Perceptually-Informed No-Reference Image Harmonisation

Alan Dolhasz[1][0000−0002−6520−8094], Carlo Harvey[1][0000−0002−4809−1592], and Ian Williams[1][0000−0002−0651−0963]

Digital Media Technology Lab, Birmingham City University, Birmingham, UK
{alan.dolhasz, carlo.harvey, ian.williams}@bcu.ac.uk
http://dmtlab.bcu.ac.uk

Abstract. Many image synthesis tasks, such as image compositing, rely on the process of image harmonisation. The goal of harmonisation is to create a plausible combination of component elements. The subjective quality of this combination is directly related to the existence of human-detectable appearance differences between these component parts, suggesting that consideration for human perceptual tolerances is an important aspect of designing automatic harmonisation algorithms. In this paper, we first investigate the impact of a perceptually-calibrated composite artifact detector on the performance of a state-of-the-art deep harmonisation model. We first evaluate a two-stage model, whereby the performance of both pre-trained models and their naive combination is assessed against a large data-set of 68128 automatically generated image composites. We find that without any task-specific adaptations, the two-stage model achieves comparable results to the baseline harmoniser fed with ground truth composite masks. Based on these findings, we design and train an end-to-end model, and evaluate its performance against a set of baseline models. Overall, our results indicate that explicit modeling and incorporation of image features conditioned on a human perceptual task improves the performance of no-reference harmonisation algorithms. We conclude by discussing the generalisability of our approach in the context of related work.

Keywords: image compositing · harmonisation · artifact detection · end-to-end compositing · deep learning

1 Introduction

Image harmonisation is an important task in image compositing and synthesis, aiming to minimise appearance-based differences between individual elements of a composite, in order to produce a perceptually plausible end result [32]. An image composite commonly consists of at least one *object*, inserted into a background image, referred to as the *scene*. As the object and scene are commonly captured under different environmental conditions, visible appearance mismatches between them may exist, due to differences in illumination, camera intrinsics, post-processing, encoding or compression. Thus, the goal of image

harmonisation is to minimise such differences and create a realistic result. This process can be performed manually by compositing artists, however, many automatic approaches have been proposed, including alpha matting - linear combinations of object and scene pixel values [23], gradient-domain optimization techniques [22, 1, 20], statistical appearance transfer [25, 19] and multi-scale methods [4, 5, 29].

With the advent of deep learning (DL), automatic image synthesis techniques have garnered renewed interest and afforded considerable improvements in state-of-the-art image compositing and harmonisation techniques. Methods using variants of convolutional autoencoders (AEs) have been successfully used to directly approximate the harmonisation function, in a supervised learning setting. Notably, Tsai et al. (2017) [30] use a convolutional AE in a multi-task setting to both segment and harmonise an input image, provided the target object mask. Another approach [7] uses a generative adversarial network (GAN) to perform both colour and geometric transformations, pre-training their model on synthetically-generated data. Conditional GANs have also been applied in this context, by learning to model joint distributions of different object classes and their relationships in image space. This allows for semantically similar regions to undergo similar transformations [2]. A more recent method combines state-of-the art attention mechanisms and GAN-based architectures with explicit object-scene knowledge implemented through masked and partial convolutions and provide a dedicated benchmark image harmonisation dataset, dubbed iHarmony [8].

A common requirement of these state-of-the-art techniques is the provision of binary object/scene segmentation masks at input, both during training and inference. These masks serve as an additional feature, identifying the corresponding image pixels that require harmonisation. As such, these methods are applicable to scenarios where new composites are generated, and these masks are available. However, in cases where these ground truth masks are not available, these techniques can not be easily applied without human intervention, limiting their application to scenarios such as harmonisation of legacy composites. Moreover, existing methods do not explicitly leverage human perception - the usual target audience of image composites. This includes human sensitivity to different local image disparities between object and scene, shown to correlate with subjective realism ratings [12]. Lastly, binary object masks used in these techniques provide only limited information about the nature of the required corrections, indicating only the area where corrections are needed. This can result in the harmonisation algorithm over- or under-compensating in different local regions of the composite.

In a recent pilot study [11], the authors argue that explicit modeling of the *perception* of compositing artifacts, in addition to their improvement, would allow for harmonisation algorithms to be used in a *no reference* setting, whereby the input mask is not required at inference time. Thus, the model performs both the *detection* and *harmonisation* task. They also show that combining two off-the-shelf, pre-trained models – a detector [10] and a harmoniser [30] – can

achieve comparable results to mask-based state-of-the-art harmonisation algorithms. This enables design of end-to-end harmonisation networks without the need for input object masks, allowing automatic harmonisation of content for which masks are not readily available. The authors also claim that the explicit encoding of the location and perceptual magnitude of errors in the model could allow the process to take advantage of the benefits of multi-task learning, feature sharing and attention mechanism in terms of generalisation [24, 26]. The potential applications of such automatic compositing systems are wide-ranging, including improvement of legacy content, detection of image manipulations and forgery, perceptually-based metrics and image synthesis.

In this paper, we recapitulate and extend this work to an end-to-end model designed, trained and evaluated from scratch. First, we present the original proof-of-concept two-stage compositing pipeline [11]. This consists of a *detector* network, which outputs masks corresponding to regions in an input image requiring harmonisation, and a *harmoniser* network, which corrects the detected regions. We then evaluate the performance of the harmoniser based on using object masks predicted by the detector, versus using ground truth object masks. Based on the evaluation of the two-stage model, we then propose a single end-to-end model, and compare its performance to a set of baselines trained from scratch on the challenging iHarmony dataset, as well as the synthetic COCO-Exp dataset from the original study [11]. We show that our end-to-end model outperforms the baselines on both datasets. This indicates the usefulness of the pre-trained perceptual features to the compositing task using two different end-to-end architectures. To our knowledge, this is the first work investigating an end-to-end combination of a DL-based feature extractor, conditioned on a perceptual task, with an image harmonisation network to perform no reference image harmonisation.

The remainder of the paper is structured as follows: Section 2 introduces related work and discusses state-of-the-art techniques, Section 3 describes the original methodology adopted for the two-stage model evaluation, Section 4 presents the results of this evaluation and Section 5 discusses the findings [11]. In Section 6 we detail the methodology, architecture and optimisation details of the proposed end-to-end models, which are evaluated in Section 7. Finally, in Section 8, we review our findings in the context of the original study and wider application to image harmonisation. We also discuss the strengths and weaknesses of our approach, before concluding and considering future research directions in Section 9.

2   Related Work

2.1   Image Compositing & Harmonisation

Automatic image compositing and harmonisation are both active and challenging problems in the domain of image understanding, synthesis and processing. While, image compositing concerns the entire process of combining regions from different source images into a plausible whole, image harmonisation focuses on the problem of matching the various appearance features between the object and

scene, such as noise, contrast, texture or blur, while assuming correctly aligned geometric and illumination properties [29].

Similarly to the problem of image in-painting, compositing and harmonisation are both ill-posed problems [16]. For a given region requiring correction, many different arrangements of pixels could be deemed plausible. This is in contrast to problems where the solution is unique. Depending on the content and context of an image composite, some scene properties, and thus required object corrections, may be inferred from the information contained within the image or its metadata, such as the characteristics of the illuminant [27], colour palette, contrast range or the camera response function. Other properties, such as an object's albedo, texture or shape are often unique to the object and cannot be derived directly from contextual information in the scene. While methods for approximation of these properties do exist [15], they are difficult to integrate into end-to-end systems and can be challenging to parametrise. The recent successes in DL have motivated a number of approaches [30, 2, 7, 8] which attempt to exploit the huge amount of natural imagery available in public datasets in order to learn the mapping between a corrupted composite image and a corrected composite, or natural image.

2.2   Multi-task Learning, Feature Sharing & Attention

Due to the abundance of natural image data and the ill-posed nature of the compositing problem, DL approaches are well-suited for this task. However, supervised DL methods require large amounts of annotated data in order to learn and generalise well. This requirement grows along with the complexity of a problem and the desired accuracy. In order to tackle this issue, many architectural considerations have been proposed, many of which focus on learning good feature representations, which generalise well between tasks.

Multi-task learning approaches rely on performing multiple related tasks in order to learn better feature representations. In recent years many tasks in image understanding have achieved state-of-the-art performance by incorporating multi-task learning [14], for example in predicting depth and normals from a single RGB image [13], detection of face landmarks [36] or simultaneous image quality and distortion estimation [17]. This is afforded by the implicit regularisation that training a single model for multiple related tasks imposes [6], and the resulting improved generalisation. Feature sharing approaches combine deep features from related domains or tasks in order to create richer feature representations for a given task. This is similar to the multi-task paradigm, however instead of sharing a common intermediate feature representation, features from one or multiple layers of two or more networks are explicitly combined. The Deep Image Harmonisation (DIH) model [30] adopts both these paradigms, by combining the tasks of image semgentation and harmonisation and sharing deep features of both task branches. Finally, attention mechanisms [9] can also be used to learn the relative importance of latent features for different combinations of task and input sample.

2.3   No more masks

State-of-the-art image harmonisation methods focus largely on improving composites in scenarios where the identity of pixels belonging to the object and scene are known a priori. For example, the DIH approach [30] uses a AE-based architecture to map corrupted composites to corrected ones, incorporating a two-task paradigm, which attempts to both correct the composite, as well as segmenting the scene. However, this approach does not explicitly condition the network to learn anything more about the corruption, such as its magnitude, type or location. Instead object location information is explicitly provided at input, using a binary mask. A similar approach [7] inputs the object mask at training time, while also introducing mask segmentation and refinement within a GAN architecture, in addition to learning of geometric transformations of the object. The segmentation network, as part of the adversarial training process, discriminates towards ground truth binary masks as an output - omitting any perceptual factor in the discrimination task. This achieves improved results compared to the AE, however at the cost of a more complex architecture and adversarial training. Due to the many dimensions along which combinations of object and scene may vary, compositing systems should be equipped to encode such differences before attempting to correct them. Kang et al. (2015) [17] show that a multi-task approach is an efficient way to ensure that distortions are appropriately encoded by the model. Other approaches to this problem include self-supervised pre-training to enforce equivariance of of the latent representation to certain input transformations [34], which has been used to train perceptually-aligned local transformation classifiers [10], also used in the proposed model.

3   Two-Stage Model: Methodology

3.1   Motivation

Whilst multi-task learning has been shown to be efficient in the coupled process of detecting and correcting arbitrary pixel level transformations within images, perceptually-based encoding of artifacts within masks has not yet been shown to be effective in the image harmonisation field. Before approaching the multi-task model, it is necessary to prove empirically that this end-to-end process is viable. Thus we first design a two-stage approach using two existing standalone networks for both detection and harmonisation to test the efficacy of these perceptual masks in this domain.

3.2   Approach

Our overarching goal is the design of an end-to-end automatic compositing pipeline, capable of detection and correction of common compositing artifacts, without the need for specification of an object mask. In order to evaluate the effectiveness of this approach, we assess predicted, perceptually-informed object

masks, rather than ground truth object masks, as input to the deep harmonisation algorithm. We then measure similarity between ground truth images and composites corrected with the harmonisation algorithm, using either the original synthetic binary masks M_s or the perceptually-based masks predicted by the detector M_p. Accordingly, we refer to composites harmonised using ground truth masks as C_s, and composites generated by the end-to-end system as C_p.
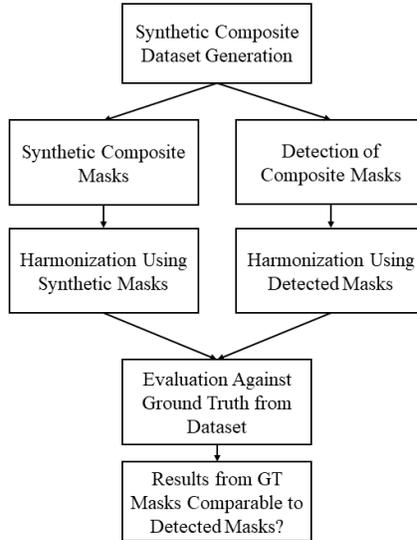

Fig. 1: Illustration of research methodology adopted in the two-stage model evaluation. Reprinted from [11].

We evaluate the hypothesis that the performance of an end-to-end detection and harmonisation model is comparable to a harmonisation model using manually created object masks. Confirmation of this hypothesis would support our case for incorporating explicit detection of composite artefacts into end-to-end image composite harmonisation systems. Our research methodology is summarised in Figure 1.

3.3   Detector and Harmoniser Models

Both the detector (referred to as the PTC henceforth) [10] and the harmoniser (referred to as the DIH) [30] are deep, image-to-image, fully convolutional autoencoder networks. The PTC takes a single image as input and generates a 2-channel output mask, which encodes probabilities for each pixel, p, in the input image as being affected by a negative (channel 0) or a positive (channel 1) perceptually suprathreshold exposure offset. We combine these two suprathreshold channels by taking a pixel-wise maximum $\max(p_0, p_1)$. This way we generate
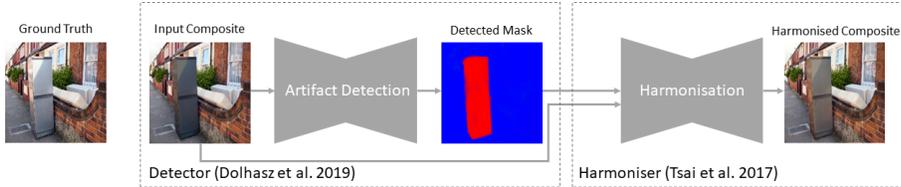
Fig. 2: System overview: illustration of the detector and harmoniser combined into a two-stage composite harmonisation system. A synthetic composite image is first supplied to the detector, which outputs a 2-channel mask indicating detected negative and positive (not pictured here) exposure shifts. This mask is converted to a single-channel representation by taking a maximum over predicted pixel-wise probabilities and fed to the harmonisation network, which then produces a harmonised composite, which we compare against the ground truth. Reprinted from [11].

a single mask of the same resolution as M_s, with the difference that each pixel encodes the probability of a suprathreshold exposure offset. We do not apply any modifications to the DIH and adopt the authors' original trained implementation. The final detector+harmoniser (PTC+DIH) system can be see in Figure 2.

3.4   COCO-Exp Dataset

To perform a fair comparison, we follow the composite generation approach of [30]. Specifically, we sample pairs of images containing objects belonging to the same semantic category (e.g. person, dog, bottle etc.) from the MSCOCO dataset [21]. Using their corresponding object masks, we perform statistical colour transfer based on histogram matching, proposed by [25]. This process can be see in Figure 3. This colour transfer is performed between object regions of the same semantic category. As the detector is only conditioned for exposure offsets, we perform colour transfer only on the luminance channel of Lab colourspace. We generate a total of 68128 composites and corresponding ground truth images. We also extract corresponding ground truth masks for comparison against the masks predicted by the detector. For the sake of brevity, we refer to this dataset as *COCO-Exp* throughout the remainder of this paper.

3.5   Similarity Metrics

To evaluate each of the two approaches, we calculate similarity metrics between ground truth images C_{gt} and harmonised images, corrected by the methods under test: C_s (harmonised using ground truth masks), and C_p (harmonised using predicted masks). We adopt the objective metrics used in the original work, i.e. Mean Squared Error (MSE):
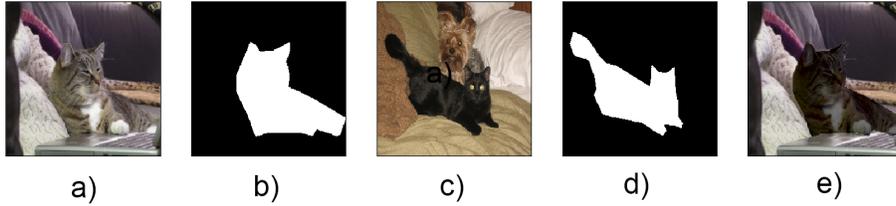
Fig. 3: Dataset generation process adapted from [30]: a) source image sampled from MSCOCO, b) corresponding object mask, c) target image, d) target image object mask, e) result of luminance transfer [25] of source - c), to target - e. Reprinted from [11].

$$MSE = \frac{1}{N} \sum_{i=0}^{n} (Y_i - \hat{Y}_i)^2 \qquad (1)$$

where Y is the ground truth and $\hat{Y}$ is the harmonised image (either C_p or C_s), and Peak Signal-to-Noise ratio (PSNR):

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right) \qquad (2)$$

here R is the maximum possible pixel intensity - 255 for an 8 bit image. In addition, we leverage the Learned Perceptual Image Patch Similarity (LPIPS) [35], which measures similarity based on human perceptual characteristics. We denote these errors with subscripts referring to the method the composite was fixed with, e.g. MSE_p for MSE between the ground truth image and corresponding composite fixed using predicted masks; MSE_s for MSE between ground truth and a composite fixed using the original MSCOCO masks.

3.6   Evaluation Procedure

Using our generated composite dataset we first evaluate the DIH with ground truth masks. We then use the same dataset to generate predicted object masks using the PTC and feed these along with the corresponding composite images to the DIH. We obtain two sets of corrected composites: composites corrected using the ground truth masks C_s and composites fixed using masks predicted by the PTC C_p. We then calculate similarity metrics between the ground truth images used to generate the composites in the first place, and each of the two sets of corrected images C_s and C_p. These are reported in the following section.

4   Two-Stage Model: Results

The results of our evaluation can be seen in Figure 4, which shows distributions of each of the similarity metrics calculated between ground truth images and

composites fixed using C_s and C_p respectively. Mean similarity metrics can be seen in Table 1. Overall, masks predicted by the detector yield higher average errors across all three metrics compared to the ground truth masks, however the magnitude of these differences is small for each of the metrics. Figure 5 shows distributions of image-wise error differentials for both techniques.

Metric	DIH	PTC+DIH
MSE	19.55	22.65
PSNR	35.81	35.18
LPIPS	0.0227	0.0292

Table 1: Means of similarity metrics for both techniques evaluated against ground truth: DIH, and the PTC+DIH. Lower is better for LPIPS and MSE, higher is better for PSNR. Reprinted from [11].
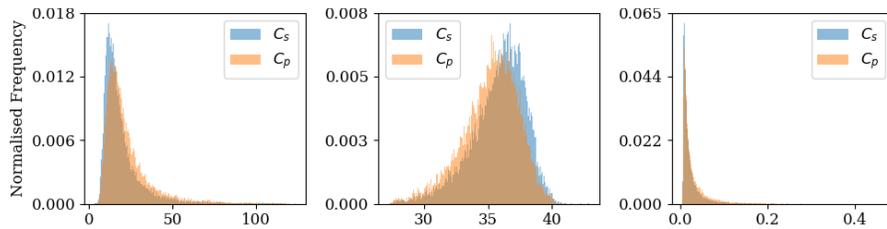

Fig. 4: Similarity metric distributions for both C_s (composites corrected with synthetic ground truth masks) and C_p (corrected with masks predicted by the detector) (a) MSE, (b) PSNR and (c) LPIPS. Larger values indicate poorer performance for MSE and LPIPS, better for PSNR. Reprinted from [11].
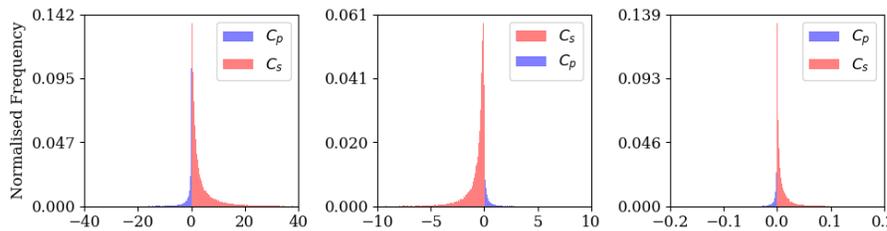

Fig. 5: The image-wise error differentials for C_p-C_s, for each of the three metrics: (a) MSE, (b) PSNR and (c) LPIPS. Note, negative values for MSE and LPIPS indicate images for which C_p (composites corrected with masks predicted by the detector) achieves lower error than C_s (composites corrected with synthetic ground truth masks). For PSNR, the obverse is true. Reprinted from [11].

5   Two Stage Model: Discussion

Our results indicate that using detected, instead of ground truth object masks can yield comparable results when performing automatic image composite harmonisation. Errors obtained using ground truth masks are on average lower compared to those obtained using predicted masks, however in a number of cases the situation is reversed. Figure 6 illustrates examples of failure cases, where Figures $6c$ $6d$ show cases of the DIH over-compensating, while the PTC+DIH combination achieves a more natural-looking result. We stress that these results were obtained with no additional training. Further investigation indicates particular scenarios where this occurs. In some cases, the harmonisation algorithm applies an inappropriate correction, rendering a higher error for C_s compared to the un-harmonised input. Then, if M_p does not approximate M_s well, is blank (no detection) or its average intensity is lower than that of M_s, the additional error induced by the harmonisation algorithm is minimised, rendering lower errors for C_p. This can be seen in both images in $6d$. This indicates the benefit of a perceptually motivated approach to mask prediction, allowing the influence over the weight of the transformation applied by the harmoniser. We also notice that the deep harmonisation network tends to apply colour transformations regardless of whether they are required. In some cases, the perceptually-based masks mitigate this problem. Images showing examples of comparable performance of the two methods can be found in Figure 7. Subfigures c and d show the results of harmonisation using the approaches under test and subfigures e and f show M_p and M_s respectively.
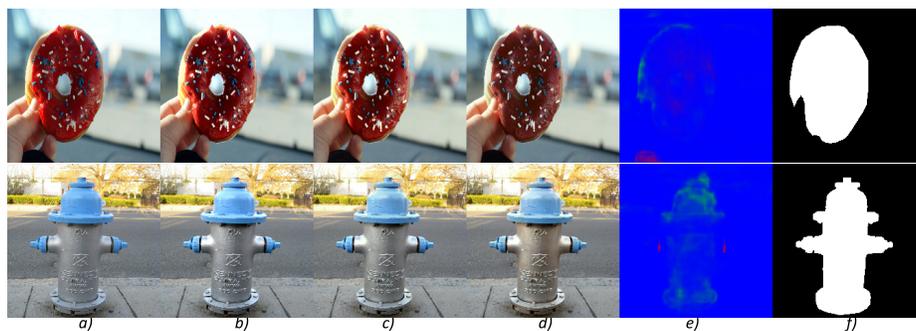

Fig. 6: Examples of the DIH with ground truth masks over-compensating, and applying colour shifts to compensate a luminance transform, resulting in suboptimal output. From left: $a)$ ground truth, $b)$ input composite, $c)$ output of PTC+DIH, $d)$ output of DIH with ground truth masks, $e)$ masks predicted by PTC, $f)$ ground truth masks. Reprinted from [11].

Due to the nature of the PTC currently operating solely on luminance transforms, a further benefit to the multi-task learning paradigm is the generalisability
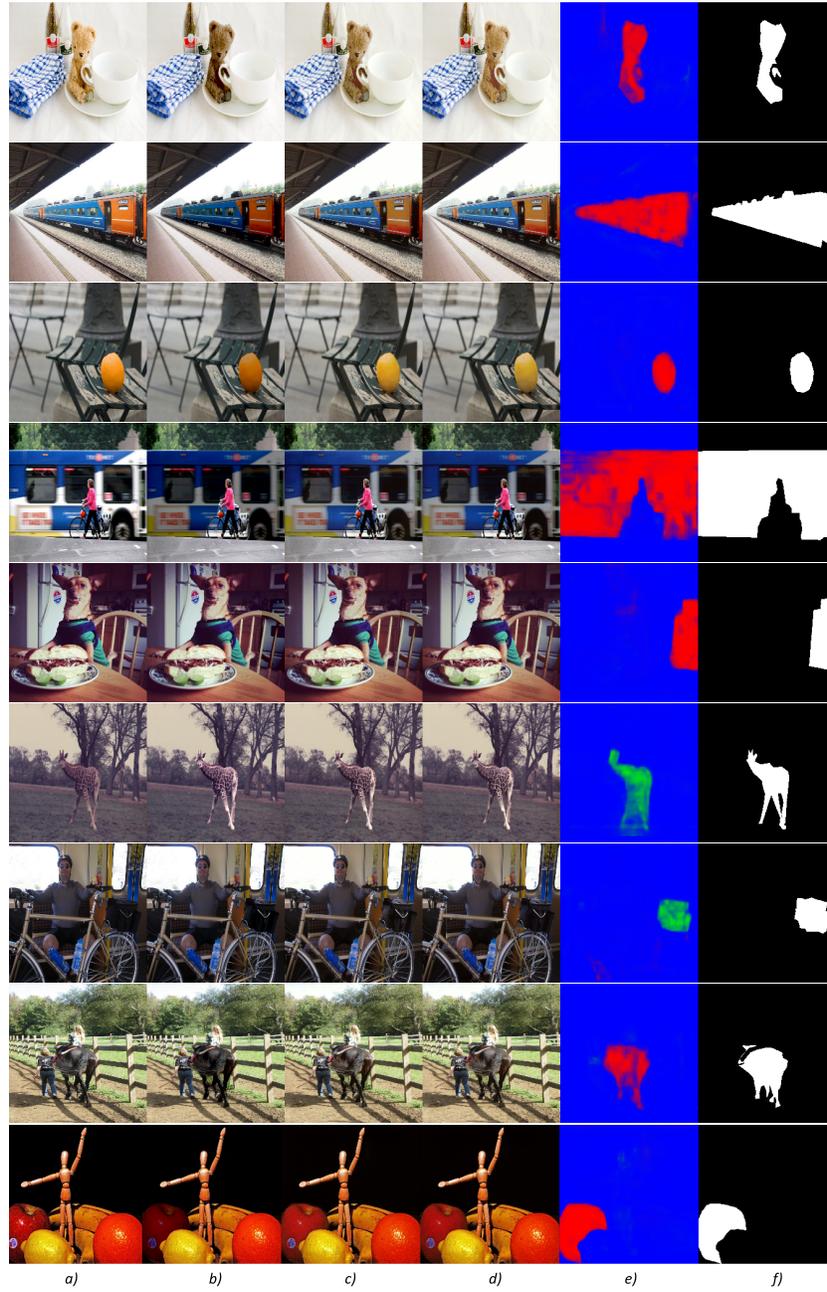
Fig. 7: Comparison of harmonisation outputs from our evaluation. From left to right: a) ground truth, b) input composite, c) corrected with PTC+DIH (C_p), d) corrected with ground truth masks + DIH (C_s), e) Detected masks (M_p), f) ground truth masks (M_s). Masks in colour indicate the raw output of the PTC, where the direction of detected luminance shifts is indicated - red for negative and green for positive shifts. Reprinted from [11].

to arbitrary pixel level transforms, for example colour shifts. The binary masks accepted by harmoniser networks currently do not separate across these transforms, they treat them all homogeneously. A perceptually motivated approach to the predicted mask can encode, on a feature-by-feature basis, the perceptual likelihood of harmonisation required. This is not to say, necessarily, that deep harmonisation networks cannot learn this behaviour, but provision of further support to encode this non-linearity at the input to the network, and/or by explicit optimisation at the output, would likely benefit performance and improve generalisation [6]. This is conceptually similar to curriculum learning improving convergence in reinforcement learning problems [3], or unsupervised pre-training techniques improving convergence in general.

6   End-to-End Model: Methodology

In Section 4 we illustrated that perceptually-based detection of local image transformations can be leveraged to generate composite masks, achieving comparable results to ground truth masks when evaluated on an image harmonisation task using a state-of-the-art harmonisation model. This indicates that an end-to-end model combining both these tasks could be used to perform *no reference* harmonisation, removing the need for provision of object masks for both training and inference, as opposed to current state-of-the-art approaches. Joint training would also allow for overall performance improvements and enable different combinations of the source models to be evaluated. Thus, to perform a fair evaluation, we implement the end-to-end model and the state-of-the-art baseline from scratch, and train both on the iHarmony dataset [8].

6.1   Model Architectures

The end-to-end model is designed by combining the DIH and PTC models. First, we implement the DIH model in Tensorflow, according to the authors' specification and perform random initialisation. We remove one outer layer of the DIH model, following [8], in order to accomodate for the lower resolution of the PTC and perform all training using a resolution of 256×256.

We evaluate two approaches to combining the source models. The first approach, *PTC-DIH* combines the models sequentially, whereby the PTC generates a mask from the input image, which is then concatenated with the input and fed to the DIH model, as illustrated in Figure 2. We replace the original 3-class softmax output of the PTC, and replace it with a single-channel sigmoid output, to match the input of the DIH model. We also add up- and downsampling operations in order to adapt the input image to the 224×224 resolution of the PTC, and its output to the 256×256 input of the DIH.

The second approach, *PTC-att-DIH*, inspired by self-attention mechanisms [31], relies on combining the latent features of both models through an attention-like dot product:

$$a_{joint} = fc_3\Big(\sigma\big(fc_1(a_{ptc})\big)\cdot fc_2(a_{dih})\Big) \tag{3}$$

where a_{ptc} is a vector of flattened activations from the bottleneck layer of the PTC, a_{dih} is a vector of activations from the last convolutional layer of the DIH encoder, fc_n are fully-connected layers with 512 neurons each, and σ is a softmax activation.

In both the PTC-DIH and PTC-att-DIH the encoder of the PTC is frozen during training, as in [10], however in the case of PTC-DIH, the decoder of the PTC is allowed to learn, while in the PTC-att-DIH only the encoder is used. The PTC does not receive any additional supervisory signals, such as ground truth object masks, or scene segmentation, only the end-to-end MSE harmonisation loss.

The performance of our joint model is evaluated against two baselines - the vanilla DIH (without semantic segmentation branch), which requires input masks (*DIH-M*), and a no-mask version of the same model (*DIH-NM*), where masks are not provided as input during training. To ensure a fair comparison, we train all models from scratch, using the same dataset and evaluate their performance on the COCO-Exp dataset from Section 3.4 and the iHarmony validation set. We motivate this by the fact that the original PTC implementation is only conditioned on exposure shifts, so a comparison across both datasets can illustrate the performance for simple exposure shifts (COCO-Exp) versus more complex colour transformations (iHarmony). If the perceptually-based features learned by the PTC generalise well across image features, an improvement should be seen over the naive DIH-NM model when evaluated on both these datasets.

6.2   Optimization Details

All of our models are trained for 50 epochs using the entire training set of the iHarmony dataset, consisting of 65742 training images and evaluated using the validation set, consisting of 7404 validation images. The Adam optimizer [18] with default parameters and an initial learning rate of 0.001 is used. We set the batch size to 32 and enforce a 256×256 resolution. We apply pre-processing to all input images scaling the pixel intensity range from $[0, 255]$ to $[-1, 1]$. For each training run, we select the model minimising validation loss for further evaluation.

7   End-to-End Model: Results

This section presents the evaluation of the proposed models on both the validation set of the iHarmony dataset, as well as the COCO-Exp dataset generated for the preliminary study.

Table 2 shows average MSE and PSNR values for both datasets and each of the models. We find that both of our proposed end-to-end models improve performance on both the iHarmony and COCO-Exp datasets, as compared to

the naive baseline, when performing harmonization with no input mask. This suggests the PTC features are relevant to the image harmonisation task. Overall, the PTC-DIH achieves best performance in harmonisation with no input mask, outperforming the PTC-att-DIH and the DIH-NM baseline.

Model	iHarmony		COCO-Exp	
	MSE	PSNR	MSE	PSNR
DIH-M	89	32.56	201	32.18
DIH-NM	153	30.93	276	31.12
PTC-att-DIH	151	31.02	264	31.37
PTC-DIH	**124**	**31.39**	**214**	**31.61**

Table 2: Test metrics for all evaluated models, across the two datasets used in our experiments. Lower is better for MSE, higher is better for PSNR. Best results using no input mask in bold. Results for the input-mask-based baseline (DIH-M) shown for reference. Higher is better for PSNR, lower is better for MSE.

Figure 8 illustrates the performance of all models under evaluation for several images from the COCO-Exp dataset. Specifically, in each row the input and ground truth are shown in Figures 8a and 8b respectively. Figures 8c, 8e and 8g show the harmonised outputs of the DIH-NM, PTC-att-DIH and PTC-DIH models respectively, while Figures 8d, 8f and 8h are difference image heatmaps between the input and the harmonised output predicted by each model. These heatmaps provide an illustration of the magnitude, direction and location of the applied correction. Upon inspection of similarity metrics, the harmonised outputs and the difference heatmaps, it can be seen that the PTC-DIH model outperforms both the baseline (DIH-NM) and the latent-space-based combination of both models (PTC-att-DIH). This can be seen clearly when comparing the difference images: the PTC-DIH applies corrections more consistently across the region of the target object, compared to the two alternatives. Figure 9 compares the performance of the PTC-DIH to the mask-based DIH-M model for 3 versions of an input image from iHarmony. It can be noticed that the output of both the PTC-DIH and DIH-M closely follow that of the reference. The area corrected by the PTC-DIH aligns with the ground truth mask. Small differeneces in the output images can be noted, particularly around edges, where the PTC-DIH sometimes contribues to softness and smearing (e.g. Fig.9e, middle row). We found this was often related to artifacts around the edges of objects and near edges of images produced by the PTC. Nonetheless, despite the lack of input mask, the PTC-DIH achieves consistent and comparable results for each of the image variations and, in some cases, avoids the colour shifts induced by the DIH (e.g. compare columns d) and e) with column c) of Figure 9), as discussed in Section 5.

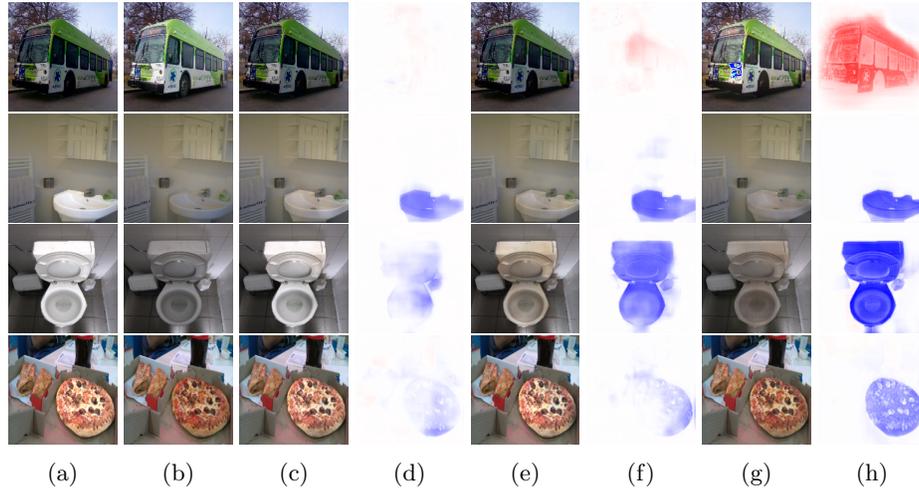| | | | | | | | |
| (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) |

Fig. 8: Comparison of outputs from each model under evaluation for a range of images from the COCO-Exp dataset. *a)* input image *b)* ground truth *c)* DIH-NM result *d)* Difference image between input and output for DIH-NM *e)* PTC-att-DIH result *f)* difference image between input and output for PTC-att-DIH g) PTC-DIH result h) PTC-DIH difference image. In difference images, red indicates that $\hat{y}_{i,j} - x_{i,j} > 0.0$ whereas blue indicates the opposite.

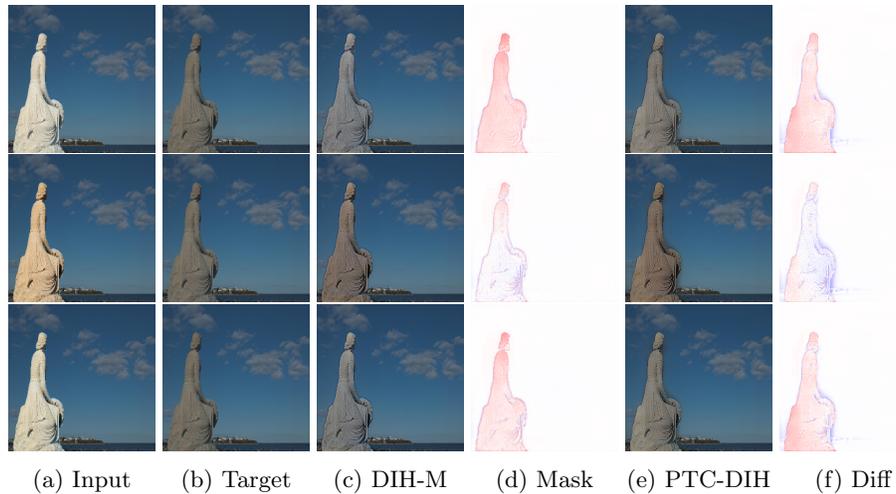| | | | | | |
| (a) Input | (b) Target | (c) DIH-M | (d) Mask | (e) PTC-DIH | (f) Diff |

Fig. 9: Comparison between the corrections applied by PTC-DIH, and the mask-based DIH-M models for multiple variants of the same image. a) input composite, b) ground truth image c) output of DIH-M, d) Difference heatmap between output of DIH-M and ground truth, e) output of PTC-DIH, f) Difference heatmap between output of PTC-DIH and ground truth.

Examples of failure cases can be seen in Figure 10. The top two rows illustrate the most common failure case, where the region requiring harmonisation is not detected, and thus not corrected by the model. The top row illustrates this scenario for a larger object size, while the middle row does so for a small object (one of the sheep near the bottom of the image). The bottom row shows a scenario where the harmonisation is performed on the correct object, however the amount of correction is insufficient. In addition, the model applies harmonisation to a part of the image not requiring harmonisation (the screen). This behaviour is likely due to the fact that the PTC was originally conditioned on exposure shifts, resulting in higher sensitivity to over-exposure, compared to other image distortions.

The impact of object size on harmonisation performance of all models is summarised in Table 3 for both the iHarmony and COCO-Exp datasets. Because the MSE is calculated across the entire image, errors are overall lower for smaller objects. However, when comparing the MSE of harmonised images against their baseline MSE (calculated between the input image and ground truth), the relative MSE improvements are greatest for larger objects. This trend is present across both datasets. The PTC-DIH achieves lowest errors in each object size category across both datasets. Notably, for objects in the COCO-Exp dataset with areas ranging 20-40% of the image size, the PTC-DIH model achieves lower errors than the mask-based DIH-M baseline. This illustrates the impact of the PTC being conditioned on only exposure shifts, but also indicates that these features are useful when transferred to a different type of transformations, such as those in iHarmony.

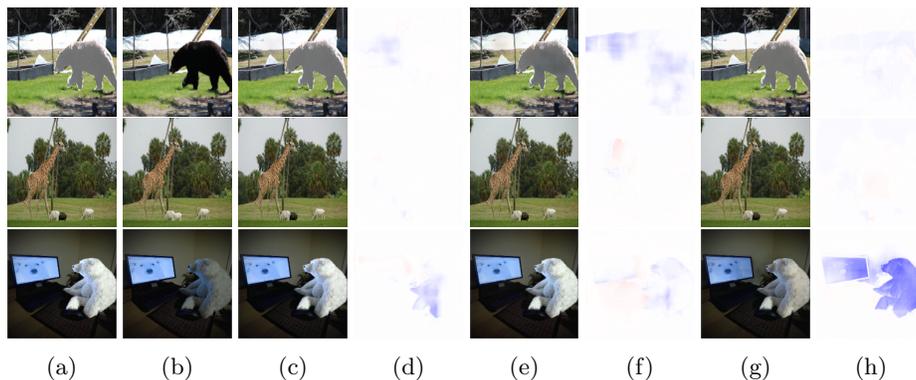(a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)

Fig. 10: Examples of failure cases. *a)* input image *b)* ground truth *c)* DIH-NM result *d)* Difference image between input and output for DIH-NM *e)* PTC-att-DIH result *f)* difference image between input and output for PTC-att-DIH g) PTC-DIH result h) PTC-DIH difference image.

iHarmony

Object Size	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
DIH-M	33.0	116.1	206.5	335.05	456.2	485.48	484.58	705.12
MSE orig.	47.1	235.02	449.84	642.75	1170.31	1222.97	1151.83	1752.12
DIH-NM	50.73	192.22	360.98	497.42	919.29	1058.39	888.11	1534.94
PTC-att-DIH	50.36	190.2	370.65	462.72	884.22	1001.85	933.02	1659.24
PTC-DIH	**45.02**	**150.04**	**311.72**	**359.99**	**623.03**	**895.33**	**720.82**	**1464.62**

COCO-Exp

Object Size	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
DIH-M	73.74	401.55	655.11	785.35	927.68	1042.68	1119.19	1129.01
MSE orig	86.11	524.29	878.42	1131.53	1503.27	1802.57	2072.08	2097.13
DIH-NM	94.3	502.63	828.69	1045.05	1373.97	1661.55	1876.75	1958.01
PTC-att-DIH	93.26	492.65	802.24	986.49	1271.15	1510.16	1684.99	1806.24
PTC-DIH	**82.35**	**410.08**	**647.13**	**778.76**	**946.99**	**1084.28**	**1240.54**	**1295.38**

Table 3: Average MSE on the iHarmony and COCO-Exp datasets for each of the evaluated models, grouped by area of harmonised object as a fraction of image size. *MSE orig* is the MSE between unharmonised inputs and ground truth. Bold values indicate lowest error for each object size, given no mask input. DIH-M model shown for reference.

8   Discussion

The results of both experiments indicate that, in the context of image harmonisation, perceptually-based detection of harmonisation targets can be used to remove the requirement for input object masks. While the proposed approach does not outperform baseline mask-based approaches, it performs significantly better than the state-of-the-art baseline when trained with no input masks. Furthermore, despite the PTC being only conditioned on exposure shifts, its combination with the DIH model improves results on both datasets, suggesting that the perceptually-based features learned by the PTC are useful to the harmonisation task. This is reinforced by the fact that even combining PTC and DIH features in latent space affords a modest improvement over the baseline. Some bias towards exposure shifts is nonetheless noticeable - largest improvements across both datasets occur for achromatic objects (e.g. the sink or toilet in Fig. 8). This could be addressed by training the PTC on a wider range of local transformations. The problem of object size and its impact on harmonisation accuracy is likely connected to the fact that larger objects tend to contribute to the MSE more, compared to smaller objects. The MSE for a small object requiring a 0.5 stop exposure shift will be lower than that of a larger object requiring the same shift. To alleviate this, when training with input masks, the MSE can simply be scaled by the mask size [28], however with no input mask, estimation of target object area becomes nontrivial and presents and interesting direction for further research.

Not unlike the original DIH implementation, the proposed end-to-end model can suffer from gradient artifacts along mask edges, particularly when the initial error to be corrected is large. This issue could be addressed by adopting masked convolutions and utilising self-attention mechanisms, as in [8] or by explicitly incorporating gradient information, as in [33]. While we plan to address these issues in future work, the advantages of our proposed model demonstrated in this work still hold in the context of image harmonisation with no input mask. Following [10], we argue that in order to improve image harmonisation performance, particularly in scenarios where input masks are not available, detection of target regions for harmonisation should leverage intermediate representations equivariant to the transformations of the input to be harmonised. Input masks used in state-of-the-art harmonisation algorithms mimic this role - they encode the presence and location of all input transformations requiring harmonisation as a local binary feature, thus receiving a form of an extra supervisory signal. Our results show that explicitly incorporating the artifact detection paradigm into the harmonisation process can be beneficial, while alleviating the requirements for presence of object masks at inference time.

9   Conclusions & Future Work

In this paper, we have evaluated a novel method for performing image harmonisation without the need for input object masks. Our approach leverages two

state-of-the-art models - an artifact detector and a harmoniser - which, when combined, produce competitive results to mask-based models. We first perform a two-stage evaluation of the original pre-trained models, and based on evaluation results, extend this to a custom end-to-end model in two variants, trained from scratch on the challenging iHarmony dataset. We show that both variants of our end-to-end model outperform the baselines when evaluated on two different datasets. These findings indicate that information about location and magnitude of composite artifacts can be useful in improving the performance of existing compositing and harmonisation approaches. We motivate this by illustrating that ground truth object masks commonly used in harmonisation algorithms essentially substitute the process of detecting local transformations and inconsistencies requiring correction. Accordingly, our results show that the requirement for provision of object masks for such algorithms can be relaxed or removed entirely by the explicit combination of composite artifact detection with their correction. This provides a basis for investigation in future work of joint modeling of both the detection and correction of composite image artifacts, e.g. under a multi-task learning paradigm, where a joint latent representation is conditioned both to be equivariant with respect to input transformations and to encode the structure of the image. In such a scenario, input masks may be used during the training stage, but would not be necessary during inference.

References

1. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. In: ACM Transactions on Graphics (ToG). vol. 23, pp. 294–302. ACM (2004)
2. Azadi, S., Pathak, D., Ebrahimi, S., Darrell, T.: Compositional gan: Learning conditional image composition. arXiv preprint arXiv:1807.07560 (2018)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
4. Burt, P., Adelson, E.: The laplacian pyramid as a compact image code. IEEE Transactions on communications **31**(4), 532–540 (1983)
5. Burt, P.J., Adelson, E.H.: A multiresolution spline with application to image mosaics. ACM transactions on Graphics **2**(4), 217–236 (1983)
6. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
7. Chen, B.C., Kae, A.: Toward realistic image compositing with adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8415–8424 (2019)
8. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8394–8403 (2020)
9. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing **29**, 4759–4771 (2020)
10. Dolhasz, A., Harvey, C., Williams, I.: Learning to observe: Approximating human perceptual thresholds for detection of suprathreshold image transformations.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4797–4807 (2020)

11. Dolhasz., A., Harvey., C., Williams., I.: Towards unsupervised image harmonisation. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,. pp. 574–581. INSTICC, SciTePress (2020). https://doi.org/10.5220/0009354705740581

12. Dolhasz, A., Williams, I., Frutos-Pascual, M.: Measuring observer response to object-scene disparity in composites. In: 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). pp. 13–18. IEEE (2016)

13. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)

14. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117. ACM (2004)

15. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. arXiv preprint arXiv:1704.00090 (2017)

16. Guillemot, C., Le Meur, O.: Image inpainting: Overview and recent advances. IEEE signal processing magazine 31(1), 127–144 (2013)

17. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). pp. 2791–2795. IEEE (2015)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

19. Lalonde, J.F., Efros, A.A.: Using color compatibility for assessing image realism. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)

20. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM transactions on graphics (tog). vol. 23, pp. 689–694. ACM (2004)

21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

22. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on graphics (TOG) 22(3), 313–318 (2003)

23. Porter, T., Duff, T.: Compositing digital images. In: ACM Siggraph Computer Graphics. vol. 18, pp. 253–259. ACM (1984)

24. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1), 121–135 (2017)

25. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications 21(5), 34–41 (2001)

26. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)

27. Shi, W., Loy, C.C., Tang, X.: Deep specialized network for illuminant estimation. In: European Conference on Computer Vision. pp. 371–387. Springer (2016)

28. Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. arXiv preprint arXiv:2006.00809 (2020)

29. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Transactions on Graphics (TOG) **29**(4), 125 (2010)
30. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3789–3797 (2017)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
32. Wright, S.: Digital compositing for film and video. Routledge (2013)
33. Wu, H., Zheng, S., Zhang, J., Huang, K.: Gp-gan: Towards realistic high-resolution image blending. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2487–2495. ACM (2019)
34. Zhang, L., Qi, G.J., Wang, L., Luo, J.: Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2547–2555 (2019)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
36. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European conference on computer vision. pp. 94–108. Springer (2014)