denote in the following as the external predictor or the evaluator) is used to evaluate the performance of our approach, as a proxy for human assessment. We highlight that Sc and Ev can be used as two independent memorability scoring functions, since $\mathcal{M}$ and $\mathcal{E}$ are disjoint. The validation set is used to implement the early stopping. To evaluate the performance of our Scorer models Sc and Ev, following [21], we compute the rank correlation between predicted and actual memorability on LaMem test set. We obtain a rank correlation of 0.63 with both models, while [21] achieves a rank correlation of 0.64 training on the whole LaMem training set. As reported in [21], this is close to human performance (0.68).

The test set of LaMem (10k images) is then used (i) to learn the proposed seed Selector and (ii) to evaluate the overall framework (and the Selector in particular). In detail, we split LaMem test set into train, validation and test for our Selector with proportion 8:1:1, meaning 8,000 for training and 1,000 for validation and test. The training set for the Selector was already introduced as $\mathcal{G}$. The validation set is used to perform early stopping, if required. We denote the test set as $\mathcal{V}$.

Regarding the seeds, we estimated the memorability of all paintings of DeviantArt using Sc and selected the 50 most and the 50 least memorable images as seeds in the set $\mathcal{S}$. The memorability scores of the deviantArt images range from 0.556 to 0.938.

For the user study, we randomly selected 66 images from $\mathcal{V}$ and we assigned to each of them a style image randomly extracted from $\mathcal{S}$. Then, for each image-seed pair we collect the actual memorability scores following the memorability game protocol described in [21]. The purpose of our user study is twofold: (i) to show that our method is able to increase the memorability of arbitrary images and (ii) to demonstrate that the external predictor Ev is a good proxy for a user evaluation.

4.1.3 Evaluation Metrics. We evaluate the performance of our method in predicting the memorability increase of an image-seed pair using two different performance measures: the mean squared error (MSE) and the accuracy A, which are defined as follows:

$$\mathrm{MSE}^{\mathtt{X}} = \frac{1}{SV}\sum_{s=1}^{S}\sum_{v=1}^{V}\left(m_{vs}^{\mathtt{X}} - \mathtt{Sl}_s(\mathbf{I}_v^{\mathcal{V}})\right)^2 \tag{5}$$

and

$$A^{\mathtt{X}} = \frac{1}{SV}\sum_{s=1}^{S}\sum_{v=1}^{V}(1 - |HS(m_{vs}^{\mathtt{X}}) - HS(\mathtt{Sl}_s(\boldsymbol{I}_v^{\mathcal{V}}))|) \tag{6}$$

where a generic image $\mathbf{I}_v^{\mathcal{V}}$ is taken from a set of (yet) unseen images $\mathcal{V} = \{\mathbf{I}_v^{\mathcal{V}}\}_{v=1}^{V}$ and the seed S is taken from a set of style seeds. HS is the Heaviside step function, which sets the input variables to 0 or 1, respectively when their initial values is lower or higher than zero [26].The evaluation is performed based on the internal predictor (Sc), the external predictor (Ev) or the human assessment (H) as indicated with $\mathtt{X} \in \{\mathtt{Sc}, \mathtt{Ev}, \mathtt{H}\}$.

4.1.4 Baseline. To the best of our knowledge this is the first work showing that it is possible to automatically increase the memorability of a generic image. For this reason, a direct and quantitative comparison with previous studies is not possible. Indeed, the recent work [21] showed that it is possible to compute accurate memorability maps from images, which can be used as bases for further image manipulations. They also observed that using a memorability map for removing image details, such as through a cartoonization process, typically lead to a memorability decrease. Complementarily, we aim to effectively increase image memorability without modifying the high level content of the images. Therefore, the approach by [21] does not directly compare with ours. The only potential competitor to our approach would be [20], except that the method is specifically designed for face photographs. Indeed, the proposed approach aims to modify the memorability while keeping other attributes (age, gender, expression) as well as the identify untouched. Therefore, the principle of [20] cannot be straightforwardly transferred to generic images. Consequently, we define an *average* baseline $\mathcal{B}$ that consists on ranking the style seeds according to the average memorability increase for the training set, defined as:

$$\bar{m}_s^{\mathtt{Sc}} = \frac{1}{G}\sum_{g=1}^{G} m_{gs}^{\mathtt{Sc}}. \tag{7}$$

In other words, we are comparing the proposed image-dependent seed selector with an image-independent seed selector. The latter consists in selecting, for a test image, the seed that maximizes the memorability gap on average over the generating set.

	A^{Sc}		A^{Ev}		MSE^{Sc}		MSE^{Ev}	
$\bar{\omega}$	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube
0.01	**63.21**	57.12	**60.96**	56.01	**0.0113**	0.0138	**0.0119**	0.0137
0.1	64.49	**64.70**	61.07	**62.22**	**0.0112**	0.0114	**0.0117**	0.0119
0.5	64.41	**67.18**	61.06	**64.38**	0.0112	**0.0102**	0.0117	**0.0106**
1	64.41	**67.75**	61.06	**64.70**	0.0112	**0.0102**	0.0117	**0.0108**

Table 1. Performance of S-cube compared to baseline $\mathcal{B}$ at varying percentage of training data $\bar{\omega}$, measured in terms of (left) accuracy A and (right) mean squared error (MSE). Performances have been evaluated using both the internal Sc and the external Ev predictor.

	A^{Ev}		MSE^{Ev}	
S	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube
20	60.66	**63.15**	0.0114	**0.0111**
50	61.09	**63.51**	0.0116	**0.0109**
100	61.06	**64.38**	0.0117	**0.0106**

Table 2. Performance of S-cube compared to baseline $\mathcal{B}$ at varying the cardinality S of the style set $\mathcal{S}$ measured in terms of Accuracy A^{Ev} and Mean Square Error MSE^{Ev}.

4.2 Experimental Results

In this section we assess the effectiveness of our method in suggesting "memoralizable" style seeds under different experimental setups. Our code is available online[1].

4.2.1 Increasing Image Memorability. Table 1 reports the performance of both the proposed approach (S-cube) and the baseline ($\mathcal{B}$) for different values of the average amount of image-seed pairs $\bar{\omega}$. More precisely, $\bar{\omega} = 1$ means that all image-seed pairs are used, $\bar{\omega} = 0.1$ means that only 10% is used, and so on. We report the accuracy (A) and the MSE evaluated using the internal scoring model Sc and the external scoring model Ev. Generally speaking our method outperforms the baseline if enough image-seed pairs are available. We argue that, as it is well known, deep architectures require a *sufficient* amount of data to be effective. Indeed, when $\bar{\omega} = 0.01$, the network optimization procedure attempts to learn a regression model from the raw image to a 100-dimensional space with, on average, only one of these dimensions propagating the error back to the network. Although this dimension is different for each image, we may be facing a situation in which not enough information is propagated back so as to effectively learn a robust regressor. This situation is coherent when the scoring method changes from Sc to Ev. We can clearly observe an increase in performance when using Sc, as expected. Indeed, since the seed Selector has been trained to learn the memorability gap from Sc, the performance is higher when using Sc instead of Ev. This result further motivates the need of having an external Scorer Ev, trained on an independent set of images, to evaluate the performance of our method. In this series of experiments and in the following, unless otherwise specified, we set $\alpha = 2$.

[1] https://github.com/AliaksandrSiarohin/mem-transfer

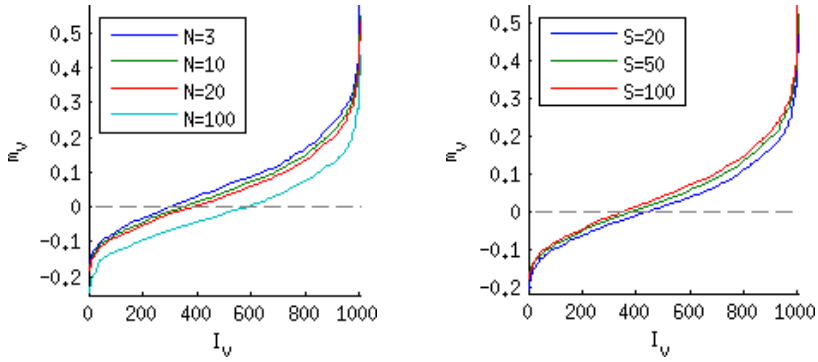

Fig. 5. Sorted average memorability gaps $\bar{m}_v$ obtained with our method S-cube (left) averaging over varying number of top N seeds and (right) at varying the cardinality S of the seed set, with $N = 10$. Abscissa corresponds to test image index, ranked by $\bar{m}_v$. The wider is $\bar{m}_v$, the better.

Furthermore, we studied the behavior of our framework when varying the size S of the seed set. Results are shown in Table 2. The parameter $\bar{\omega}$ is set to 0.5. Specifically, we select two sets of 50 and 20 seeds out of the initial 100, randomly sampling these seeds half from the 50 most and half from the 50 least memorable ones. In terms of accuracy, the performance of both the proposed method and the baseline remain pretty stable when decreasing the number of seeds. However, a different trend is observed for the MSE. Indeed, while the MSE of the proposed method increases when reducing the number of seeds (as expected), the opposite trend is found for the baseline method. We argue that, even if the baseline method is robust in terms of selecting the bests seeds to a decrease of the number of seeds, it does not do a good job at predicting the actual memorability increase. Instead, the proposed method is able to select the best seeds and better measures their impact, especially when more seeds are available. This is especially important if the method wants to be deployed with larger seed sets.

We also assess the validity of our method as a tool for effectively increasing the memorability of a generic input image $\mathbf{I}_v$. In Figure 5 (left) we report the average memorability gap $\bar{m}_v$ over the top N seeds retrieved, with $N = 3, 10, 20$ and all the seeds. For display purposes, we rank the images of test set $\mathcal{V}$ by their average memorability gap, so that in Figure 5 (left), the curve closer to the upper-left corner corresponds the the best method. It can be noted that $\bar{m}_v$ achieves higher values when smaller sets of top N seeds are considered, as an indication that our method effectively retrieves the most memorabilizing seeds. In Figure 5 (right) we report the average memorability gaps $\bar{m}_v$ obtained over the test set $\mathcal{V}$ with S-cube, considering $N = 10$ and a varying number of style seeds S. As expected, a larger number of seeds results into a higher increase in term of memorability.

Finally, we investigated the link between the memorability score of the style seeds and the corresponding average memorability increase obtained for each seed. A low correlation is found between these two sets, meaning that the seed style alone is not predictive for the memorability increase and that the optimal ranking is dependent also on the image. In details, we found a Pearson rank correlation ρ=0.277 in the case of $\alpha = 2$. These results further motivate the intuition behind the proposed method: increasing the memorability of an image means selecting the style which better matches its visual appearance.

4.2.2 S-cube as a generic framework. In this subsection we show additional results in order to demonstrate that the proposed S-cube is a general framework and different choices can be made

	A^{Ev}		MSE^{Ev}	
	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube
Ulyanov *et al.* [43]	61.06	**64.38**	0.0117	**0.0106**
Huang *et al.* [15]	70.08	**75.12**	0.0142	**0.0112**
AlexNet [24]	61.06	**64.38**	0.0117	**0.0106**
VGG16 [38]	61.06	**63.49**	0.0117	**0.0111**

Table 3. Performance of S-cube compared to the baseline $\mathcal{B}$ under different implementation choices, evaluated in terms of Accuracy and MSE and using the external predictor (Ev): (*top*) using the style transfer methods in [43] and in [15] and (*bottom*) using different architectures for the Selector.

	Huang *et al.* [15]	Ulyanov *et al.* [43] α=0.5	Ulyanov *et al.* [43] α=2	Ulyanov *et al.* [43] α=10
Huang *et al.* [15]	1.000	0.789	0.875	0.938
Ulyanov *et al.* [43] α=0.5	-	1.000	0.936	0.836
Ulyanov *et al.* [43] α=2	-	-	1.000	0.934
Ulyanov *et al.* [43] α=10	-	-	-	1.000

Table 4. Pearson correlations between memorability gaps of each image-seed pair in the test set $\mathcal{V}$, obtained using different style transfer methods.

to implement the main system components. In particular, in our experiments we consider the Synthesizer and the Selector.

For the *Synthetizer*, as discussed in Section 3.2.2, we used two different style transfer methods, one proposed by Ulyanov *et al.* [43] and the other by Huang *et al.* [15]. Performances in terms of Accuracy and MSE for S-cube and the baseline $\mathcal{B}$ are reported in Table 3. Better performance are obtained using our S-cube with respect to the baseline. These results demonstrate that our method can integrate different style transfer methods, suggesting that S-cube can be easily upgraded (and further improved) when novel style transfer techniques are made available in literature.

Similarly, for the *Selector* we consider different deep networks. In particular, in Table 3 we report the performance of the S-cube when using two different architectures, VGG16 [38] (pre-trained on ImageNet) and AlexNet (Hybrid-CNN [47]). The details of these architectures can be found in the original papers [38, 47]). From the Table it can be observed that S-cube consistently outperforms the baseline.

Furthermore, we investigated the impact of using different style transfer techniques on the style ranking for each image in the test set $\mathcal{V}$. To this aim, we computed the Pearson correlation between the memorability gaps obtained with S-cube using different style transfer methods. These values are reported in Table 4. As it can be seen, a strong correlation, i.e. ρ =0.875, is found when considering Huang *et al.* [15] and Ulyanov *et al.* [43] with $\alpha = 2$. A correlation value greater than 0.8 is found when considering Ulyanov *et al.* [43] with different α values.

4.2.3 Analyzing the impact of the degree of stylization α. We also performed additional experiments to study the impact of the hyper-parameter α on the performance of the method. In these experiments we consider the style transfer approach in [43] for implementing the Synthesizer. In Table 5 we show the performance of S-cube when the system is trained using a pre-defined α value.

		A^{X}		MSE^{X}	
	α	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube
X = Sc	0.5	62.70	**63.37**	0.0102	**0.0100**
	2	64.41	**67.75**	0.0112	**0.0102**
	10	67.99	**73.25**	0.0125	**0.0104**
X = Ev	0.5	58.30	**59.50**	0.0107	**0.0103**
	2	61.06	**64.70**	0.0117	**0.0108**
	10	68.31	**71.71**	0.0132	**0.0111**

Table 5. Performance of S-cube at varying α. Performance is reported in terms of Accuracy and MSE using both the internal (Sc) and the external (Ev) predictor.

$\mathcal{A}$	Top 3	Top 10	Top 20	Top 30	All
$\{0.5\}$	0.0574	0.0377	0.0217	0.0143	-0.0085
$\{2\}$	0.0739	0.0567	0.0440	0.0352	-0.0096
$\{10\}$	0.0695	0.0651	0.0573	0.0493	-0.0251
$\{0.5, 2, 10\}$	**0.0742**	**0.0688**	**0.0606**	**0.0516**	**0.0064**

Table 6. Average memorability increases obtained when considering the top N memoralizable style seeds retrieved with S-cube and different sets $\mathcal{A}$.

In all the cases S-cube performs better than the baseline $\mathcal{B}$, both when considering the internal and the external predictor.

In Table 6 we show the performance obtained with the extended version of our method which estimates the best style seed and the Synthesizer parameter α for a given image. We report the average memorability increases over the test set $\mathcal{V}$, obtained when averaging over the top N best style seeds retrieved for each test image ($N = 3, 10, 20, 30$ and 100). Our method achieves the best performance when it is possible to choose the optimal α value for each image-seed pair. In other words, by actively selecting α it is possible to achieve a higher increase in terms of memorability. In order to give an idea of how well our method is performing, we computed the Upper Bound (UB) of the memorability increase, achievable for a given set of images and styles, in the case of α=2 and Top 3. Specifically, we run the Stylizer with for the set of image-style pairs, we ranked for each test image the obtained stylized images according to the memorability measured by the internal Scorer Sc, we selected the Top 3 for each test image and we averaged the corresponding memorability increases based on the external Scorer Ev. The following results are obtained: Baseline ($\mathcal{B}$): 0.0694; S-cube: 0.0739; UB: 0.0804.

To further analyze the impact of the degree of stylization, we also conduct an experiment to see if there is correlation between high/low values of the parameter α and certain types of style seeds. Specifically, we run an experiment to verify whether it exists for each seed a tendency to be assigned to a specific α value among the three considered (α=0.5, 2 or 10). The probability for each style seed to be assigned to α equal to 0.5, 2 or 10 is depicted in Fig.6. It is easy to see that this probability changes according to the style seed. In other words, in order to increase the memorability of a test image, some seeds tend to be selected with a low α, other with a high value. In Fig. 7, we reported the top 10 style seeds which are usually assigned with a low value of stylization coefficient ($\alpha = 0.5$) (see Fig.6-top) and the top 10 styles which are usually assigned with a high value of stylization

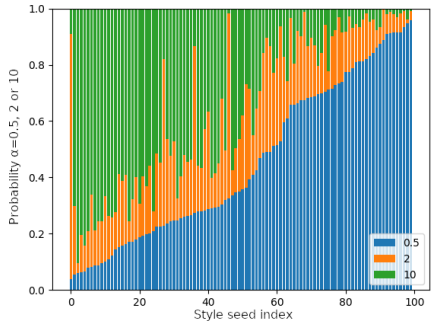

Fig. 6. Bar plot showing, for each of the 100 style seeds used, the probability to be selected with (blue) $\alpha = 0.5$, (orange) $\alpha = 2$ or (green) $\alpha = 10$ by our method S-cube. The seed styles are sorted by crescent probability of being selected with $\alpha = 0.5$.

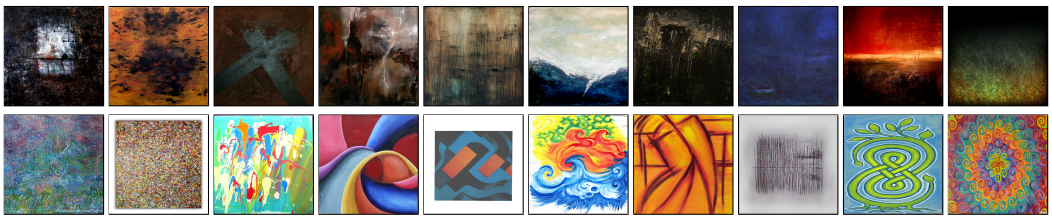

Fig. 7. (Top) 10 top seeds having a higher probability to be assigned to α=0.5 and (bottom) 10 top seeds having a higher probability to be assigned to α=10.

coefficient ($\alpha = 0.5$) (see Fig.6-bottom). Indeed it can be seen that images from the first set look mostly dark and gloomy, while those from the second set are more colorful and bright. Also, in the first set, lines and edges are less defined with respect to the second set.

4.3 User study

To further demonstrate the effectiveness of our method, in this section we report the results of a user study. We first provide the details of the memory game protocol we implemented to collect memorability scores from users, then we illustrate the results of our study.

4.3.1 Memory game protocol. We collected the actual memorability scores for a verification set $\mathcal{Z}$ which was randomly sampled from the test set $\mathcal{V}$, as detailed in Section 4.1. To this aim, we follow the protocol of the efficient memory game described in [21] applying some modifications in order to adapt the game to our scenario of image stylization. We build a memory game session by randomly selecting 66 target images, 12 vigilance repeats and 59 fillers out of the 1,000 images of the LaMem test set $\mathcal{V}$. For each image, we randomly sampled an associated style out of $\mathcal{S}$. For each session, we made sure that 33 targets are shown in their original version, while the other 33 are displayed in their corresponding stylized version. Also, we make sure that each image is displayed strictly in its original or in its stylized version, with α=2. Each target repeat is shown after a minimum of 35 to a maximum of 150 images. Vigilance repeats were shown within 7 images from the first showing. Vigilance repeats ensure that the user is focused on the game. While for each session the set of targets, vigilance repeats and fillers is preserved, we created 100 image sequences with a different image sorting, randomly assigned, so that no memory bias is introduced

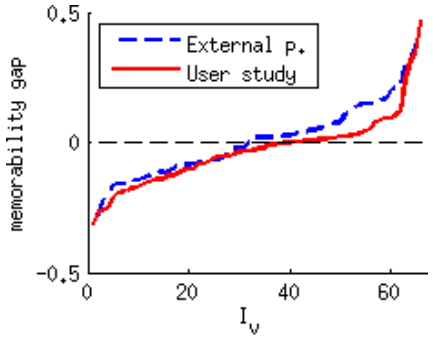

Fig. 8. Sorted memorability gaps for the image-seed pairs in the set $\mathcal{Z}$ measured using the predicted and actual memorability scores collected through the user study.

	A^X		MSE^X		ρ^X	
	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube	$\mathcal{B}$	S-cube
X = Sc	**72.73**	65.15	0.0155	**0.0138**	0.517^{***}	$\mathbf{0.570}^{***}$
X = Ev	**68.18**	63.64	0.0183	**0.0162**	0.479^{***}	$\mathbf{0.544}^{***}$
X = H	**60.61**	59.09	0.0188	**0.0153**	0.191^{*}	$\mathbf{0.453}^{***}$

Table 7. Performance of our method compared to baseline $\mathcal{B}$, evaluated on the set $\mathcal{Z}$, using the internal predictor (Sc), the external predictor (Ev) and the actual memorability scores collected through the user study (H). Performances are reported in terms of Accuracy (A), MSE and Pearson's correlation ρ. (***p-value < 0.005 and *p-value > 0.05).

by showing each image at different times of the game session. When a new player starts the game, a random sequence among the possible 100 is considered.

In order to recruit volunteers for the game, we sent invitation emails to several mailing lists of university groups and trusted associations. Before accessing the game, participants were asked to provide personal information, such as age, gender and nationality (the latter was optional). We recruited over 200 players, both male and female (respectively 42.3% and 57.7%), aged between 17 and 62 years old, and from a large variety of countries.[2] We made sure that each volunteer played only once and we discarded the results of players who did not qualify to the game by detecting less than 25% - *i.e.* less than 4 - of the vigilance repeats. Finally, for each of the 132 target images (66 original and the corresponding 66 stylized versions), we computed the actual memorability scores by aggregating the performance of over 80 players.

4.3.2 Results. The internal and external predictors are trained in natural images. One fair question is whether or not the outcome of these predictors is still valid for stylized images. In particular, we would like to assess the correlation between the external predictor (use to evaluate the proposed approach) and the human scores. To this aim we compute the values of the Pearson's correlation coefficient and MSE considering the automatically predicted and the actual memorability scores of the images in the verification set $\mathcal{Z}$. The Pearson's correlation[3] values corresponding the original and the stylized images in $\mathcal{Z}$ are respectively $\rho = 0.66$ and $\rho = 0.50$ (we recall that

[2] Austria, Belgium, Bosnia, Brasil, Chile, Croatia, Estonia, France, Germany, Grece, Italy, India, Iran, Latvia, Lithuania, South Korea, Malaysia, Mexico, Mongolia, Montenegro, The Netherlands, Poland, Portugal, Romenia, Russia, Serbia, Singapore, Spain, Sweden, Turkey, U.K, Ukraine, U.S.A., etc.

[3] We also computed the Spearman's rank correlation, which systematically leads to the same conclusions than the Pearson's correlation coefficient.

Fig. 9. (Left) Least and (right) most memorable abstract art paintings from LaMem dataset [21]. The actual memorability scores are reported below each image.

human performance is $\rho = 0.68$). These results demonstrate a high correlation between actual and predicted memorability scores. For the stylized set, a drop in performance can be observed which can be probably ascribed to the domain shift existing between original and stylized images. Indeed, the Scorer model is trained on LaMem dataset which does not include stylized images. In fact, while LaMem includes abstract art images, they represent a small subset. Thus, the learned memorability predictor will be obviously more accurate in the case of generic images. A similar trend can be observed for the MSE: 0.0121 for the original images and 0.0132 for the stylized ones. Indeed, a 16.7% increase in the error is motivated by the lower accuracy of the external predictor on the stylized set.

Furthermore, in Figure 8 we plot the sorted memorability gaps obtained with the external predictor and the user study for the 66 image-seed pairs in the verification set. Overall, a similar trend can be observed, with the external predictor which only slightly overestimates the memorability gaps.

Finally, in Table 7 we show the performances of our method compared to the baseline $\mathcal{B}$ in predicting the memorability increase for the images in our verification set $\mathcal{Z}$. Note that in this case, differently from previous experiments, for each image in the verification set we only have a single corresponding style and therefore we consider only the associated predicted memorability score. The performance is measured in terms of Accuracy, MSE and Pearson's ρ, and evaluated using the internal predictor (Sc), the external predictor (Ev) and the actual memorability scores collected though the user study (H). It can be seen that S-cube outperforms the baseline in all cases in terms of MSE and ρ. In the case of Accuracy, an advantage is observed for the baseline. A possible interpretation of these results is that, for this particular set of images, $\mathcal{B}$ probably performs better in predicting the direction of the memorability increase, while it still performs worst at estimating the absolute value of the variation. This is shown by the low Pearson's coefficient ρ, which reaches non statistically significant values in the case of H.

4.4 Qualitative Analysis and Discussion

We also performed a qualitative analysis in order to gain further insights on the link between memorability and style. Specifically, in this qualitative analysis we bring our focus on the link between memorability and complexity, which to the best of our knowledge has been poorly explored so far, compared to the link with other attributes like interestingness [12] or colors [42]. We discuss some visual results from two datasets, respectively abstract art images and the data used in our user study.

4.4.1 Abstract art and memorability. Our first study focuses on the set of abstract art images included in the LaMem dataset. These images are annotated with the actual memorability scores. The goal of our study is to analyze qualitatively the specific patterns which can explain differences in memorability. This image set allows us to focus only on the visual style of images, discarding high level semantics which naturally influence memorability (*e.g.* the presence of faces tend to make images more memorable). In details, we extracted from LaMem dataset the 280 abstract art paintings which are part of the Affective dataset [28]. From this set, we manually discarded the

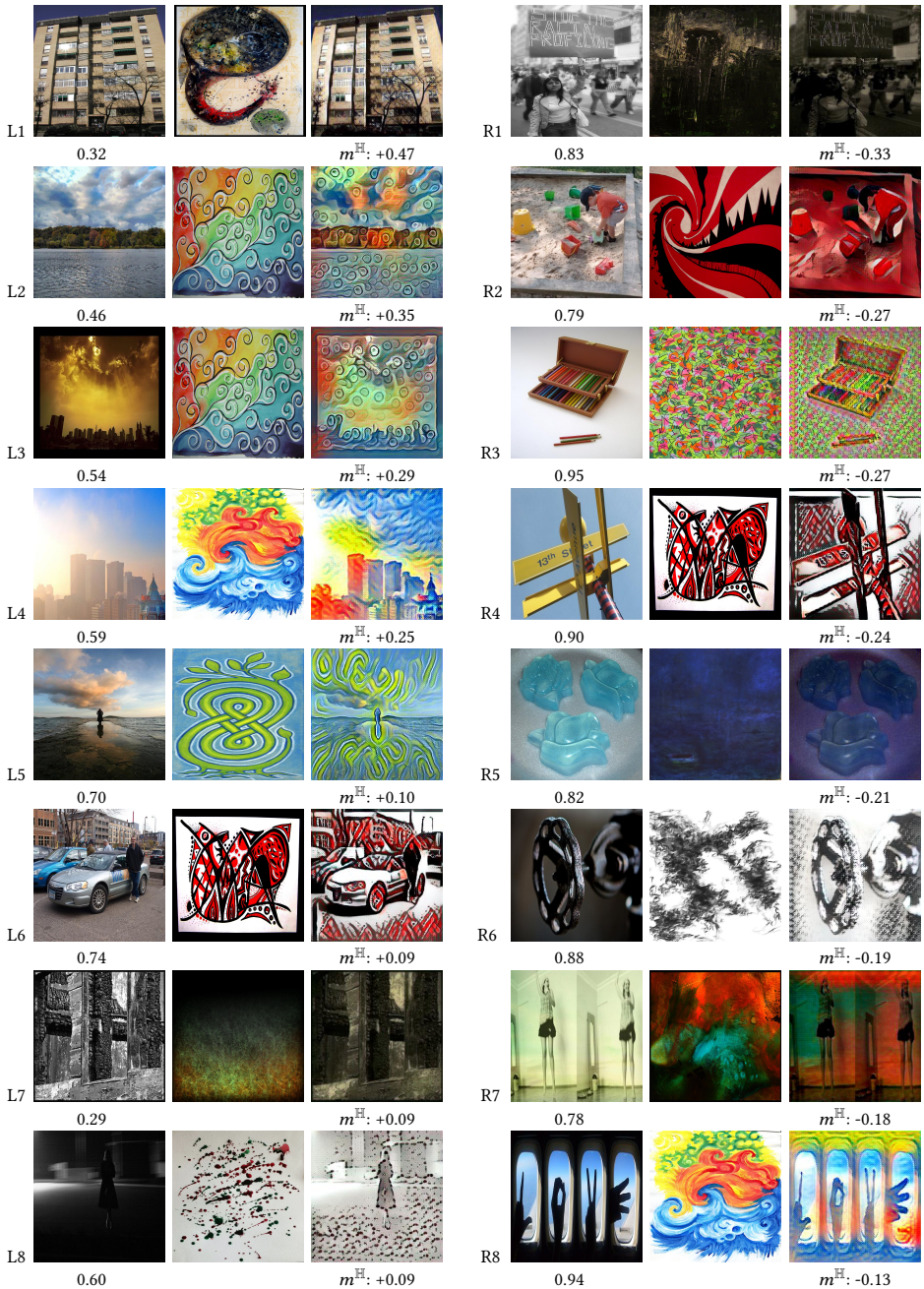

Fig. 10. Sample results of the user study: for each block (left) original input image, (center) style seed and (right) corresponding synthesized image. The actual memorability scores and gaps are reported respectively below each original and stylized image. Both cases where style transfer produces an increase (L1-L8) and a decrease (R1-R8) in memorability are shown.

paintings which contain people or objects, in order to completely cut out any semantic information. After this process we selected a total of 187 images, with a memorability score ranging from 0.37 to

0.98. In Figure 9 we show the five least and the five most memorable abstract art images, together with the corresponding memorability scores. It can be observed that images on the left exhibit certain common patterns (such as higher complexity, darker colors, absence of straight lines) while those on the right tend to be simpler and with more pleasant colors. This finding suggests that complexity and dark colors may have a negative influence on memorability. In other words, it appears that simple paintings, especially those with pleasant colors, tend to be more memorable. This may be explained with the fact that complex patterns require more time to be assimilated by the viewer. In the memorability game, each image is displayed for exactly the same amount of time: complex patterns probably get a lower chance to be decoded and remembered. This finding is in line with previous studies in the literature [4].

4.4.2 Increasing Image Memorability. In the third study we conduct a qualitative analysis on the images considered in our user study. Figure 10 reports some sample image-seed pairs for which we measured the largest memorability increases (L1-L8) and decreases (R1-R8). The collected memorability scores and memorability gaps $m^{\mathbb{H}}$ are reported below the original and the stylized image. In the cases of Figure 10.L2-L5, typical non distinctive urban and outdoor scenes are transformed into a corresponding more colorful and joyful version. These new stylized versions of the images are probably perceived with a higher valence, arousal and, definitely, lower naturalness. In the case of Figure 10.L6, the stylization clearly introduces complexity to the image. Still, the main objects in the scene (*i.e.* a man next to a car) are clearly visible. Similarly, in the case of Figure 10.L8, the stylization process highlights the presence of a person in the scene, thus probably increasing its memorability.

The samples in Figure 10(right) illustrate that only increasing strangeness is not a sufficient condition for increasing image memorability. The drop in memorability in most of these cases may be explained with the fact that stylization diminish the "readibility" of the image, thus reducing the possibility for the observer to decode and retain the visual information observed in only 1 second. For example, the text on the sign in the stylized versions of Figure 10.R1 is no longer readable, while in Figure 10.R2 the child in the playground is almost no longer recognizable. In these cases the stylization process reduces the semantic information of the image, thus diminishing its memorability. Similarly, in Figure 10.R8, the stylization introduces colorful elements but makes the objects in the scene harder to recognize. In Figure 10.R6, the stylized image is a sort of sketch of the original one. In this case a possible explanation for the memorability decrease is the fact that the black and white style makes images not particularly distinctive and hard to be remembered. Finally, it is interesting to observe that the same style applied to different images can induce opposite effects in terms of memorability variations. This confirms the validity of our framework where we select the best style for each given image, as there can be no universal style that is effective in memoralizing all images. The samples in Figure 11 correspond to the images obtained with *S*-cube. Specifically, we reported sample results where the optimal degree of stylization is automatically found to be $\alpha = 0.5$, 2 or 10, respectively for the top, central and bottom rows of Figure 11. The memorability increase is achieved by modifying the style of the images while retaining their original high-level content.

5 CONCLUSIONS

Visual memorability is a complex phenomenon which depends on multiple factors. In this work we showed that we can increase the probability of an image to be remembered by varying its style. Specifically, we presented a novel approach to increase image memorability based on an editing-by-filtering paradigm. In details, we proposed a deep learning framework made by three components: the Scorer, the Synthesizer and the Selector. The novelty of our approach relies on

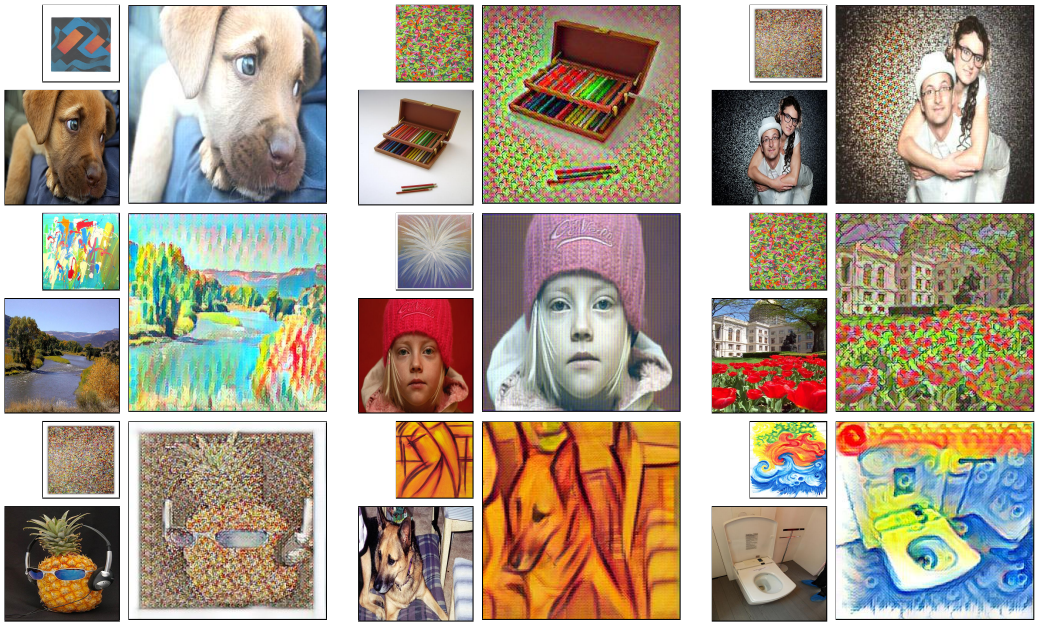

Fig. 11. Sample images obtained using S-cube: (top) $\alpha = 0.5$, (central row) $\alpha = 2$ and (bottom) $\alpha = 10$.

the fact that, given an input image, the Selector is able to automatically compute the style which guarantees the higher increase in term of memorability which is then provided to the Synthesizer for generating a stylized version of the original image. The effectiveness of our approach both in increasing memorability and in selecting the top memoralizable styles has been evaluated on a public benchmark.

While this work focused on memorability, our approach is highly flexible and we believe that, by replacing the deep network implementing the Scorer, it can potentially be applied to other perceptual attributes such as aesthetic quality or evoked emotions. Future works will be devoted to extend our framework in this direction.

Other possible extensions for this work are devising an approach for computing the degree of stylization α within a continuous range and designing a deep architecture where the style seeds are directly provided as input data to the Selector instead of considering only the memorability gaps. We expect that these modifications would lead to improved performance.

While the literature on predicting subjective attributes like memorability from visual data is quite huge, few works have focused on the problem of synthesizing images such as to modify these properties. In this work we considered neural style transfer techniques but we believe that other deep learning models, such as Generative Adversarial Networks (GANs) [10], can be exploited for this purpose. Future works will dig in this direction.

6 ACKNOWLEDGMENTS

We gratefully acknowledge Fondazione Caritro for supporting SMARTourism project and NVIDIA Corporation for the donation of the TitanX GPUs.

REFERENCES

[1] Peter P Aitken. 1974. Judgments of pleasingness and interestingness as functions of visual complexity. *Journal of Experimental Psychology* 103, 2 (1974), 240.

[2] Afsheen Rafaqat Ali and Mohsen Ali. 2017. Automatic Image Transformation for Inducing Affect. *British Machine Vision Conference (BMVC)* (2017).

[3] Daniel E Berlyne. 1960. Conflict, arousal, and curiosity. McGraw-Hill Book Company.

[4] Daniel E Berlyne. 1963. Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology/Revue canadienne de psychologie* 17, 3 (1963), 274.

[5] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2015. Intrinsic and extrinsic effects on image memorability. *Vision research* 116 (2015), 165–178.

[6] Alex J Champandard. 2016. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768* (2016).

[7] Russell Eisenman. 1966. Pleasing and interesting visual complexity: Support for Berlyne. *Perceptual and motor skills* 23, 3_suppl (1966), 1167–1170.

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Alvin G Goldstein and June E Chance. 1971. Visual recognition memory for complex configurations. *Attention, Perception, & Psychophysics* 9, 2 (1971), 237–241.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*.

[11] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. 2013. Visual interestingness in image sequences. In *ACM Multimedia*.

[12] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The Interestingness of Images. In *IEEE International Conference on Computer Vision (ICCV)*. 1633–1640.

[13] Raisa Halonen, Stina Westman, and Pirkko Oittinen. 2011. Naturalness and interestingness of test images for visual quality evaluation. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 78670Z–78670Z.

[14] Li He, Hairong Qi, and Russell Zaretzki. 2015. Image color transfer to evoke different emotions based on color combinations. *Signal, Image and Video Processing* 9, 8 (2015), 1965–1973.

[15] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *IEEE International Conference on Computer Vision (ICCV)* (2017).

[16] Juan Huo. 2016. An image complexity measurement algorithm with visual memory capacity and an EEG study. In *SAI Computing Conference (SAI), 2016*. IEEE, 264–268.

[17] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems (NIPS)*.

[18] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482.

[19] Aditya Khosla. 2017. Predicting human behavior using visual media. (2017).

[20] Aditya Khosla, Wilma Bainbridge, Antonio Torralba, and Aude Oliva. 2013. Modifying the memorability of face photographs. In *IEEE International Conference on Computer Vision (ICCV)*.

[21] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *IEEE International Conference on Computer Vision (ICCV)*.

[22] Aditya Khosla, Jianxiong Xiao, Phillip Isola, Antonio Torralba, and Aude Oliva. 2012. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*. ACM.

[23] Hye-Rin Kim, Henry Kang, and In-Kwon Lee. 2016. Image Recoloring with Valence-Arousal Emotion Model. In *Computer Graphics Forum*, Vol. 35. 209–216.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*.

[25] James L. McGaugh Larry Cahill. 1995. A Novel Demonstration of Enhanced Memory Associated with Emotional Arousal. (1995).

[26] David R Lide. 2018. Handbook of mathematical functions. In *A Century of Excellence in Measurements, Standards, and Technology*. CRC Press, 135–139.

[27] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep photo style transfer. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[28] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM international conference on Multimedia*.

[29] Stephen Maren. 1999. Long-term potentiation in the amygdala: a mechanism for emotional learning and memory. *Trends in neurosciences* 22, 12 (1999), 561–567.

[30] Weijie Mao Mengjuan Fei, Wei Jiang. 2018. Creating memorable video summaries that satisfy the userâĂŹs intention for taking the videos. 275 (2018).
[31] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[32] Elizabeth A. Phelps. 2004. Human emotion and memory: interactions of the amygdala and hippocampal complex. *Current opinion in neurobiology* 14, 2 (2004), 198–202.
[33] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2018. Artistic style transfer for videos and spherical images. *International Journal of Computer Vision* (2018), 1–21.
[34] Andreza Sartori, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. 2015. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM Transactions on Interactive Intelligent Systems* 5, 2 (2015), 8.
[35] Sumit Shekhar, Srinivasa Madhava Phaneendra Angara, Manav Kedia, Dhruv Singal, and Akhil Sathyaprakash Shetty. 2017. Techniques for enhancing content memorability of user generated video content. (2017). US Patent 9,805,269.
[36] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. 2018. Avatar-Net: Multi-scale Zero-shot Style Transfer by Feature Decoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[37] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2017. How to Make an Image More Memorable? A Deep Style Transfer Approach. In *International Conference on Multimedia Retrieval (ICMR)*.
[38] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR* (2015).
[39] Mohammad Soleymani. 2015. The quest for visual interest. In *ACM international conference on Multimedia*.
[40] Lionel Standing. 1973. Learning 10000 pictures. *The Quarterly journal of experimental psychology* 25, 2 (1973), 207–222.
[41] Lionel Standing, Jerry Conezio, and Ralph Norman Haber. 1970. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science* 19, 2 (1970), 73–74.
[42] Noah Sulman Thomas Sanocki. 2011. Color Relations Increase the Capacity of Visual Short-Term Memory. *Perception* 40, 6 (2011).
[43] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. *ICML* (2016).
[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[45] Wenguan Wang and Jianbing Shen. 2017. Deep cropping via attention box prediction and aesthetics assessment. In *IEEE International Conference on Computer Vision (ICCV)*.
[46] Hang Zhang and Kristin Dana. 2018. Multi-style generative network for real-time transfer. *ECCV Workshops* (2018).
[47] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *Advances in neural information processing systems (NIPS)*.

Received February 2007; revised March 2009; accepted June 2009