Towards Unsupervised Image Harmonisation

Alan Dolhasz[a], Carlo Harvey[b] and Ian Williams[c]
Digital Media Technology Lab, Birmingham City University, Birmingham, UK
{*alan.dolhasz, carlo.harvey, ian.williams*}@bcu.ac.uk

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

Abstract: The field of image synthesis intrinsically relies on the process of image compositing. This process can be automatic or manual, and depends upon artistic intent. Compositing can introduce errors, due to human-detectable differences in the general pixel level transforms of component elements of an image composite. We report on a pilot study evaluating a proof-of-concept automatic image composite harmonisation system consisting of a state-of-the-art deep harmonisation model and a perceptually-based composite luminance artifact detector. We evaluate the performance of both systems on a large data-set of 68128 automatically generated image composites and find that without any task-specific adaptations, the end-to-end system achieves comparable results to the baseline harmoniser fed with ground truth composite masks. We discuss these findings in the context of extending this to an end-to-end, multi-task system.

1 INTRODUCTION

Image compositing is a common task in image processing where an *object* from one image is extracted and inserted into another image, referred to as the *scene*, with the aim of creating a plausible, realistic result (Wright, 2013). Due to inherent disparities in appearance between the object and scene, commonly resulting from differences in illumination, camera intrinsics, post-processing, encoding or compression, component elements of a composite often require post-processing in order to create a compelling and realistic final result. To address these issues, a wide range of automatic compositing techniques have been proposed. These include alpha matting - linear combinations of object and scene pixel values (Porter and Duff, 1984), gradient-domain optimization techniques (Pérez et al., 2003; Agarwala et al., 2004; Levin et al., 2004), visual appearance transfer (Reinhard et al., 2001; Lalonde and Efros, 2007) and multi-scale methods (Burt and Adelson, 1983a,b; Sunkavalli et al., 2010).

More recently, deep learning (DL) based approaches have achieved considerable success in the domain of image compositing. Notably Tsai et al. (2017) adopt the denoising autoencoder (DAE) (Vincent et al., 2008) attempting to learn the compositing function directly from image data, including semantic information derived from ground truth semantic segmentation labels. Chen and Kae (2019) leverage a generative adversarial network (GAN) to learn both colour-based and geometric transformations in order to perform compositing of arbitrary objects into arbitrary scenes. Conditional GANs have also been adopted to address this problem, by learning to model joint distributions of different object classes and their interactions in image space (Azadi et al., 2018), as well as performing colour and gradient blending between composite elements, which have been semantically aligned.

These existing methods are not without limitations. Firstly, they focus on creation of new composites, thus requiring object/scene segmentation masks to be available at input. This limits their use for cases where these are not available, such as improvement of existing image composites. Secondly, they do not explicitly leverage human perceptual characteristics, such as their sensitivity to various image artifacts or magnitude of mismatch between object and scene (Dolhasz et al., 2016). Finally, the masks supplied to such algorithms provide only a binary indication whether a given pixel belongs to the object or the scene. This implies the entire region requires correction and induces a generic transformation, such as colour transfer, uniformly across the region. This can result in the harmonisation algorithm over-compensating and generating a suboptimal out-

[a] https://orcid.org/0000-0002-6520-8094
[b] https://orcid.org/0000-0002-4809-1592
[c] https://orcid.org/0000-0002-0651-0963

put, even compared to the unprocessed input composite.

We argue that *perceptual detection* of composite artifacts should be *explicitly* modelled in deep-learning-based image compositing and harmonisation. Our reasoning behind this is as follows. Firstly, it enables design of end-to-end harmonisation systems without the need for manually supplied object masks, allowing harmonisation of composites for which masks are not available. Secondly, the explicit encoding of the location and perceptual magnitude of errors in the output of the model allows the process to take advantage of the benefits of multi-task learning in terms of generalisation (Ruder, 2017; Ranjan et al., 2017). 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.

Consequently, in this study, we design a proof-of-concept end-to-end compositing pipeline consisting of a *detector* network, which outputs masks corresponding to regions requiring harmonisation, and a *harmoniser* network, which corrects the detected regions. We then assess the impact of object masks predicted by the detector on the accuracy of the harmoniser, compared to using ground truth object masks. Our study adopts two existing networks - the Deep Harmonisation algorithm proposed by Tsai et al. (2017) as the *harmoniser* network, and a perceptually-based fully convolutional network proposed by Dolhasz et al. (2019) as the *detector* network.

We show that our prototype end-to-end system, using the the detector network without any task-specific adaptations or re-training, produces results which are comparable to those obtained using ground truth masks. To our knowledge this is the first work investigating the combination of a deep-learning-based detection model with a composite harmonisation one to both detect and fix composites. We are currently developing a complete, end-to-end version of the model, trained specifically for this purpose.

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 and processing. Image compositing concerns the entire process of combining regions from different source images into a plausible whole, while 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 (Sunkavalli et al., 2010).

Similarly to the problem of image in-painting, compositing and harmonisation are both ill-posed problems (Guillemot and Le Meur, 2013). 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 inferable from the information contained within the image or its metadata, such as the characteristics of the illuminant (Shi et al., 2016), 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 approximating these do exist (Gardner et al., 2017), they are difficult to integrate into end-to-end systems and can be difficult to parameterise. The recent successes in DL have motivated a number of approaches (Tsai et al., 2017; Azadi et al., 2018; Chen and Kae, 2019) 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

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. Two popular DL paradigms, unsupervised learning and multi-task learning, are often used to address the issues of data labeling and model generalisation.

In recent years many tasks in image understanding have achieved state-of-the-art performance by incorporating multi-task learning Evgeniou and Pontil (2004), for example in predicting depth and normals from a single RGB image (Eigen and Fergus, 2015), detection of face landmarks (Zhang et al., 2014) or simultaneous image quality and distortion estimation Kang et al. (2015). This is afforded by the implicit regularization that training a single model for multiple related tasks imposes (Caruana, 1997) and the resulting improved generalisation.

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. Tsai et al. (2017) use a DAE-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, they do not explicitly condition the network to learn anything more about the corruption, such as its magnitude, type or location. Instead they provide location information at input time, using a binary mask. Chen and Kae (2019) uses a similar approach - inputting the object mask at training time, however also introducing mask segmentation and refinement within the GAN, in addition to geometric transformations. 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 DAE, 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 assess such differences before attempting to correct them. Kang et al. (2015) shows that a multi-task approach is an efficient way to ensure that distortions are encoded by the model.

3 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 design an end-to-end approach using two existing standalone networks for both detection and harmonisation to test the efficacy of these perceptual masks in the 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 propose to 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.

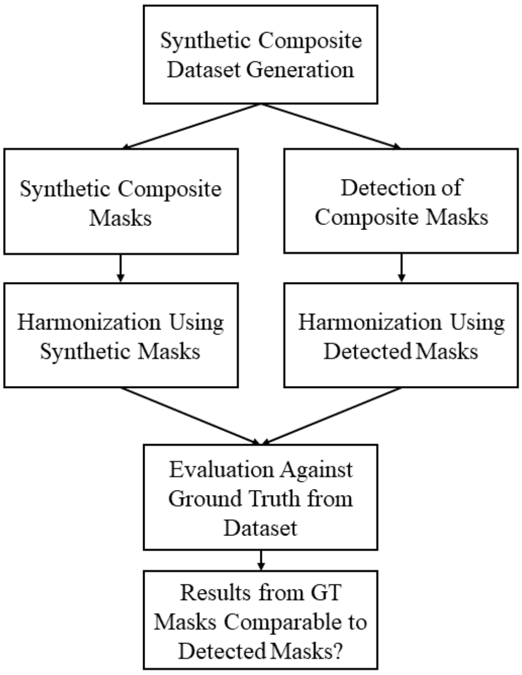
Synthetic Composite Dataset Generation
Synthetic Composite Masks
Detection of Composite Masks
Harmonization Using Synthetic Masks
Harmonization Using Detected Masks
Evaluation Against Ground Truth from Dataset
Results from GT Masks Comparable to Detected Masks?

Figure 1: Illustration of research methodology adopted in this work.

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 (Dolhasz et al., 2019) and the harmoniser (Tsai et al., 2017) are deep, image-to-image, fully convolutional autoencoder networks. The detector takes a single image as input and generates a 2-

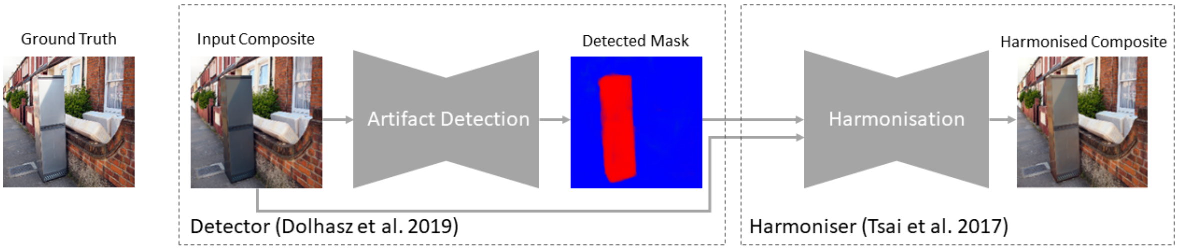

Figure 2: System overview: illustration of the detector and harmoniser combined into an end-to-end composite harmonisation system. A synthetic composite image is first supplied to the detector, which outputs a 2-channel mask indicating detected negative and positive (not pictured here) luminance shifts. This mask is converted to a single-channel representation by taking a maximum over predicted pixel-wise probabilities and fed to the harmoniser network. The harmoniser then produces a harmonised composite, which we compare against the ground truth.

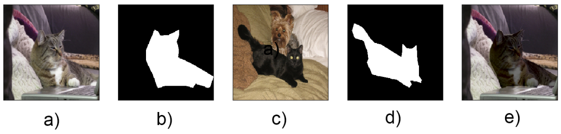

Figure 3: Dataset generation process adapted from Tsai et al. (2017): a) source image sampled from MSCOCO, b) corresponding object mask, c) target image, d) target image object mask, e) result of luminance transfer (Reinhard et al., 2001) of source - c), to target - e

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 luminance offset. We combine these two suprathreshold channels by taking a pixel-wise maximum $\max(p_0, p_1)$. This way we generate a single mask in the same format as M_s from MSCOCO, where each pixel encodes the probability of a suprathreshold luminance offset. We do not apply any modifications to the harmoniser and adopt the authors' original trained implementation. The final detector+harmoniser system can be see in Figure 2.

3.4 Dataset

To perform a fair comparison, we follow the composite generation approach of Tsai et al. (2017). Specifically, we sample pairs of images containing objects belonging to the same semantic category (e.g. person, dog, bottle etc.) from the MSCOCO dataset (Lin et al., 2014). Using their corresponding object masks, we perform statistical colour transfer based on histogram matching, proposed by Reinhard et al. (2001). 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 luminance 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 the ground truth masks for comparison against the masks predicted by the detector.

3.5 Similarity Metrics

To evaluate each of the two approaches, we calculate similarity metrics between ground truth images C_{gt} and composites corrected by the methods under test: C_s and C_p. We adopt the objective metrics used in the original work, i.e. Mean Squared Error (MSE):

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

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):

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

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) (Zhang et al., 2018), 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 Procedure

Using our generated composite dataset we first evaluate the harmoniser with ground truth masks. We

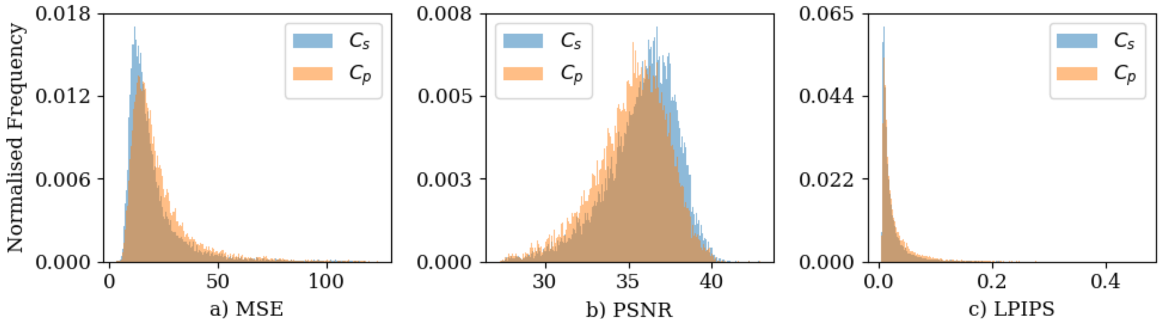

Figure 4: A comparison of the distributions for both C_s (composites corrected with synthetic ground truth masks) and C_p (corrected with masks predicted by the detector) with the number of images in each bin for each metric value. This is shown for: (a) MSE, (b) PSNR and (c) LPIPS. Larger values of MSE and LPIPS indicate poorer performance, whilst this is true for smaller values of PSNR.

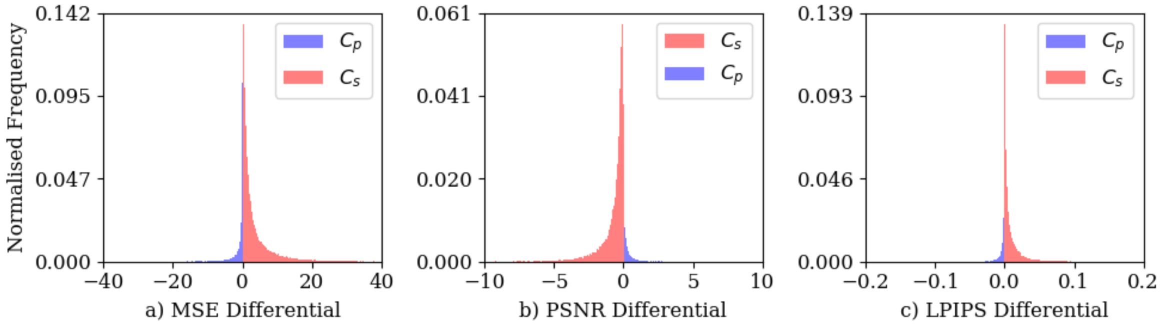

Figure 5: The image-wise error differentials for C_p-C_s. This is shown for each of the three metrics: (a) MSE, (b) PSNR and (c) LPIPS. Note, negative values for MSE and LPIPS indicate images for which C_p (composites corrected with masks predicted by the detector) achieves lower error than C_s (composites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

then use the same dataset to generate predicted object masks using the detector and feed these along with the corresponding composite images to the harmoniser. We obtain two sets of corrected composites: composites corrected using the ground truth masks C_s and composites fixed using masks predicted by the detector] 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 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	harmoniser	detector + harmoniser
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: harmoniser, and the detector+harmoniser. Lower is better for LPIPS and MSE, higher is better for PSNR.

5 DISCUSSION

Our results indicate that using detected, instead of ground truth object masks can yield comparable results when performing automatic image composite

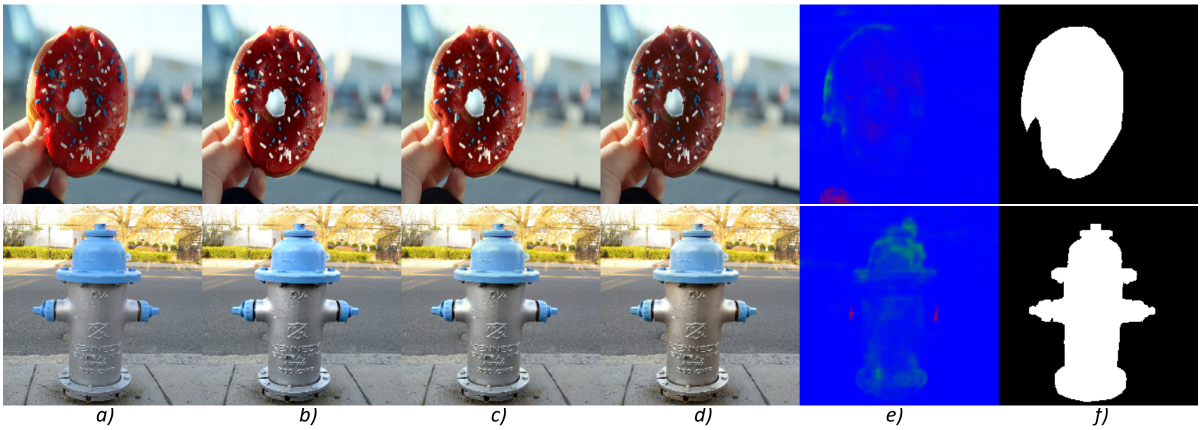

Figure 6: Examples of the harmoniser with ground truth masks over-compensating, and applying colour shifts to compensate a luminance transform, resulting in suboptimal output. From left: *a)* ground truth, *b)* input composite, *c)* output of detector+harmoniser, *d)* output of harmoniser with ground truth masks, *e)* masks predicted by detector, *f)* ground truth masks

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. For example, Figure 6*c* and *d* shows cases of the harmoniser over-compensating, while the detector+harmoniser 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 unharmonised 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 6*d*. 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 apporaches under test and subfigures *e* and *f* show M_p and M_s respectively.

Due to the nature of the detector network currently operating solely on luminance transforms, a further benefit to the multi-task learning paradigm is the generalisability 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 further support to encode this non-linearity at the input to the network and/or by explicit optimisation at the output, particularly in a multi-task context, would likely benefit performance and improve generalisation (Caruana, 1997).

6 CONCLUSION

These findings, obtained by combination of off-the-shelf models, not modified or re-trained for this specific task, indicate that information about location and magnitude of composite artifacts can be useful in improving the performance of existing compositing and harmonisation approaches. Furthermore, 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 modelling of both the detection and correction of composite image artifacts, e.g. under a multi-task learning paradigm.

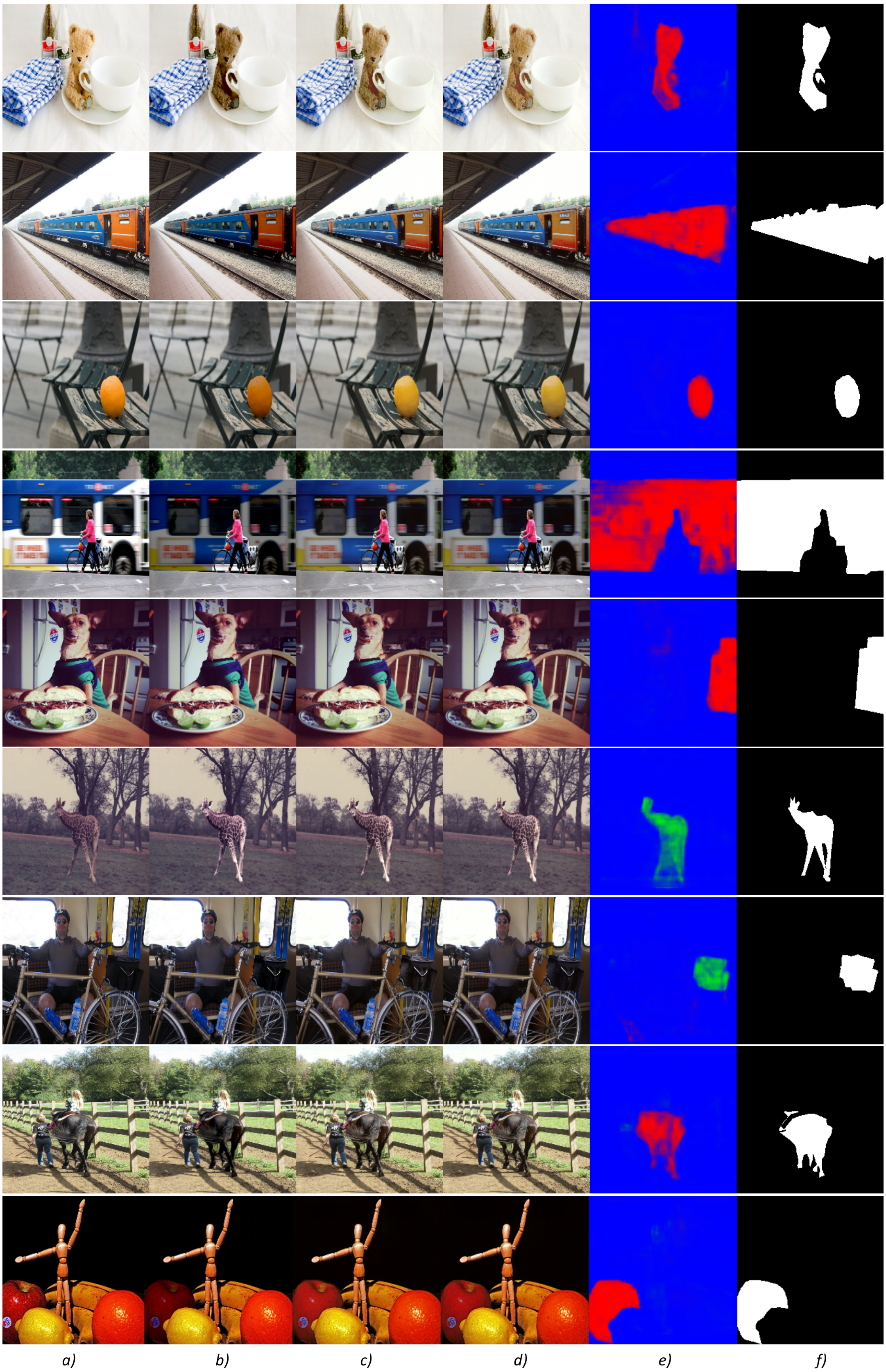

Figure 7: Comparison of harmonisation outputs from our evaluation. From left to right: a) ground truth, b) input composite, c) corrected with detector+harmoniser C_p, d) corrected with ground truth masks + harmoniser C_s, e) Detected masks M_p, f) ground truth masks M_s. Masks in colour indicate the raw output of the detector, where the direction of detected luminance shifts is indicated - red for negative and green for positive shifts.

REFERENCES

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., and Cohen, M. (2004). Interactive digital photomontage. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 294–302. ACM.

Azadi, S., Pathak, D., Ebrahimi, S., and Darrell, T. (2018). Compositional gan: Learning conditional image composition. *arXiv preprint arXiv:1807.07560*.

Burt, P. and Adelson, E. (1983a). The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540.

Burt, P. J. and Adelson, E. H. (1983b). A multiresolution spline with application to image mosaics. *ACM transactions on Graphics*, 2(4):217–236.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chen, B.-C. and Kae, A. (2019). Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Dolhasz, A., Harvey, C., and Williams, I. (2019). Learning to observe: Approximating human perceptual thresholds for detection of suprathreshold image transformations. *arXiv preprint arXiv:1912.06433*.

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

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

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

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

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

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

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

Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. In *ACM transactions on graphics (tog)*, volume 23, pages 689–694. ACM.

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

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

Porter, T. and Duff, T. (1984). Compositing digital images. In *ACM Siggraph Computer Graphics*, volume 18, pages 253–259. ACM.

Ranjan, R., Patel, V. M., and Chellappa, R. (2017). 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.

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

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

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

Sunkavalli, K., Johnson, M. K., Matusik, W., and Pfister, H. (2010). Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):125.

Tsai, Y.-H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., and Yang, M.-H. (2017). Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Wright, S. (2013). *Digital compositing for film and video*. Routledge.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer.