

Normalized Moments for Photo-realistic Style Transfer

TREVOR D. CANHAM

(B.Sc., Rochester Institute of Technology, United States, 2018)

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Graduate Program in Electrical Engineering and Computer Science
York University
Toronto, Ontario

May 2024

©Trevor Canham, 2024

Abstract

Style transfer, the operation of matching appearance features between source and target images, is a complex and highly subjective problem. Due to the profundity of the concept of artistic style, the optimal solution is ill-defined, so the variety of approaches that have been proposed represent partial solutions to varying degrees of efficiency, usability and appearance of results. In this work a photo-realistic style transfer method for image and video is proposed that is based on vision science principles and on a recent mathematical formulation for the deterministic decoupling of features. As a proxy for mimicking the effects of camera color rendering or post processing, the employed features (the first through fourth order moments of the color distribution) represent important cues for visual adaptation and pre-attentive processing.

The method is evaluated on the above criteria in a series of application relevant experiments and is shown to have results of high visual quality, without spatio-temporal artifacts, and validation tests in the form of observer preference experiments show that it compared very well with the state-of-the-art (deep learning, optimal transport, etc.) The computational complexity of the algorithm is low, and a numerical implementation that is amenable for real-time video application is proposed and demonstrated. Finally, general recommendations for photo-realistic style transfer are discussed.

Acknowledgements

My sincerest thanks to my parents to whom I owe everything, to my friends for keeping me from becoming dull, to Dr. Michael Brown for being a great advisor, to Dr. Richard Murray and Dr. Marcus Brubaker for being a great committee and to everyone who was a part of my life in Spain when this work began. In particular, thanks to Dr. Marcelo Bertalmío and Dr. Javier Portilla who trusted me with their ideas.

Contents

Abstract	II
List of Publications	V
List of Figures	VI
List of Tables	IX
1 Introduction	1
1.1 Motivation	1
1.2 Research Problem and Contribution	2
1.3 Report Organization	4
2 Preliminaries	5
2.1 Moments	5
2.2 Visual Sensitivity to Moments	9
3 Related Work	12
4 Using Decoupled Moments for Photorealistic Style Transfer	16
4.1 Method Overview	16
4.2 Transferring Decoupled Moments	21
4.3 Method usability	22
5 Experiments	25
5.1 Qualitative Comparison	26
5.2 Psychophysical experiment	32
5.3 Applications	33
5.3.1 Broadcast Source Homogenization	33

5.3.2	Video	34
5.3.3	Demo software	36
5.3.4	Color grading workflow case study	37
5.3.5	Recommendations & limitations	38
6	Summary and Future Work	42
6.1	Summary	42
6.2	Future Research Directions	42
	References	44

List of Publications

Trevor D. Canham, Adrián Martín, Marcelo Bertalmío, and Javier Portilla. “Using Decoupled Features for Photo-realistic Style Transfer”. In *SIAM Journal on Imaging Sciences*, vol. 16, no. 3, pp. 1687-1726, 2023, doi: 10.1137/22M1512491.

List of Figures

2.1	A visualization of the terms of equation 2.1 for various k values on a two dimensional zero-mean gaussian probability distribution $f(x)$. The symmetry/asymmetry of odd/even moments can be observed as well as the relative weighting of tone scale areas for each statistical moment.	6
2.2	The effect of decoupled moments on image appearance and distribution shape.	7
4.1	Comparing the results between each moment the transfer is progressively refined, bringing the source closer to the target image, emphasized in the zoomed regions of every second row. From left to right: original source image; source mean matched to target on far right; source mean and variance matched; source mean, variance and skewness matched; source mean, variance, skewness and <i>ortho-kurtosis</i> matched; target image. First and third example source and target provided by ARRI [1], second example source and target from [2], last example source and target made available by NASA under CC-BY 2.0. Second row source and target from [2]. Fifth and seventh row source and target provided by ARRI [1].	18
4.2	Three-step feature transfer using de-normalization. The extracted features (concentric layers, here labeled as the luminance's decoupled moments) capture the style of the image, while the normalized image (central circle) has no style information. Here the target's features are visualized by means of a random image with the same features ("style without scene"). Conversely, the normalized source has pre-fixed values for those features ("scene without style"). The transfer process combines these components to obtain the result.	19

4.3	Process for converting radiance (display linear) images with sRGB primaries to IPT space, where we use the symbol “ $\odot$ ” for representing an element-wise scalar operation for the vector coefficients (here a power).	21
4.4	From left to right, varied source images, target images, and the results of the default configuration of our method. First row source and target images, and the second row target image uploaded to Flickr by Kristoffer Trolle (CC-BY 2.0). Second row source images uploaded to Flickr by Mk Feeny (CC-BY 2.0). Third row images sourced from [2]. Bottom row images provided by ARRI [1].	23
4.5	From left to right, source image with cropped region bounded in green, target image with cropped region bounded in green, default method transfer result with cropping, default method transfer result without cropping. Images sourcing left to right, top to bottom: European Southern Observatory (ESO), European Space Agency (ESA), ARRI Camera Group (complete second row), Labelled Faces in the Wild (complete third row) [2], MK Feeny, ESA. Licensed via CC-BY 2.0.	24
5.1	Experimental image test set, featuring highly chromatic and stylized scenes. Fourth image made available by TVR Romania, the remainder are made available by Pexels.	26
5.2	From left to right, source image with ARRI WCG <i>LogC₄</i> encoding, target in Rec. 709 encoding, the result of Ke et al. [3], demonstrating that the method’s constrained transform cannot replicate basic non-linear transforms common to motion picture imaging pipelines. The results of the proposed method are shown for comparison.	27
5.3	From top to bottom, sources, targets, proposed result, Ke et al. [3], Luan et al. result [4], Li et al. result [5], Mechrez et al. result [6] (taking Luan et al. results as input), Huang result [7].	28
5.4	From top to bottom, sources, targets, proposed result, Rabin result [8], Ferradans result [9], Blondel result [10].	29
5.5	From top to bottom, sources, targets, proposed result, Li result [11], Wu result [12], Xie result [13].	30
5.6	From top to bottom, sources, targets, proposed result, Grogan result [14], Photoshop result [15], and Zabaleta & Bertalmío result [16].	31

5.7	Broadcast application example. From top to bottom, user source images, studio targets, control (no style transfer applied), proposed result, Grogan et al. result [14], Photoshop result [15], and Zabaleta & Bertalmío result [16]. The user images are sourced from [2], the red studio image was provided by Premiere sports, and the blue studio image was provided by TVR Romania.	35
5.8	Demo software GUI.	37
5.9	The “hero” look, which defines the color processing for the whole project is derived in the top row from source and target stills. The resulting LUT can then be applied to a variety of scenes.	39
5.10	Additional corrections to avoid jarring cuts between shots with different lighting conditions. In the bottom row, the analysis of the source and target is limited to the green boxes through cropping.	40

List of Tables

5.1	Psychophysical experiment results, where the percentage which the method on the column label was preferred over the method on the row label, averaged over 20 image combinations and 19 observers.	33
-----	--	----

Chapter 1

Introduction

1.1 Motivation

Cinema colorists, whose job it is to beautify images after they have been shot, manually explore a multi-dimensional space of image states to find novel visual experiences. Generally speaking, these tools are comprised of simple mathematical operations applied to individual color channels (like additive offset, multiplicative gain and exponential functions), bear similarity to various adaptive operations in the human visual system (HVS) and in the context of the standard dynamic range (SDR) format are good approximations of these behaviors. However, as the dynamic range of capture and display devices increases, so does the need for tools which target more specific areas of the tone scale (the range of intensity levels from black to white) as the different luminance regimes become more perceptually distinct [17].

A common goal of many color correction processes is to reflect the color and texture rendering behavior of motion picture film. More generally, inspiration for how new images should be styled is most commonly derived from pre-existing imagery of the same format, although there are always exceptions where paintings or real life visual experiences are employed as targets. With that said, a tool which would allow colorists to directly match the visual impression of a target image for use as a starting point should be highly desirable. Such tools have long been proposed, but too often fail to meet the high image quality standards of cinema or are otherwise unusable by colorists. Their absence in current industrial workflows in spite of their abundance in the research community is evidence to this point.

1.2 Research Problem and Contribution

The umbrella term of style transfer (ST) refers to the class of methods in image processing where appearance characteristics of a target image are extracted and replaced in a source. An artist’s style is difficult to define because it can be comprised of both tangible and intangible features. We have the ability, for example, to model the effect of a camera or lens, or even brushstrokes and canvas material properties. Image features can even be mapped to abstract representations which mimic those of a body of work. However, the glue that holds everything together is the artist’s intangible embodied experience. It is through this lens that they conduct their selective process, connecting and incorporating disparate influences and omitting distractions. Even if personal history, preferences and previous work are known, such a filter would be difficult to predict. Automated tools run the risk of hurting the art community by trivializing this contribution, and so it should be stated that this is an ill-posed and highly subjective problem without the possibility of achieving an optimal answer. Instead, there are only incomplete answers which vary in their usability for a particular application.

The original works in this area, from Kotera et al. [18] and later Reinhard et al. [19], aimed to accomplish ST simply by imposing the mean and standard deviation of the target image onto the source. In doing so, the authors could achieve a new version of the source image which reflected some of the target properties like color balance, luminance level, and contrast using simple algebraic operations. This was a promising tool because its computation was extremely simple and its effects were predictable and robust (with respect to the variety of usable source/target combinations), allowing artists to learn how to use it without becoming frustrated by frequent artifacts. However, in most cases the targeted features failed to describe the style characteristics users were expecting to transfer, like high level scene details (e.g. changing seasons, like the drought effect achieved in the movie “O Brother, Where Art Thou?”, which was the first film with a fully digital color correction). Also, contrast and brightness have always been easy for professional colorists to control manually, so the novelty was not sufficiently significant to justify a change in workflow.

Moving forward, Morovic and Sun [20], and later Pitié [21] proposed the use of optimal transport (OT) methods, which aimed to derive a color mapping between two distributions. In some special cases, the methods were able to produce strong

“alternate reality” effects, allowing the approach to gain popularity. However, the complexity of the resulting maps often produced artifacts like heavy banding and gamut clipping, prohibiting the approach from achieving the robust performance of the preceding moments transfer strategy. This marked the beginning of a trend in the field where sensational results were prioritized over usability in a general context, resulting in a departure from the needs of the cinema and photographic communities.

This progressed further with the introduction of neural style transfer (NST) by Gatys et al. [22], whose method employed convolutional neural networks (CNNs) to extract a large number of low level features from a target image and impose them onto the corresponding features of the source. This approach worked especially well in replicating modern painting styles (e.g. impressionist, cubist, surrealist), and the results generated a wave of popularity which dwarfed that of OT and continues to the present day, with scores of papers published yearly attempting to advance the application. While this development has opened up a novel and fruitful dialogue between art and technology fields, the fundamental problems of these methods (and those of machine learning in general [23]), which make them difficult to use by artists in a photo-realistic context, also persist: a stark loss of performance outside of a few limited examples, and a lack of predictability and corrective recourse due to their black box nature.

Generalizing across these major directions in the field, the primary design element which separates one method from another is the selection of features to isolate and transfer. In terms of quantitative comparison the majority of works only address computational efficiency, rendering some feature sets more desirable than others. However, they also have an affect on a number of usability criteria. A major example is that of the interpretability of features or the degree to which features are correlated to meaningful, tangible elements of the art generation process (e.g. brushstroke styles or photographic filters), making them easier or harder to identify in a target image, predict their appearance in the context of a source and to conceptualize and integrate into a pre-existing workflow. This property also relates to the degree to which the tool can be parameterized or made modular such that artists can customize or refine the results for their purposes.

Features also differ in how easily they can be divorced from the specific context of the target and re-appropriated for the source. Methods are expected to work in such a way that intense, identifiable characteristics of the target can be transferred

to the context of the source without carrying any residual fingerprint of the target’s scene content. Feature intensity is often a trade-off with method reliability as more constrained transforms are less prone to artifacts.

The work in this thesis presents an ST framework which obtains a small set of interpretable feature descriptors, each of which is linked to fundamental parameters of visual adaptation and response. Following the example set by Reinhard et al. [19], the statistical transfer approach is extended to include higher order moments skewness and kurtosis, which has not been attempted previously because they are highly correlated. However, by making use of a methodology from Martinez-Enriquez et al. [24] transforms are found along an orthogonal pathway on their respective feature manifolds which aim to objectively match these target features in the source while making minimal change to the image, avoiding artifacts. The hypothesis is that these features will be highly interpretable since they play an important role in the visual processing of scenes and are present in the common lexicon of image quality evaluation. In doing so, a robust tool is presented with which artists can build familiarity and can be integrated into their workflows. To facilitate this we provide a standalone GUI software which allows for modular control of transferred features, control of image regions for analysis via cropping, and 3D LUT generation such that color and tone transforms can be extracted and imported to any professional image editing platform.

1.3 Report Organization

In what follows, an overview of the background concepts of statistical moments and their relevance to visual processing is given in Chapter 2. Next, the related works in the area are addressed in greater depth in Chapter 3. Following this, the proposed method is described in Chapter 4. Next, a series of experiments to validate the proposed method in context of the related work and the application described above are described in Chapter 5. Finally, the work is summarized and future work is discussed in Chapter 6.

Chapter 2

Preliminaries

This chapter provides an overview on the concept of moments (what they are and how they can be normalized) and their application as features for image analysis and vision modelling.

2.1 Moments

The shape of a function can be roughly described in terms of its moments. When applied to different domains, these shape descriptors take on a specific meaning, representing physical properties. The generalized definition of a given moment M_k of some function $f(\vec{x})$ about its average value μ_x is as follows.

$$M_k = \int_{-\infty}^{\infty} (\vec{x} - \mu_x)^k f(\vec{x}) d\vec{x} \quad (2.1)$$

This concept is illustrated in Figure 2.1 where the terms of equation 2.1 for various k values are plotted, taking a two dimensional zero-mean gaussian distribution as $f(\vec{x})$.

Most people are familiar with the zeroth, first and second order moments (the sum, mean, and variance of a function respectively) and can relate them to phenomena in various domains of analysis (e.g. economics, demographics, etc.) When applied to color images, these moments are used to analyze the probability distribution, or the number of pixels present at each intensity level on a color channel. Here, the first and second moments relate to the brightness and contrast of the image respectively. The fact that these concepts are generally understood is reflected in that they are assigned to picture quality controls on consumer devices, which affect the image roughly by adjusting the first and second order moments through additive offset and multiplicative

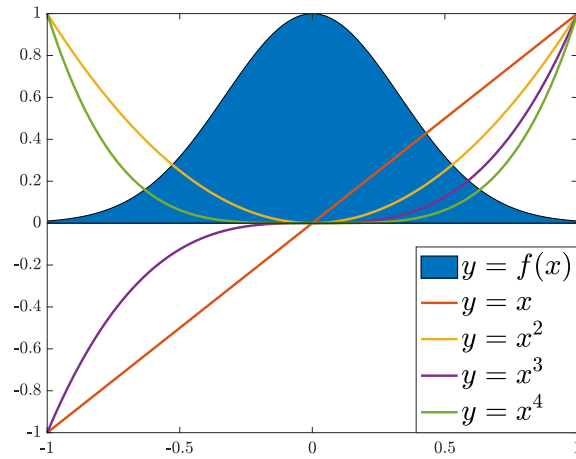


Figure 2.1: A visualization of the terms of equation 2.1 for various k values on a two dimensional zero-mean gaussian probability distribution $f(x)$. The symmetry/asymmetry of odd/even moments can be observed as well as the relative weighting of tone scale areas for each statistical moment.

gain controls.

Higher order moments like the third (skewness) and fourth (kurtosis) are harder to interpret, as they progressively target smaller areas of the distribution towards the tails and become more correlated to each other. Skewness, for example, is a measure of the asymmetry of a distribution about its mean. Distributions which have a high (positive) skew value will lean towards the lower end of the domain and will have a stretched out tail towards the high end, while a distribution with the same mean but a negative skew will have the opposite orientation. While only present in the vocabulary of specialists, the skewness of image values may be discussed in terms of key (low key being an image with a positive skewness where most of the distribution is concentrated in the shadows or high key being an image with negative skewness where most of the distribution is concentrated in the highlights. Odd moments are highly correlated to each other, so the skewness of a distribution is highly correlated to the mean. In imaging terms, key is highly correlated to brightness (in practice, a low key image will generally be darker than a high key one) so the two image properties are challenging to differentiate unless adjusted in a decoupled manner with the proposed method (Figure 2.2).

Kurtosis, on the other hand, deals with the sum size of the tails on the higher and lower ends of the distribution, as opposed to skewness' relative size of one versus the other. Generally this is used as a descriptor of the relative presence of outliers in

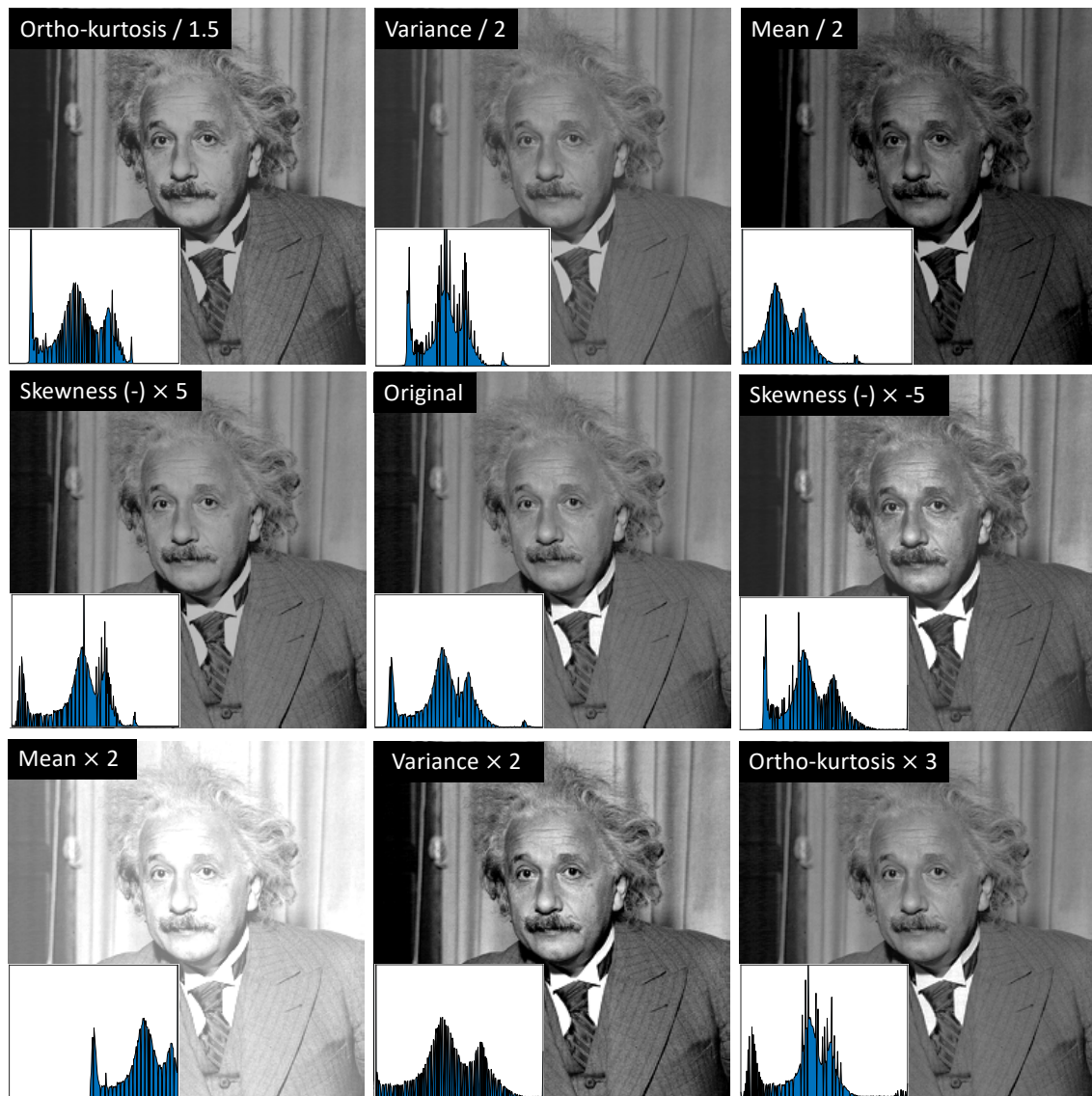


Figure 2.2: The effect of decoupled moments on image appearance and distribution shape.

a dataset, so in many analysis applications this value is not particularly relevant as outliers are often ignored. However, due to the discrete range of images caused by the sensitivity limits of capture devices and the reproduction capabilities of displays, large concentrations of pixels at the white and black limits are common. In fact, image editors commonly clip (fix to the maximum and minimum) large regions of image detail to simulate high contrast visual experiences on displays with limited dynamic range, demonstrating their relevance to image analysis.

The asymmetry of the first and third order moments and the symmetry of the second and fourth order moments can be explained by their relative polynomial order (visualized in Figure 2.1). This means that mean and skewness preserve the sign of the input, whereas variance and kurtosis values are always positive. Statistical moments with arbitrarily high k values behave in a similar way to the third and fourth order moments (respectively, for even and odd k values), but emphasize $\vec{x}$ values further from the mean as the asymptotic region about the origin of polynomial functions increases with order. For this reason, higher than fourth order moments are rarely used in analysis and do not have names like the lower ones.

Observing Figure 2.1 the range of sensitivity of the successive moments overlaps heavily. Since coupled features are inherently inefficient, successive normalization steps must be taken such that they are unaffected when working with one another (i.e. their gradients throughout the space of possible input vectors $\vec{x} \in \mathbb{R}$ are orthogonal such that $\nabla \vec{x}_J \times \nabla \vec{x}_j = 0$ for moments J and j). For example, the mean is to be adjusted additively (so as to not affect the zeroth order moment of sum) to a normalization point of zero given an input vector $\vec{x}$, where it will be unaffected by multiplicative operations related to the variance. Similarly, the variance has a natural normalization point of 1 where it can remain unaffected by exponential operations related to the higher order moments, whose normalization points follow a similar pattern to the first two moments corresponding to their even/odd parity. For example since skewness can be positive or negative, its natural normalization point is 0 like the mean, and is found by numerically searching for a function that returns this result for $\vec{x}$. The resulting normalized distribution is balanced about this mid point.

The relationship between the third and fourth standardized central moments is more complex than the lower order moments because of various incompatibilities (e.g. a high sample skewness value is algebraically incompatible with a low sample kurtosis value [25]). A number of authors propose alternative measures of kurtosis which are

approximately free from these complications and are more independently meaningful as a result, which attempt to locate some minimum point on the kurtosis gradient in vector space by taking advantage of various patterns and relations uncovered through study of the feature space [26, 27]. However, these measures are usually only decoupled in a limited subspace of possible input vectors. In response to this, Martinez-Enriquez & Portilla propose a general methodology which, given a set of coupled features, produces a set of similar features but with the property of having mutually orthogonal gradients throughout the whole domain (in favorable cases, like the moments) [28]. This method is employed here in the context of style transfer to increase the efficiency of the feature transfer process and the descriptive power of the chosen features in comparison to the standardized central moments.

2.2 Visual Sensitivity to Moments

Vision scientists, inspired by the field of informatics, have devoted significant attention to the problem of identifying the information processing constraints of the human visual system (HVS) in order to represent information with minimal bandwidth. The resulting models and experimental methodologies have been successfully applied to numerous applications. A popular example is the measurement of basic perceptual units of texture such that it can be generated from simple instructions. This work is intimately related to the problem of style transfer, as it requires the definition of content-invariant primaries of visual appearance to be measured and swapped between images.

A strong backing exists for the use of statistical moments as features in the texture segmentation and visual search literature. In the foundational work of Julesz [29], differences in the mean and variance of noise patterns were found to be “immediately and effortlessly perceptible, without scrutiny” in comparison to more complex features which required timescales in the range of cognition to discriminate. These features which immediately “pop out” when observers are presented with a visual search task have been classified as pre-attentive visual processing [30]. This stage involves a parallel survey of the entire visual field to quickly segment a scene and direct attention towards regions of interest. Anecdotally, this can be related to the experience of our attention being caught by objects on the road when driving. Mis-identification and hallucination phenomena suggest this is a separate stage of processing that is fed by

depleted, summary information, allowing for rapid parallel processing.

Chubb et al. [31] propose a model in which neural adaptation processes extract summary statistics of the luminance distribution of scene regions such that they can be grouped together and segmented to direct attention. Specifically, they extend Julesz’s model to include mechanisms which are not only sensitive to the mean and variance, but also propose a third mechanism which is sensitive to higher order moments. The significance of higher than second order moments of the luminance distribution is also reflected in the work of Pratt et al. [32] who found that with alternative statistical techniques than the Markov-based strategy used by Julesz, patterns which differ only in skewness while keeping the first two statistical moments the same could be discriminated by observers. This is supported by the recent discovery of neural structures which are sensitive to skewness [33]. The functional explanation for this sensitivity provided by Chubb et al. is the separation of highlight and shadow regions, facilitating general foreground/background segmentation. If the image display is a relevant sub-region of the viewer’s scene segmentation processes, the global statistics of the image should be highly correlated to the viewer’s pre-attentive segmentation of the screen from the viewing environment.

While the features of pre-attentive response are not limited to the moments of the intensity distribution, the significance of this feature is corroborated in the context of texture discrimination via experiments with Portilla & Simoncelli’s method for texture synthesis [34]. The model employs four sets of parameters to generate images including the marginal moments up to kurtosis, the local auto-correlation, filter magnitude correlation and cross-scale phase statistics. Balas [35] generated stimuli in which one of these parameter sets is lesioned (i.e. ignored in synthesis) and shown to observers in a discrimination task versus the texture generated with the full set. In all experimental variations (representing periodic, structured and 3D asymmetric texture families), stimuli pairs with lesioned moments were the most noticeable by a significant margin.

Following the assumption that pre-attentive responses represent summaries of neural adaptation, the method’s behavior can alternatively be conceptualized as transferring the global adaptive response from the target image to the source. The foundational works of Attneave [36] and Barlow [37] suggest that neural structures in the visual pathway are optimized to accommodate the statistical properties of stimuli which we are most commonly exposed to (i.e. natural scenes). A thorough review of this concept is provided by Simoncelli & Olshausen [38].

In summary, the human visual system has a limited bandwidth and energy budget with which to transmit information, and reduces redundancy by decorrelating signals and distributing them across pathways and response levels. One specific example of this behavior is the separation of visual signals into positive (ON) and negative (OFF) contrast channels, roughly about the midpoint of the distribution [39]. This value is the most commonly occurring response level in normal distributions, so by assigning it a null response, or low intensity responses for values closest to it, the visual system's encoding is further optimized. Given the rapidly changing stimulus distribution due to saccades (rapid eye movements) and changing scenery, this redistribution of signals must be accomplished quickly through adaptation. This implies that the intensity distribution of the stimulus (which can be more efficiently characterized by its moments) is a strong predictor of some part of the visual system's adaptive response. So if a viewer is presented with two images where these properties are matched, they should experience no shift in that particular adaptation process when they are presented in sequence.

With the various uses for the moments in the visual system and efficient encoding in mind, the existence of similar normalization routines to those presented in section 2.1 are plausible in the HVS. This is observed in the example of signal splitting between negative and positive contrast pathways. Higher order adaptation processes like those in response to contrast may similarly find a normalization point at which more detailed information can be efficiently extracted. For this reason a representation in which the moments of a signal are normalized is a plausible approximation of adaptation to that signal, while the decoupled extracted moments are a plausible approximation of pre-attentive signals.

Chapter 3

Related Work

In the most general sense, ST methods work by measuring features in source and target images and matching the source features to the target. An early foundational example is the method of Reinhard et al. [19], which first transforms images into a decorrelated color representation (defining luminance L and chroma α, β groups), then extracts the mean and standard deviation values from each channel in the source and target and finally matches them via equation 3.1, where C corresponds to the color channel, T and S correspond to target and source image, (i, j) signifies the current pixel location, and σ and μ correspond to the standard deviation and mean, respectively.

$$C'_{(i,j)} = \sigma_T^C / \sigma_S^C \times (C_{s(i,j)} - \mu_S^C) + \mu_T^C \quad (3.1)$$

Aside from the proposed method, a number of approaches have recently been posited which are effectively extensions of the statistical transfer concept of Reinhard et al. However, instead of extending the transfer mechanism, they add pre-processing stages which isolate image regions into groups to be passed through equation 3.1. For example, in the method of Wu and Xue [12], the image is segmented into salient and non-salient features and the mean and standard deviation are matched between the independent regions. Another instance is the method of Xie et al. [13], where the image is segmented into color object regions by way of K-means clustering and the statistics are matched between the corresponding clusters of the target and source.

Instead of isolating and matching a small number of significant features, neural style transfer (NST) addresses the problem in a comparatively brute force approach, where a large number of trivial features are extracted and the corresponding ones are matched between target and source. NST is by far the most popular avenue of study

in the field, with the proliferation of works in recent years exceeding that of any other by at least an order of magnitude. However, if one takes a closer look at these works it can be observed that they are for the most part incremental, experimenting with different flavors of network architecture, training strategies, feature derivation, or loss functions.

A recent review from Jing et al. [40] also demonstrates that almost none of these works attempt to produce photo-realistic results. The review also points out the common drawbacks of these methods which are summarized as follows. First, the methods must be trained, requiring significant storage and computational resources and resulting in a high economic and environmental cost at scale, with some methods recommending the use of 50 or more graphics processing units [41]. As a result, the photo-realistic results of many existing methods cannot be replicated because their code bases, photo-realistic models or datasets are not publicly available. Second, they are overfitted to their training sets, such that inputs which deviate from the set will lead to unacceptable results considering professional quality standards. Third, after being trained, many methods require the input images to be segmented and semantically labeled which can be difficult to accomplish [42, 43]. Finally, methods in this category are black boxes because their features are too many and too insignificant, and thus are difficult to parameterize and are complicated to tune in any meaningful way to refine results or correct artifacts. For representative results of the artifacts that can occur, the reader is referred to the experimental section (Figure 5.3).

After encountering these caveats, five significant methods which offer photo-realistic results and could be tested in this context were identified. Among them is the method of Luan et al. [4], which made modifications to the nature of the feature transfer process used by Gatys et al. [22] to maintain integrity and avoid “painterly” artifacts. Later, Li et al. [5] offered an advancement to this work, where the stylization process was followed by a smoothing step to ensure spatial consistency. Another notable recent work was that of Mechrez et al. [6] which sought to refine the output of other neural style transfer techniques by reinforcing the boundaries of the original image. The work of Huang et al. [7] also constitutes a significant contribution to photo-realistic NST. Despite the fact that they tailored their application for image translation, a slightly different topic which aims to transfer the general content of an input image to a source (not only “style”), their work demonstrated the capability to make photo-realistic transfers of the variety explored in this paper.

More recently, Ke et al. [3] train a 3×3 color normalization mapping between sets of edited versions of each input image and use the source mapping to normalize the input and the target mapping to de-normalize, returning the output. The results reflect the global chromatic content of the target applied to the source, and artifacts are avoided by constraining the transform to a linear color mapping and limiting the training set to artist approved renditions. However, this constraint results in an inability to accomplish transforms between color encoding spaces which is a common motion picture color correction operation, demonstrated in Figure 5.2. They also offer the contribution of the two-step normalization and de-normalization process, but it should be noted that this is not novel (it appears already in the work of Reinhard et al. [19]) or intrinsically meaningful as it is tied to its training set’s range of styles sourced from a collection of LUTs.

The works of the optimal transport (OT) category are similar to NST in the sense that they make use of a loss function to minimize differences between source and target images, but differ in that they create a direct general mapping function to transfer between color distributions instead of matching features. Morovic and Sun [44] presented the original OT method for ST, which created a complex (one-to-many) mapping between color clusters in the source and target images. The method presented by Pitié et al. [21] expands this concept to a form which is reflected in today’s state-of-the-art methods, where the task is broken down into successive simultaneous 1D distribution transfer problems, and the simplification to a limited number of data clusters is no longer necessary.

A thorough, recent survey of the topic is also provided by Pitié [45], where the state-of-the-art OT methods, as well as those of other approaches, are tested and compared. Included among them was the method of Ferradans et al. [9], which is based on a regularized OT framework that relaxes the constraints on size consistency between source and target datasets and also adds a penalty to preserve the image structure in the transform. The method of Rabin et al. [8] advances on this relaxed and regularized discrete OT concept by including a pre-processing step for modeling the spatial distribution of colors within the image domain, and a process for tuning the relaxation parameters automatically. Another notable recent method is that of Blondel et al. [10], whose regularization of the OT process allows for smooth approximations between sparse color groups.

Like the methods of optimal transport, the method of Grogan et al. [14] also works

on the derivation of a color mapping between two distributions, but instead of using an OT framework it finds an $L2$ optimized thin plane splines transform to perform the color transfer. A similar recent method from Li et al. [11] segments the image based on region saliency and color clusters, and creates detail preserving dynamic lookup tables to form a mapping between image segments of the target and source.

Other methods exist which are highly relevant to our application but do not fall under the categories listed above such as the “Color Match” function of Adobe Photoshop [15]. The work of Zabaleta and Bertalmío [16] should also be highlighted, which transfers the global color appearance of a style target onto a source image via a three stage vision model. First the luminance properties are transferred via a constrained histogram equalization process, then the color properties are transferred through principle component analysis, and finally a transfer of local contrast properties is accomplished. While the method was originally tailored for cinema applications using unedited camera raw images as sources, it can easily be adapted for transfer between arbitrary source and target images.

Many works, particularly among the NST group, express a desire for or functionally require a loss function or metric to judge the accuracy of style transfer. However, given the profundity of the concept of artistic style, the inherently subjective evaluation of results, and the wide range of potentially acceptable solutions this pursuit seems somewhat misguided. In general, this is a somewhat circular process, where metrics simply judge performance based on the authors own selection of style defining features.

The proposed approach offers benefits in the perceptual relevance, simplicity and quality of features. The features (the marginal moments of the color distribution) are few, and perceptually distinct (their relevance to texture discrimination has been demonstrated [35]), making them interpretable for artists to identify in transfer pairs. The method includes a normalization process which is meaningful in a perceptual sense, as it reflects the behavior of adaptive processes, as well as in a mathematical sense, as it allows for the decoupled extraction of our chosen features. The proposed method does not require a computationally expensive training process and is not dependant on any training set. The quality of its results are also high, resulting in minimal artifacts, juxtaposing the lack of robustness observed in the results of many of the above methods when evaluated outside of their test set, demonstrated in Chapter 5).

Chapter 4

Using Decoupled Moments for Photorealistic Style Transfer

This section provides an overview of the proposed method’s design, its motivation and its default configuration. Then, the process for transferring decoupled moments is outlined. Finally, the results of the default configuration and the method’s modularity are demonstrated.

4.1 Method Overview

The crux of the style transfer problem comes in decoupling the chosen style features from the content differences between the target and the source. Oftentimes the expectation is that they must have common or analogous elements. Our intended application is to allow users to automatically develop transforms which emulate the appearance induced by the color rendering and editing process of the target in the context of the source. The direct approach to accomplish this would be a physical simulation of these processes, where their effect on the source can be inverted and the target process applied. Unfortunately the specific details of these processes are unknown and cannot be easily predicted from the resulting image as effects of lighting, content, optics, in-camera and post processing can all potentially be mistaken for each other. Instead, the current aim is to transfer their sum effect on the global visual impression of the target to the source.

Taking inspiration from the vision science literature outlined in Chapter 2, we accomplish this by employing decoupled moments of the input color distribution. Ef-

ficient representation theory demonstrates that marginal moments can be predictive of adaptation processes as they result in the normalization of signal distributions with respect to these measures [36, 37, 39, 38]. Pre-attentive visual processing points to the moments as the highest weighted feature in rapid parallel texture segmentation [29, 32, 35, 33]. Considering the fact that the image display is something that the visual system should clearly isolate (and is in fact bounded by the display hardware itself), moments, queried globally over the image strongly correlate to our rapid adaptation to the display. This direct visual interpretability makes them a useful feature for defining style, and their global nature results in minimal dependence on local scene structures.

In Figure 4.1 the effect of transferring each moment on image color appearance is demonstrated. The hierarchical nature of the moments can be observed here, with the first transfer, the mean, having the greatest impact of all and each subsequent transfer having a progressively subtler effect. Still the impact is noticeable, as the general color distribution properties of the image become closer to the target. The variance impacts the contrast, the skewness impacts the balance of the image towards the shadows and the highlights resulting at times in the clipping or crushing of detail, and the ortho-kurtosis parameter acts as a refining step, separating the concentrations of the absolute darkest and brightest pixels in the image from the highlights and shadows. In this way, the decoupled adjustment of these features is useful for exercising more precise control over different areas of the tone scale range, an increasingly important trait in modern color correction as this range is gradually being extended by emerging displays and formats.

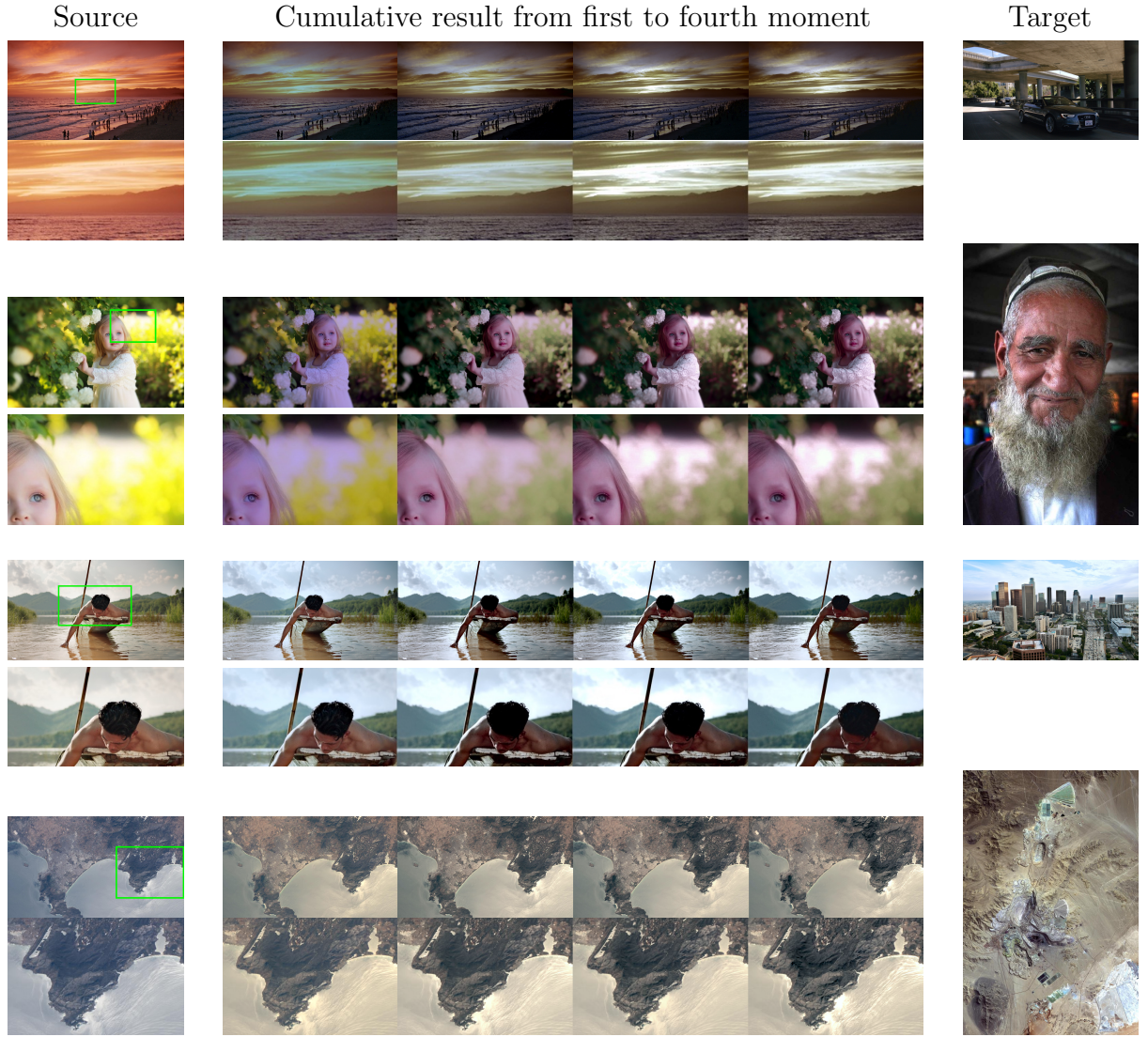


Figure 4.1: Comparing the results between each moment the transfer is progressively refined, bringing the source closer to the target image, emphasized in the zoomed regions of every second row. From left to right: original source image; source mean matched to target on far right; source mean and variance matched; source mean, variance and skewness matched; source mean, variance, skewness and *ortho-kurtosis* matched; target image. First and third example source and target provided by ARRI [1], second example source and target from [2], last example source and target made available by NASA under CC-BY 2.0. Second row source and target from [2]. Fifth and seventh row source and target provided by ARRI [1].

Due to the fact that moments become increasingly correlated at higher orders, special care must be taken to decouple them. It is for this reason that this seemingly

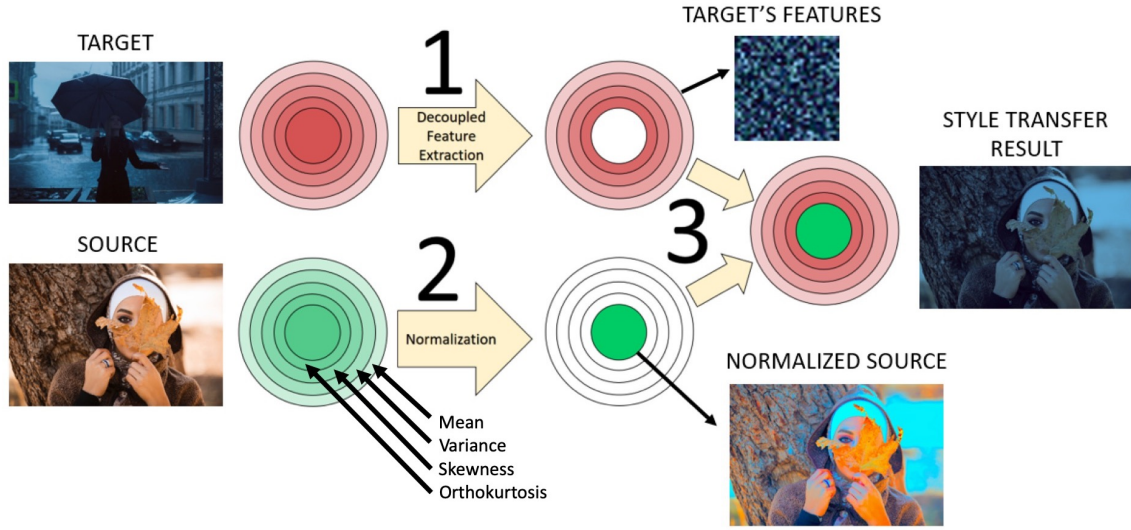


Figure 4.2: Three-step feature transfer using de-normalization. The extracted features (concentric layers, here labeled as the luminance’s decoupled moments) capture the style of the image, while the normalized image (central circle) has no style information. Here the target’s features are visualized by means of a random image with the same features (“style without scene”). Conversely, the normalized source has pre-fixed values for those features (“scene without style”). The transfer process combines these components to obtain the result.

obvious extension of the classical style transfer has not yet been explored in the literature. However, a recently proposed mathematical formulation of feature decoupling from Martinez-Enriquez et al. enables this to be accomplished [24]. Given a set of features, this methodology defines a similar set which has the property of mutually orthogonal gradients. While these features are not equivalent to their coupled counterparts, they are favorable for the given application because they increase the efficiency of the feature transfer process and the descriptive power of each individual feature, which Martinez-Enriquez & Portilla demonstrate for the applications of texture classification and synthesis [28, 46]. An overview of this process is illustrated in Figure 4.2, where the decoupled features are first extracted from the target, normalized in the source, and then the extracted target features are matched in the source. The concentric circles represent the nested nature of these steps. As in Eq. 3.1 from the classical approaches, lower order moments must first be normalized before addressing higher order ones and the matching process must occur in reverse order.

Since the vision science background outlined in Chapter 2 specifies that the relevant features are summary statistics of neural adaptation, special care must be taken to

approximate the coding conditions at this stage of the visual pathway. The original work of Reinhard et al. [19] largely devoted its focus to the definition of an encoding space where the luminance and color channels were well decorrelated to allow for transfers with minimal cross talk between color regions. Here however, additional steps are taken to reflect the phenomena of chromatic adaptation and retinal response non-linearity, in addition to decorrelated opponent channels.

With these things in mind, the proposed method is computed in the following way:

1. Start with two images as input, one source S and target T , encoded for viewing on a standard sRGB display. The first step is to undo the $1/2.2$ gamma correction of this encoding specification to yield display linear radiance values.
2. Next, we apply the gray world method [47] to simulate chromatic adaptation by taking the channel-wise mean value to estimate the illuminants L_S and L_T for each radiance image, and then we scale the radiance image corresponding to S by the scale factor L_T / L_S . In this way, we simulate adaptation to the dominant illuminant of image T in the context of image S returning S_a .
3. Following this, the illuminant adapted S_a and original T radiance images are converted to the IPT color space. This space is employed in favor of the custom space of Reinhard et al. [19], as it has a strong decorrelation not only between the channels, but also between the perceptual attributes of hue and saturation. The preservation of the hue component is important as achromatic adjustments should not result in hue shifts, which can be very noticeable in content like skin tones. The space also reflects retinal response non-linearity with a 0.43 gamma function. Its computation is outlined in Figure 4.3.
4. Once S_a and T are converted to IPT space, the intensity channel I is isolated and the first four decoupled moments of T are extracted, normalized in S_a , and the moments of T are matched in S_a .
5. This process is repeated only for the first two moments in the chroma channels P and T , returning the transferred image S'_a .
6. After the feature transfer, S'_a is converted back to sRGB and the display non-linearity is re-applied.



Figure 4.3: Process for converting radiance (display linear) images with sRGB primaries to IPT space, where we use the symbol “ $\odot$ ” for representing an element-wise scalar operation for the vector coefficients (here a power).

4.2 Transferring Decoupled Moments

Following the process illustrated in Figure 4.2, measures of the mean, variance, skewness and orthokurtosis (μ_k , generally computed as in Eq. 2.1) are first extracted from the target and normalized in the source (x_k) in the following way (source and target vectors generalized as $\vec{x}$).

$$\begin{aligned}
 x_0 &= \vec{x} \\
 \mu_1(\vec{x}) &= \mu(x_0) \\
 x_1 &= x_0 - \mu_1 \\
 \mu_2(\vec{x}) &= \mu_2(x_1) \\
 x_2 &= \frac{x_1}{\sqrt{\mu_2(x_1)^2}} \\
 \mu_3(\vec{x}) &= \mu_3(x_2) \\
 x_3 &= x_2 \left(\frac{x_1}{1 - t_0 x_1} \right) \\
 \mu_4(\vec{x}) &= \mu_4(x_3)
 \end{aligned}$$

Where t_0 is a normalizing function which results in $\mu_3(\vec{x}) = 0$. Due to the normalization operations that are applied at each step, the resulting measures μ_4 is no longer equivalent to its standardized counterpart derived from Eq. 2.1. Through this normalization process, reference values for each progressive moment k can be obtained without releasing the previous moments, as explained in Chapter 2.

Then, the normalized input and reference values, are passed to MATLAB’s ordinary differential equation solver *ODE45*, given the de-normalization algorithm of Martinez-Enriquez et al. [24], which integrates along the fourth-order gradient (where the previous three moments are invariant) until the target ortho-kurtosis value is obtained, aiming for a solution with minimal impact on the input vector. This process is then

repeated for the third moment, and the second and first moments are adjusted to their reference values as in the method of Reinhard et al. (Eq. 3.1) [19]. For greater detail, readers can access the code directly¹, where the process for computing the nested normalization and transfer is found in the *toggle* function.

Finally, a relevant aspect to note is that the feature transfer corresponds to applying a non-linear point-wise function to the luminance of each pixel, a look-up table (LUT) can be generated to freeze the point-wise transformation and perform a very fast style transfer on other input samples (e.g., frames in a video sequence). The application of this LUT to arbitrary inputs is demonstrated in section 5.3.4.

4.3 Method usability

The proposed method is applied to images in its configuration outlined above, where the mean and variance values are transferred from target to source for IPT channels, and skew and ortho-kurtosis values only on the I channel. This particular configuration was determined to be an optimal balance between avoiding artifacts and achieving a strong visual similarity to the target. In Figure 4.4, the performance of this automatic configuration on a series of images is demonstrated. It is clear to see from these examples that the results take on a pleasing and visually similar appearance to their targets while preserving important details of the original sources.

Aside from the default state, users have several avenues to optimize the transform. One simple mechanism is the inclusion of more or fewer features on various input channels. In this way, the transfer can be made to reflect the target more accurately, or better maintain the characteristics of the source. Results can also be refined through manual cropping of the source and target images. Oftentimes when selecting candidates for style transfer users focus only on the appearance of the relevant subject of the image while ignoring the background’s influence. Cropping allows users to specifically select what image areas they would like to be included. In Figure 4.5, various cases where limiting the analysis to a region of the image can be beneficial to the results are demonstrated. One can see in these examples that cropping out large bright or dark regions has an effect on the result such that no details are clipped or crushed.

¹<https://github.com/trevorcanham/dCoupST>

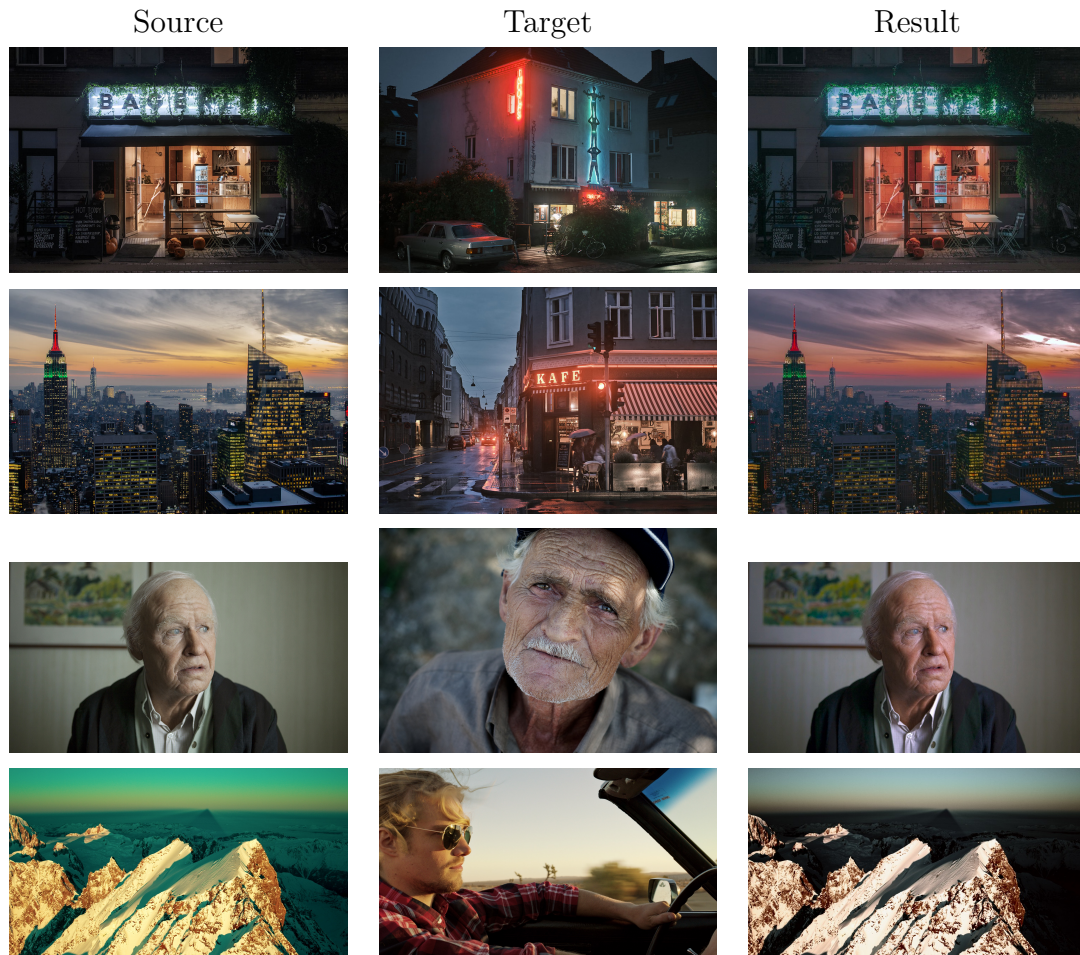


Figure 4.4: From left to right, varied source images, target images, and the results of the default configuration of our method. First row source and target images, and the second row target image uploaded to Flickr by Kristoffer Trolle (CC-BY 2.0). Second row source images uploaded to Flickr by Mk Feeny (CC-BY 2.0). Third row images sourced from [2]. Bottom row images provided by ARRI [1].

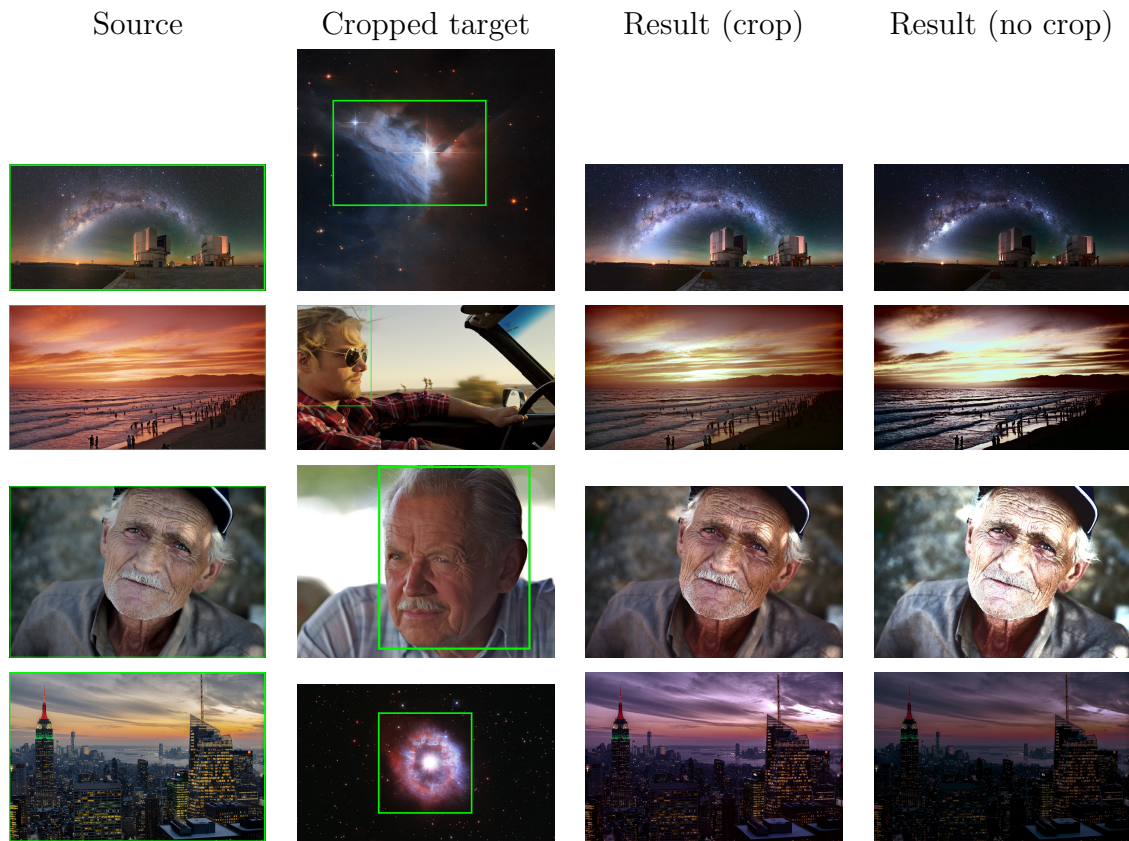


Figure 4.5: From left to right, source image with cropped region bounded in green, target image with cropped region bounded in green, default method transfer result with cropping, default method transfer result without cropping. Images sourcing left to right, top to bottom: European Southern Observatory (ESO), European Space Agency (ESA), ARRI Camera Group (complete second row), Labelled Faces in the Wild (complete third row) [2], MK Feeny, ESA. Licensed via CC-BY 2.0.

Chapter 5

Experiments

The evaluation of ST is as challenging as the definition of the problem itself and for the same reasons. The problem relies on a separation between an image’s style and its scene content, such that one image’s style can be applied over the scene content of another. The evaluation of a method’s effectiveness first requires the definition of an application relevant or arbitrary threshold which separates the two. For each ST method, this threshold is defined by the features which are extracted and matched, and each user carries their own expectation of what should be transferred from the target and what source details should stay the same. The user and method thresholds rarely align perfectly, making the judgment of ST quality particularly subjective compared to other image processing problems. For this reason, the most effective way to evaluate these methods is by characterizing their functionality through qualitative evaluations in challenging scenarios. In this section, comparisons between the proposed method and all of those discussed in the related work section are made and a psychophysical experiment comparing the most relevant approaches is presented. To augment this evaluation, larger test sets corresponding to each example shown here are available online¹. Then, direct experiments with motion picture and broadcast applications are demonstrated. Finally the method’s limitations and general recommendations for the development of photo-realistic style transfer tools is discussed.

¹https://drive.google.com/drive/folders/15nKB1jq0CCA9pm6i1fcISFSUvuYoBhac?usp=share_link



Figure 5.1: Experimental image test set, featuring highly chromatic and stylized scenes. Fourth image made available by TVR Romania, the remainder are made available by Pexels.

5.1 Qualitative Comparison

The results of the proposed method are compared here to the state-of-the-art. With the exception of Wu & Xue [12] and Xie et al. [13], which were implemented from scratch, the demo software provided by the authors of the respective works was employed in all other cases. The results of the proposed method are all computed in its default state outlined in Chapter 4. The original source images are shown in Figure 5.1.

The relevant NST approaches (Luan et al. [4], Li et al. [5], Mechrez et al. [6], Huang et al. [7], and Ke et al. [3]) were tested. Emulating how the methods would be used by a professional in image editing or color correction, neither pre-processing nor additional training steps were taken to refine the results, and so the methods are used “out of the box” as they were provided in their demo repositories. The results of Mechrez et al. are computed using the results of Luan et al. as input, as this was the primary method their work was tailored to refine. While the constrained transform of Ke et al. [3] shows clean results which are mostly free from artifacts, it fails to account for non-linear encoding operations common to color correction pipelines (Figure 5.2). Figure 5.3 shows the results of all methods in comparison to the proposed method, where one can see that the less constrained methods show egregious spatial distortions and spurious colors which make them unacceptable for professional photorealistic applications, even after automatic refinement steps are taken.

In Figure 5.4, the state-of-the-art OT approaches (Rabin et al. [8], Ferradans et al. [9], and Blondel et al. [10]) are compared to the proposed method. When applied to arbitrary sources, these methods are brittle and frequently contain egregious artifacts which range from banding to entire image sections failing to render. In some cases they have high computational complexity, rendering them impractical for use in an image editing and color grading context.

The results of Wu & Xue [12], Xie et al. [13] and Li et al. [11], shown in Figure

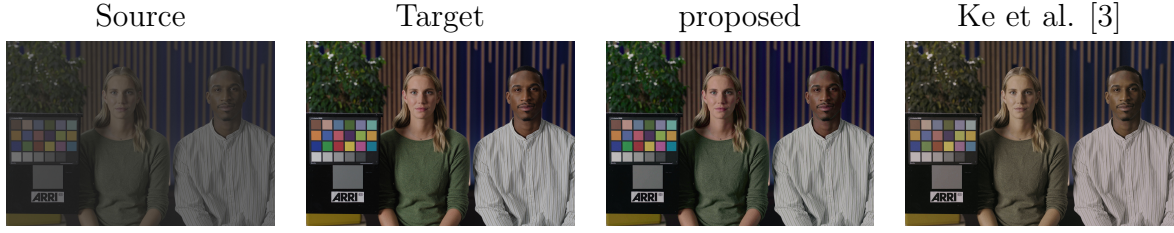


Figure 5.2: From left to right, source image with ARRI WCG $\text{Log}C_4$ encoding, target in Rec. 709 encoding, the result of Ke et al. [3], demonstrating that the method’s constrained transform cannot replicate basic non-linear transforms common to motion picture imaging pipelines. The results of the proposed method are shown for comparison.

5.5, demonstrate that these methods are also not robust outside of their test set and result in banding artifacts and poor color transfer, with the exception of the results of Wu & Xue. It should be noted however, that using the saliency threshold parameter recommended by the authors, all test images fall completely into the salient category and thus the method is simply performing the same transfer methodology as Reinhard et al. [19].

Next, the methods of Grogan et al. [14], Photoshop [15], and Zabaleta & Bertalmío [16], identified to be the most promising state-of-the-art methods for the photo-realistic application, are compared qualitatively in Figure 5.6. One can see that while these methods result in artifacts in some cases, they are for the most part correctable and limited to clipping or crushing, which can be recovered in pipelines where out of range values are allowed. The results of column 4 should be highlighted specifically, where the proposed method is able to match the color of the clipped highlights in the overhead lighting and other white regions of the source and target image, while all other methods render them with the tint of the target illumination. This example provides an effective demonstration of the proposed method’s ability to independently control distinct tone scale regions by separating clipped white areas from highlights where the illumination color is preserved. In the section that follows a psychophysical validation experiment to compare the quality of their results is described.



Figure 5.3: From top to bottom, sources, targets, proposed result, Ke et al. [3], Luan et al. result [4], Li et al. result [5], Mechrez et al. result [6] (taking Luan et al. results as input), Huang result [7].

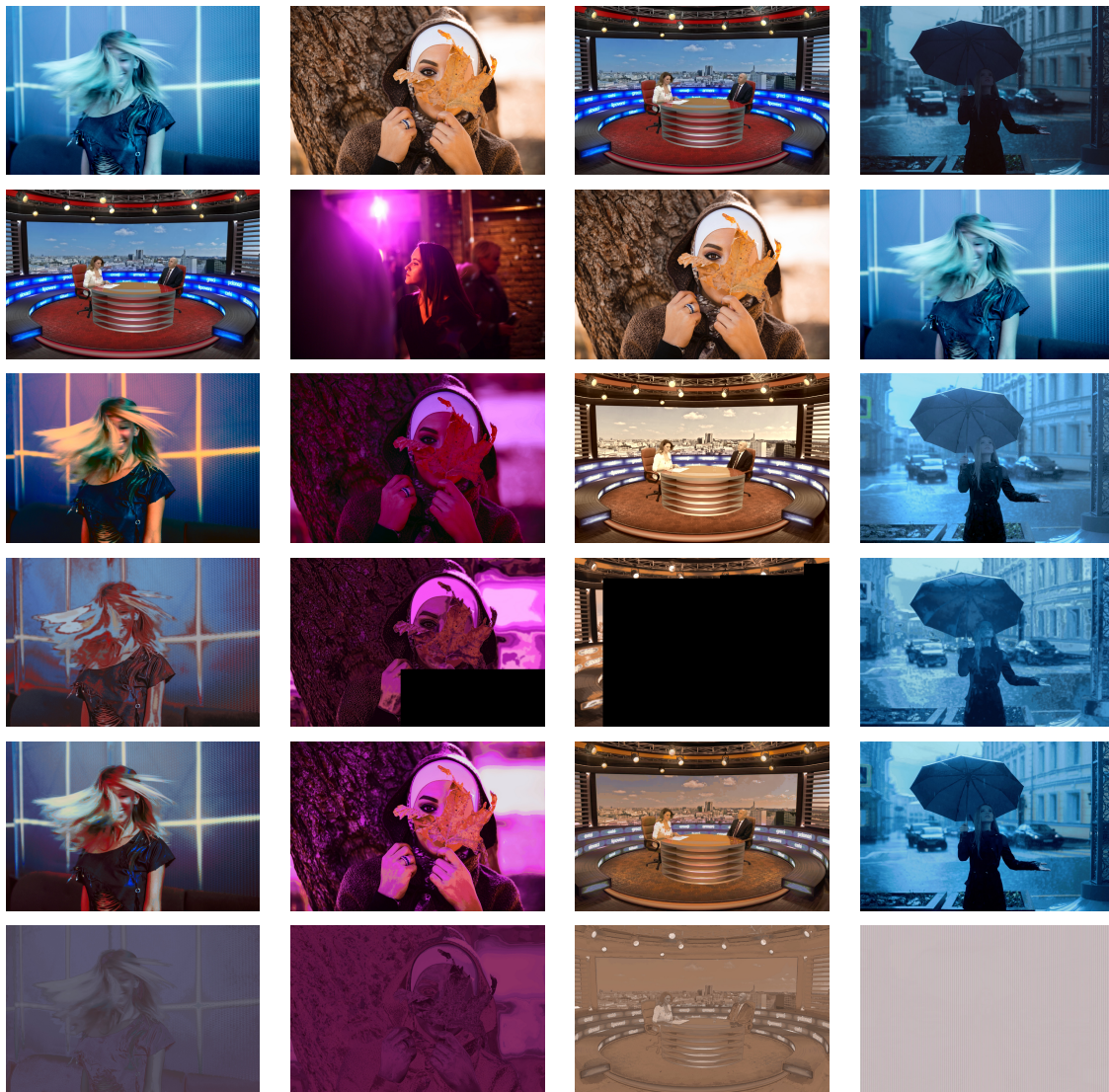


Figure 5.4: From top to bottom, sources, targets, proposed result, Rabin result [8], Ferradans result [9], Blondel result [10].

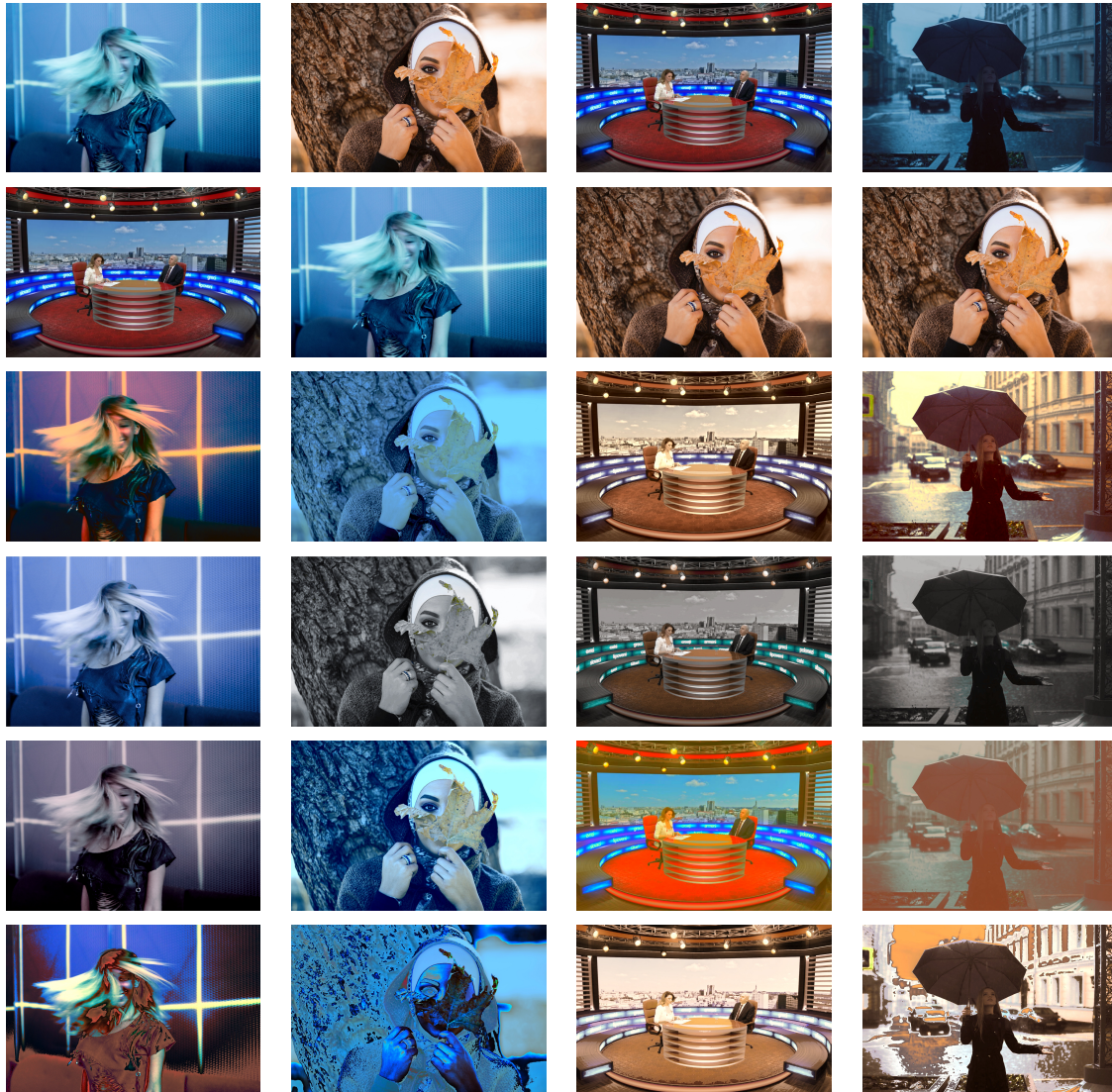


Figure 5.5: From top to bottom, sources, targets, proposed result, Li result [11], Wu result [12], Xie result [13].



Figure 5.6: From top to bottom, sources, targets, proposed result, Grogan result [14], Photoshop result [15], and Zabaleta & Bertalmío result [16].

5.2 Psychophysical experiment

A total of 19 observers, including experts and amateurs from a range of occupations with varied affiliations, nationalities, gender and age groups made subjective comparisons between the four most application-relevant methods (whose results are shown in Figure 5.6) in an online pair comparison experiment. Each method was used to make a style match between all combinations of the images in Figure 5.1. Observers were shown the original source and target images and made a two-alternate-forced-choice (2AFC) judgment between the results of two competing methods based on their ability to translate the style of the target to the source while maintaining a natural appearance, considering the original state of the source image.

The experimental image set (Figure 5.1) represents a series of highly challenging cases for style transfer. First, all of the images are highly chromatic and stylized, and include scenes with artificial lighting like dance clubs and broadcast studios. Second, all images are compressed to 8-bit per color channel detail. These features make them susceptible to gamut clipping and banding and highlight the primary challenges of working with graded images “in the wild.”

Also reflecting the qualities of an “in the wild” evaluation scenario, this experiment was conducted online using PsychoJS [48] and was distributed using the Pavlovia platform. While users were asked to run the experiment in a darkened viewing environment on a full-sized computer monitor, no other constraints were placed on the quality or calibration state of the viewing equipment. Observers were first presented with a series of instructions describing the task and then with a series of slides where the source and target images were shown at the top of the display and the results of two competing methods were displayed below. Observers were asked simply to select which of the two images they preferred. They were presented with every unique style/target combination of a set of 5 images (20 unique pairs), which were compared between four methods (proposed, [15], [14], [16]), resulting in a total of 120 evaluations. This task took observers between 10 and 20 minutes to complete which was comfortably within the acceptable range considering experimental fatigue. A full summary of the results is shown in Table 5.1.

The results of the experiment showed that, in comparison to the proposed method, the results of Photoshop [15] were preferred 32% of the time, followed by those of Zabaleta and Bertalmío [16] which were preferred 41% of the time, and the results

of Grogan et al. [14] were preferred 58% of the time. A full chart of the results is shown in Table 5.1. Looking back to Figure 5.6, one can infer an explanation for the success of Grogan et al. in the context of this experiment as the dominant color of the target image is closely reflected in all of their results. However, this is accomplished at the expense of maintaining the color discrimination boundaries between separate regions (e.g. foreground and background), resulting in the crushing of important memory colors (e.g. skin tones) to nearly monochromatic representations. In the context of professional color grading and image editing, users would consider this effect unfavorable because it prohibits further color-segmented refinement operations. In addition, the ranking reversal of Grogan et al. [14] and Zabaleta & Bertalmío [16] in this experiment versus the psychophysical validation of the latter brings to light the highly subjective nature of style transfer evaluation, where factors like the chosen image set and personal preferences of observers can heavily influence the results. For example, since many of the source images were already nearly washed out in highly chromatic light to begin with, these represented easy cases for the method of Grogan et al. to succeed and its shortcomings to be hidden.

Table 5.1: Psychophysical experiment results, where the percentage which the method on the column label was preferred over the method on the row label, averaged over 20 image combinations and 19 observers.

Methods	Photoshop [15] ↑	Grogan et al. [14] ↑	Proposed ↑	Zabaleta & Bertalmío [16] ↑
Photoshop [15] ↓	-	68%	62%	55%
Grogan et al. [14] ↓	32%	-	43%	38%
Proposed ↓	38%	58%	-	41%
Zabaleta & Bertalmío ↓ [16]	44%	63%	59%	-

5.3 Applications

5.3.1 Broadcast Source Homogenization

During the course of this work the proposed method was adapted for a novel broadcast application where the image of remote participants captured on mobile phones and webcams could be composited onto a studio broadcast. In this application it is important to mimic the color properties of the studio lighting and capture equipment in order to best facilitate the illusion that the participant is present in the studio.

The method was employed to make a color match between the user-submitted video and the broadcast stream, and then foreground extraction and super-resolution were applied to the user image to composite them onto an LED wall in the studio. This is a challenging color transfer problem, especially considering that studio broadcasters often have a virtual background or an LED wall that does not significantly impact the way the anchors are illuminated (as it is overpowered by the studio lights and is situated behind the anchors), but has a large influence on the luminance and color distribution of the target image. In order to account for this, a foreground extraction module was employed to matte out the background of the user image and replace it with the studio background such that the statistics of the background are more or less the same for the user and the anchor and the color transfer can focus on the lighting characteristics which are relevant to their appearance.

For this application, the proposed method is applied in its default setting. As a demonstration of the effectiveness of our method for this application, Figure 5.7 shows the results of replacing only the color transfer step of this process with the solutions of Grogan et al. [14], Photoshop [15], and Zabaleta & Bertalmío [16]. It can be seen that while Photoshop and Zabaleta & Bertalmío perform acceptably, the proposed method maintains the greatest degree of integrity for the original user image. This is an important quality for this application as users may be alienated if their skin tone and the colors of their clothing are misrepresented. Additionally, one can see that the method of Grogan et al. is inappropriate for this application, as the user skin tones are rendered to perfectly match studio background colors.

5.3.2 Video

It is not trivial to extend an image-based ST method to video sequences, as a simple frame-by-frame application of the method would most probably produce noticeable artifacts, with visible flickering and temporal incoherence even if the changes in content and color between frames are not significant. Another aspect to consider is that it is computationally expensive to calculate the ST transform for each frame independently when it should not change along the same sequence. In order to solve these issues, the approach used by Zabaleta & Bertalmío [16] is adopted for the video extension of the proposed method: the transform for a single representative frame of the source video is computed, encoded as a 3D LUT, and then applied to all the frames in the sequence. This guarantees temporal coherence and produces results that are free of

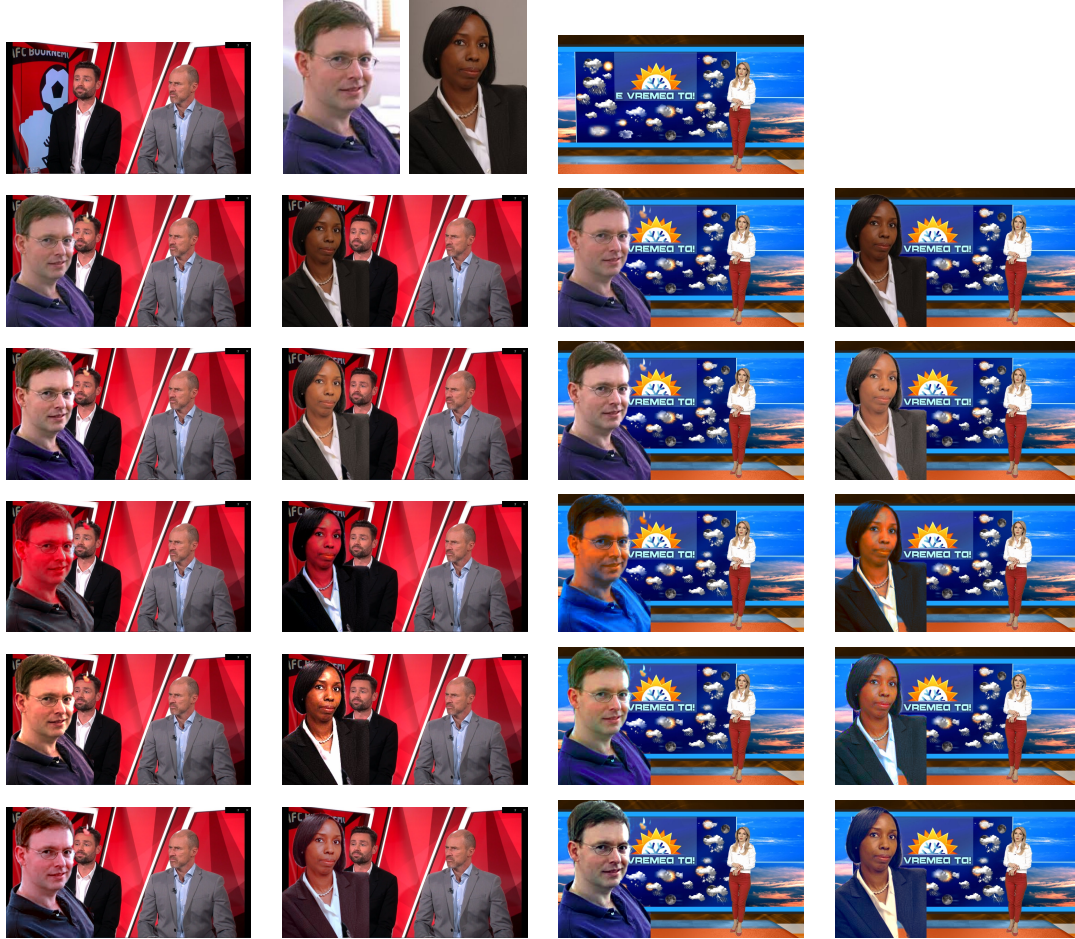


Figure 5.7: Broadcast application example. From top to bottom, user source images, studio targets, control (no style transfer applied), proposed result, Grogan et al. result [14], Photoshop result [15], and Zabaleta & Bertalmío result [16]. The user images are sourced from [2], the red studio image was provided by Premiere sports, and the blue studio image was provided by TVR Romania.

flickering artifacts.

The process of applying 3D LUTs to images is implemented in all professional image/video editing suites and is of very low computational complexity, making this process amenable for real-time implementation in the current production ecosystem. Furthermore, this development allows for the transform to be computed with a lower resolution version of the source and target images and then quickly applied to the original, full resolution source image. The computation time for the proposed method using a consumer level production workstation (2018 MacBook Pro with a 2.6 GHz intel core i7 processor and 16 GB 2400 MHz DDR4 memory, MATLAB 2021a) is 15 s for a 1920 x 1280 test image, 3.6 s at 50% resolution, and 0.8 s at 25%. While sub-sampling can cause the moments to shift subtly, informal tests show that the resulting transformations derived at 25% resolution have barely perceptible differences in comparison to the full resolution transform. So by taking advantage of LUT approximation and re-scaling, HD video sequences can be processed in real time after a ~ 1 s initial processing delay.

In order to generate the LUT, a $33 \times 33 \times 33$ 3D table corresponding to an identity transform is passed through the implementation described above, and any operations which alter the image itself are identically applied to the LUT in sequence (aside from the ortho-kurtosis transform which is accurately approximated with a spline function.) This LUT is then saved in the Adobe *.cube* format [49] such that it can be imported to any video editing or color correction platform and integrated into standard workflows. A series of demo videos is available online¹ which show that the LUT approach allows for a video style transfer to be accomplished which is free from temporal artifacts.

5.3.3 Demo software

Readers can access a standalone GUI implementation of the proposed method (shown in Figure 5.8) and the MATLAB source code². This implementation allows users to select the source and target images, toggle features to be transferred to specific color channels, perform cropping on the source and target input, encode the transform as a LUT, and batch process images for a particular transfer.

¹https://drive.google.com/drive/folders/15nKB1jq0CCA9pm6i1fcISFSUvuYoBhac?usp=share_link

²<https://github.com/trevorcanham/dCoupST>

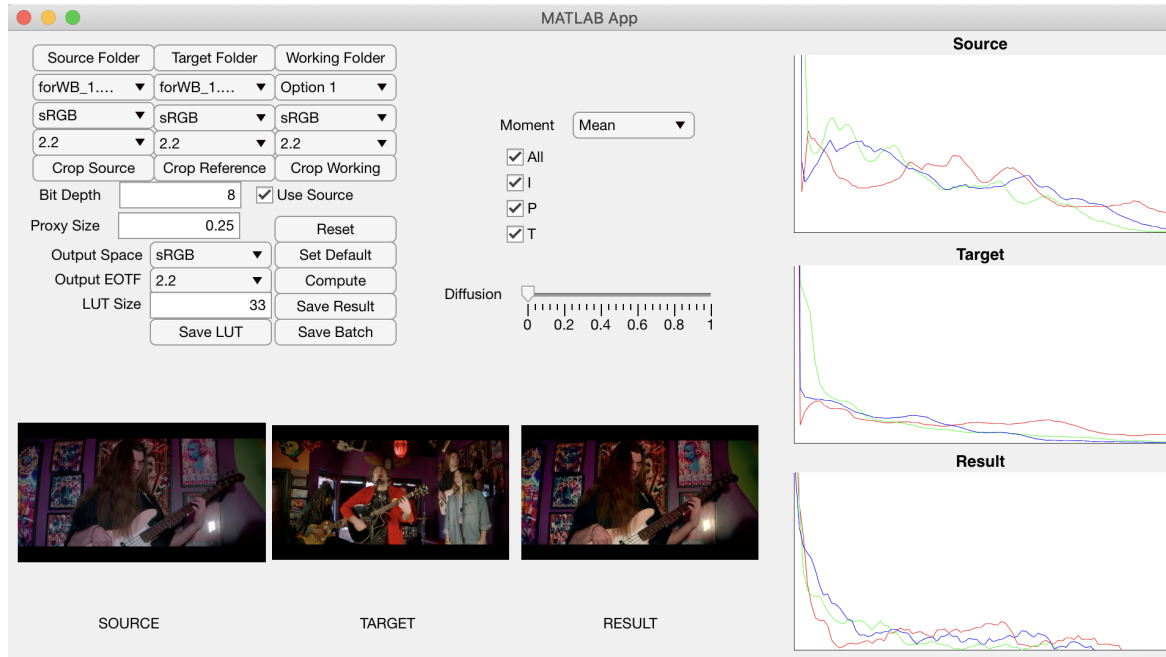


Figure 5.8: Demo software GUI.

5.3.4 Color grading workflow case study

To validate the usefulness of the demo software in a real production scenario the method was employed during the color grading process of a music video. The source footage was shot on a Sony $\alpha 7S II$ and was delivered in $Slog_3$ wide gamut encoding. The process started by selecting a series of source stills that were particularly meaningful to the story. Then, a series of potential looks were developed by matching these stills to a number of targets which had previously made a pleasurable impression on the colorist. During this process, the source footage was kept in its original encoding format to avoid clipping or quantization. However since the target images were in a display-ready format the results reflected these encoding characteristics.

After finding a good source/target match, a LUT was derived which served as the base look for the video. The effect of this LUT on a range of scenes is shown in Figure 5.9. Due to mixed lighting conditions in several scenes additional corrections had to be made to avoid jarring changes in color balance between cuts. To accomplish this, the pair of frames constituting this transition were matched, producing a LUT to be applied to the source clip. The results are shown in Figure 5.10. For the scene in the top row, similar scene contents were cropped out for matching and only the first order moment was matched to account for the shift in lighting conditions. In the bottom

row the method was used in its default state with no cropping. While the resulting image does not appear to have the exact same lighting conditions as the target image, the transition between the shots is remarkably smoother since they now share the same moments up to the fourth order. In this sense, the concept of matching these features and their relation to visual adaptation is validated. Apart from these edits, only minimal manual adjustments were made during the color correction process and so the time and technical skill required to complete this part of the production was greatly alleviated.

5.3.5 Recommendations & limitations

In order to supplement the discussion on usability, a number of useful guidelines are offered here for the best use of the proposed method and to highlight its limitations. Furthermore, we offer a series of general recommendations on the design of photo-realistic style transfer methods.

One artifact that can occur is the exacerbation of banding or other compression artifacts. For this reason, the use of high bit depth, uncompressed images is recommended for best results. Another is the clipping or crushing of detail in order to fit the target statistics. To avoid this, it is better to work in a non-destructive space, where out-of-range values are preserved and clipping does not occur so that these details can be reintegrated back into the image at the user's discretion. In addition, the cropping function can be used as in Figure 4.5 to better align the source and target images to derive a compatible transfer.

Another limitation which was highlighted in the comparison to the method of Grogan [14] is that the dominant colors' appearance in the target image is not always well preserved when integrating with the spatial details of the source image. The best results in this regard came from the use of the IPT color space. However, one might encounter a preferable alternative representation. In addition to modulating the color space, some applications may benefit from modulating the number of features transferred per color channel. For instance in the default configuration, the higher order moments are only transferred in the case of the luminance channel, but it may be beneficial in some cases to also transfer these features on the chroma channels.

As discussed previously, there exists an image compatibility factor in ST, where only a subset of possible source/target combinations will be acceptable. The range of this subset for a given source image depends on the degree to which it approaches

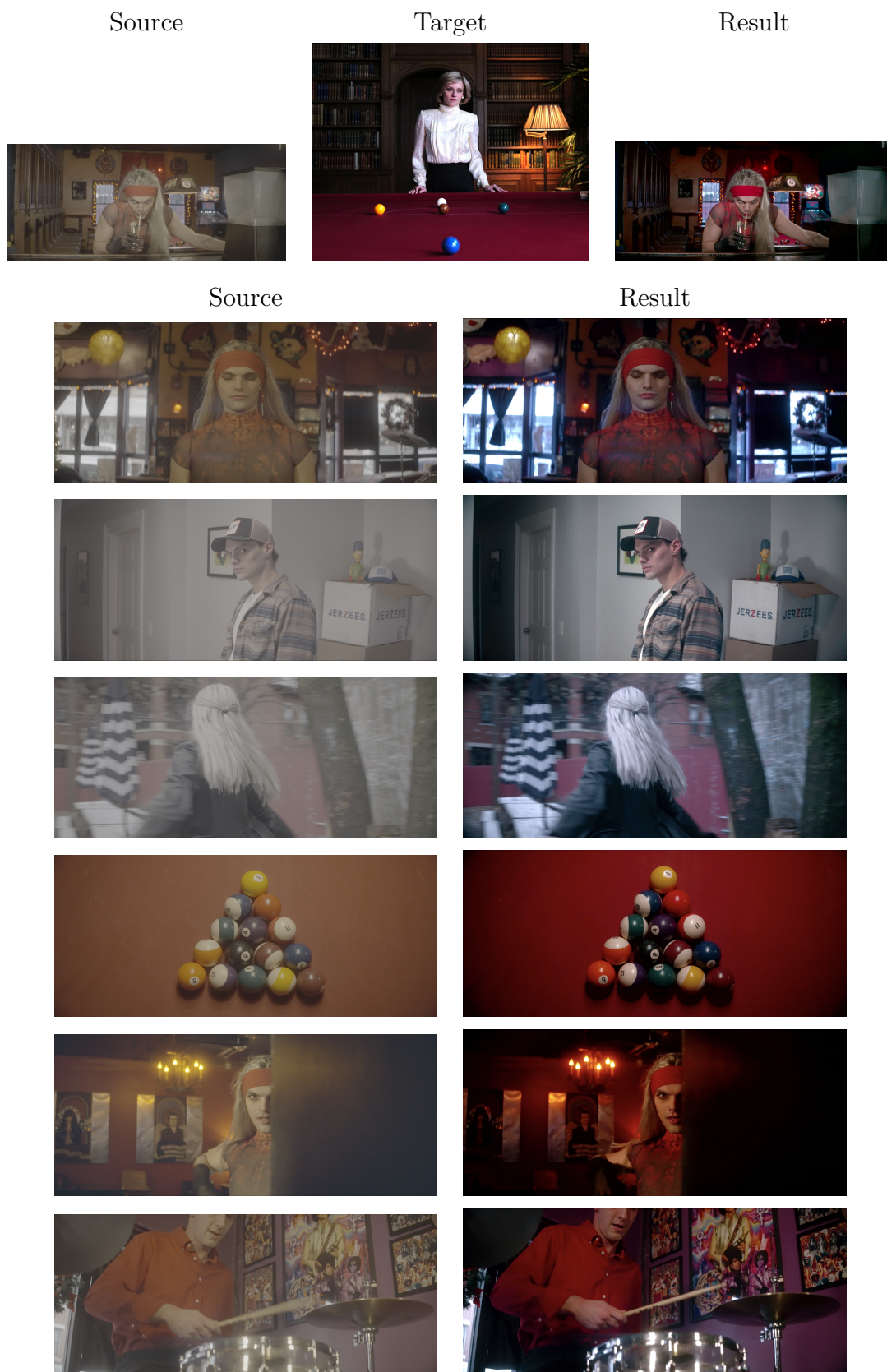


Figure 5.9: The “hero” look, which defines the color processing for the whole project is derived in the top row from source and target stills. The resulting LUT can then be applied to a variety of scenes.



Figure 5.10: Additional corrections to avoid jarring cuts between shots with different lighting conditions. In the bottom row, the analysis of the source and target is limited to the green boxes through cropping.

its gamut boundaries, as well as its distance in feature space from a potential target. For example, common transfer mismatches include target images that are significantly darker or brighter than the source, causing details to be obscured via clipping and crushing.

A trade-off between feature intensity and artifacts can be observed in all of the tested methods, but is most poignantly demonstrated between the constrained but artifact-free results of Ke et al [3]. with the intense feature adjustment of Luan et al. [4]. In general, transfers which employ more visually striking features are more likely to produce artifacts. With this in mind, a recommendation can be made to parameterize methods such that users can make simple feature-intensity adjustments and determine the ideal result for their purposes (i.e. by adjusting a weighting between source and target features, or by including more or less features as in the proposed). Methods with meaningful modular controls have a greater potential for interactivity, through which users can develop fluency. Fully automatic tools, on the other hand, are more likely to be seen as gimmicks, and can be frustrating to artists as they provide no corrective recourse in the results.

With that said, there exist high intensity features with minimal content dependency (like the global ones employed in the proposed method), which can be transferred without lingering artifacts related to the source content. On a related note, features which are intense enough to be readily identifiable in target and source images are helpful for method interpretability. Features should also be consistent enough that users can

predict their effect on the source and intentionally select targets. The strong vision science motivation and interpretability of the proposed method is beneficial to this end. Finally, features whose appearance resemble that of pre-existing workflow elements like camera rendering and color editing are easier to conceptualize and integrate into the artistic process.

This work also serves as a demonstration that advancements can still be made on classical image processing methods without the necessity of training. As mentioned in Chapter 3, many of the methods of the neural style transfer category require significant computational resources to produce a working model. At the same time, the authors of these works offer the training methodology (dataset, architecture and loss function) as a contribution and state that, although the presented results may be imperfect for all applications, users can retrain the model for their particular purposes without being fully transparent about this computational cost. The performance benefits of the results should be considered in context of the resource expenditure of these solutions. In contrast, the proposed method transfers features which are relevant to pre-existing image editing workflows in a novel and computationally simple way without requiring training.

Chapter 6

Summary and Future Work

6.1 Summary

A photo-realistic style transfer method for image and video based on vision science principles and on a recent mathematical formulation for the deterministic decoupling of sample statistics is proposed. This method extends traditional approaches by increasing the number of features transferred, including higher order statistical moments which correspond to important cues for visual adaptation processes and serve as impactful and interpretable feature descriptors. The method is validated in qualitative comparisons against the state-of-the-art via a psychophysical experiment, and is tested on a variety of applications. The practical findings from these experiments are summarized in a series of recommendations for style transfer method design in Chapter 5. Finally, the method is implemented in a demo software GUI to facilitate its adoption by post-production workers.

6.2 Future Research Directions

As future work, a tool designed to give users the ability to adjust the statistical moments addressed here independently, without the need of a target image would be of interest. Due to their decoupled and localized nature (with respect to the tone scale), it would be useful for colorists to control them directly, particularly in the context of emerging formats where the tone scale range is extended. The challenge in doing so, however, is in computing the feature adjustment such that users can receive real-time feedback from the controls. Another interesting and challenging extension

of this work would be to characterize and reproduce additional spatial and *texture* features, such as different types of noise and film grain.

The proposed method can also be applied to interact with visual perception in various artistic and scientific applications. For example, the method could be used to gradually transition over a sequence of images which trigger varying states of adaptation in the viewer, as frequent and abrupt re-adaptation may be a cause of display fatigue. Also, by using a randomly distributed noise image as source, the method could be used to apply the moments of a target image, which could then be used to pre-adapt observers before seeing the image itself (for example, in a psychophysical experiment) or to be used as a background for image presentation. The psychophysical experiments of Julesz [29] and Chubb et al. [31] were based on stimuli which varied with respect to their moments. While the stimulus generation techniques employed in these works were limited to synthetic random patterns with coupled moments that were limited to a finite range, the proposed method allows for the moments of both synthetic and natural image stimuli to be manipulated in a decoupled manner to arbitrary reference values, extending these experimental paradigms towards more practically relevant scenarios involving complex images.

Finally, building on the vision science motivation for the normalization and extraction of sample statistics, an alternative direction would be to explore further applications of intermediate representations where particular features of interest are normalized. Due to the regularized statistical properties and the decoupling of features like contrast and brightness, these representations may allow for more straightforward interpolation, compression and re-appropriation for interchange between formats (tone and gamut mapping).

References

- [1] ARRI Camera Systems, “Enhanced capture material for hdr4eu project d2.2,” 2018.
- [2] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.
- [3] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson W.H. Lau, “Neural preset for color style transfer,” in *CVPR*, 2023.
- [4] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala, “Deep photo style transfer,” in *CVPR*, 2017.
- [5] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz, “A closed-form solution to photorealistic image stylization,” in *ECCV*, 2018.
- [6] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor, “Photorealistic style transfer with screened poisson equation,” in *BMVC*, 2017.
- [7] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, “Multimodal unsupervised image-to-image translation,” in *CVPR*, 2018.
- [8] Julien Rabin, Sira Ferradans, and Nicolas Papadakis, “Adaptive color transfer with relaxed optimal transport,” in *ICIP*, 2014.
- [9] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol, “Regularized discrete optimal transport,” *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1853–1882, 2014.
- [10] Mathieu Blondel, Vivien Seguy, and Antoine Rolet, “Smooth and sparse optimal transport,” *ArXiv*, 2017.

- [11] Zhijiang Li, Zhenshan Tan, Liqin Cao, Hu Chen, Lei Jiao, and Yanfei Zhong, “Directive local color transfer based on dynamic look-up table,” *Signal Processing: Image Communication*, vol. 79, pp. 1–12, 2019.
- [12] Zongsheng Wu and Ru Xue, “Color transfer with salient features mapping via attention maps between images,” *IEEE Access*, vol. 8, pp. 104884–104892, 2020.
- [13] Bin Xie, Kang Zhang, and Ke Liu, “Color transfer based on color classification,” in *ICEMCE*, 2020.
- [14] Mairéad Grogan and Rozenn Dahyot, “Robust registration of gaussian mixtures for colour transfer,” *ArXiv*, 2017.
- [15] Adobe Systems Incorporated, *Match colors in photoshop*, Adobe Systems Incorporated, San Juan, California, United States, 2022.
- [16] Itziar Zabaleta and Marcelo Bertalmío, “Photorealistic style transfer for video,” *Signal Processing: Image Communication*, vol. 95, pp. 116240, 2021.
- [17] Steve Yedlin, “On color science,” <http://www.yedlin.net/OnColorScience/>, Accessed: 2010-09-30.
- [18] Hiroaki Kotera, Teturo Morimoto, and Ryoichi Saito, “Object-oriented color matching by image clustering,” in *The Color Imaging Conference*, 1998.
- [19] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley, “Color transfer between images,” *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.
- [20] Ján Morovic, *To develop a universal gamut mapping algorithm*, Ph.D. thesis, University of Derby, 1998.
- [21] François Pitié, Anil Kokaram, and Rozenn Dahyot, “N-dimensional probability density function transfer and its application to colour transfer,” in *ICCV*, 2005.
- [22] Leon Gatys, Alexander Ecker, and Matthias Bethge, “Image style transfer using convolutional neural networks,” in *CVPR*, 2013.
- [23] Matthew J. Colbrook, Vegard Antun, and Anders C. Hansen, “The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and smale’s 18th problem,” *PNAS*, vol. 119, no. 12, 2022.

- [24] Eduardo Martinez-Enriquez, Maria del Mar Gonzalez, and Javier Portilla, “Deterministic decoupling of global features and its application to data analysis,” 2022.
- [25] Rajesh Sharma and Rajmohan Bhandari, “Skewness, kurtosis and Newton’s inequality,” *The Rocky Mountain Journal of Mathematics*, vol. 45, no. 5, pp. 1639–1643, 2015.
- [26] David C. Blest, “A new measure of kurtosis adjusted for skewness,” *Australian and New Zealand Journal of Statistics*, vol. 45, no. 2, pp. 175–179, 2003.
- [27] Juan Francisco Rosco, Arthur Pewsey, and Michael Chris Jones, “On Blest’s measure of kurtosis adjusted for skewness,” *Communications in Statistics: Theory and Methods*, vol. 44, no. 17, pp. 3628–3638, 2015.
- [28] Eduardo Martínez-Enríquez and Javier Portilla, “Controlled feature adjustment for image processing and synthesis,” in *MMSP*, 2020.
- [29] Bela Julesz, “Visual pattern discrimination,” *IRE transactions on Information Theory*, vol. 8, no. 2, pp. 84–92, 1962.
- [30] Anne Treisman, “Preattentive processing in vision,” *Computer Vision, Graphics, and Image Processing*, vol. 31, no. 2, pp. 156–177, 1985.
- [31] Charles Chubb, Michael S. Landy, and John Econopouly, “A visual mechanism tuned to black,” *Vision Research*, vol. 44, no. 27, pp. 3223–3232, 2004.
- [32] William K Pratt, Olivier D Faugeras, and Andre Gagalowicz, “Visual discrimination of stochastic texture fields,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 11, pp. 796–804, 1978.
- [33] Matthew Yedutenko, Marcus HC Howlett, and Maarten Kamermans, “Enhancing the dark side: asymmetric gain of cone photoreceptors underpins their discrimination of visual scenes based on skewness,” *The Journal of physiology*, vol. 600, no. 1, pp. 123–142, 2022.
- [34] Javier Portilla and Eero P Simoncelli, “A parametric texture model based on joint statistics of complex wavelet coefficients,” *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

- [35] Benjamin J Balas, “Texture synthesis and perception: Using computational models to study texture representations in the human visual system,” *Vision Research*, vol. 46, no. 3, pp. 299 – 309, 2006.
- [36] Fred Attneave, “Some informational aspects of visual perception,” *Psychological review*, vol. 61, no. 3, pp. 183–193, 1954.
- [37] Horace B Barlow, “Possible principles underlying the transformations of sensory messages,” pp. 217–234, 1961.
- [38] Eero P Simoncelli and Bruno A Olshausen, “Natural image statistics and neural representation,” *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.
- [39] David H. Hubel, *Eye, Brain and Vision*, Academic Press, 2019.
- [40] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song, “Neural style transfer: A review,” *IEEE Transactions on Visualization & Computer Graphics*, vol. 26, no. 11, pp. 3365–3385, 2020.
- [41] Jie An, Haoyi Xiong, Jun Huan, and Jiebo Luo, “Ultrafast photorealistic style transfer via neural architecture search,” in *AAAI*, 2020.
- [42] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang, “Learning linear transformations for fast image and video style transfer,” in *CVPR*, 2019.
- [43] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha, “Photorealistic style transfer via wavelet transforms,” in *CVPR*, 2019.
- [44] Jan Morovic and Pei-Li Sun, “Accurate 3d image colour histogram transformation,” *Pattern Recognition Letters*, vol. 24, no. 11, pp. 1725–1735, 2003.
- [45] François Pitié, “Advances in colour transfer,” *IET Computer Vision*, vol. 14, no. 6, pp. 304–322, 2020.
- [46] Eduardo Martinez-Enriquez and Javier Portilla, “Deterministic feature decoupling by surfing invariance manifolds,” in *ICASSP*, 2020.
- [47] Gershon Buchsbaum, “A spatial processor model for object colour perception,” *Journal of the Franklin institute*, vol. 310, no. 1, pp. 1–26, 1980.

-
- [48] Jonathan Pierce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv, “Psychopy2: Experiments in behavior made easy,” *Behavior Research Methods*, vol. 51, pp. 195–203, 2019.
- [49] Adobe Systems Incorporated, “Cube lut specification version 1.0,” September 2013.