Super-Resolution Appearance Transfer for 4D Human Performances

Marco Pesavento Marco Volino Adrian Hilton
Centre for Vision, Speech and Signal Processing
University of Surrey, UK
{m.pesavento,m.volino,a.hilton}@surrey.ac.uk

Abstract

A common problem in the 4D reconstruction of people from multi-view video is the quality of the captured dynamic texture appearance which depends on both the camera resolution and capture volume. Typically the requirement to frame cameras to capture the volume of a dynamic performance ($> 50m^3$) results in the person occupying only a small proportion $< 10\%$ of the field of view. Even with ultra high-definition 4k video acquisition this results in sampling the person at less-than standard definition 0.5k video resolution resulting in low-quality rendering. In this paper we propose a solution to this problem through super-resolution appearance transfer from a static high-resolution appearance capture rig using digital stills cameras ($> 8k$) to capture the person in a small volume ($< 8m^3$). A pipeline is proposed for super-resolution appearance transfer from high-resolution static capture to dynamic video performance capture to produce super-resolution dynamic textures. This addresses two key problems: colour mapping between different camera systems; and dynamic texture map super-resolution using a learnt model. Comparative evaluation demonstrates a significant qualitative and quantitative improvement in rendering the 4D performance capture with super-resolution dynamic texture appearance. The proposed approach reproduces the high-resolution detail of the static capture whilst maintaining the appearance dynamics of the captured video.

1. Introduction

The increasing popularity of VR/AR technologies has driven a rise in the demand for 4D modelling techniques that accurately reconstruct human performances. Video-based capture systems have been developed recently to replace marker-based motion capture enabling the reconstruction of the detailed surface dynamics for complex non-rigid shapes such as people. These systems consist of a set of cameras that capture a scene from multiple viewpoints [12,14,47]. A 4D reconstruction of the performance of the captured scene can then be retrieved by applying stereo matching and feature tracking approaches. This reconstruction is a sequence of 3D geometric objects of the performer, each with its own pose and texture appearance. Unless the acquisition system has a large number of cameras [12,20,28,40,49] or the overall volume to capture is limited, the surface texture resolution resolution is limited [24]. In general, the resolution of the captured models decreases proportionally to the increasing of the capture volume size. Moreover, the colours of the appearance are significantly influenced by the lighting of the capture space, which can deteriorate the appearance if the illumination is not thoroughly controlled.

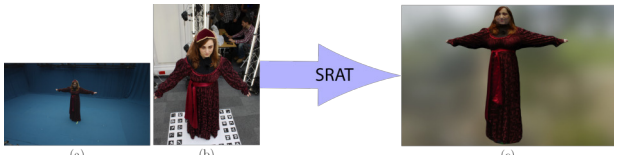


Figure 1: Overview of the SR transfer pipeline: a) Low-resolution capture frame; b) High-resolution image; c) SR result.

Whilst 4D reconstruction systems still face a number of limitations, there are alternative methods that accurately provide a static 3D shape reconstruction from multiple high-resolution DSLR camera images in a limited capture volume giving higher detail reconstruction of surface shape and appearance [7,8,30,53]. The resolution of shape detail reconstruction is dependent on the image resolution and the 3D reconstruction systems use high-resolution (HR) cameras that can acquire higher resolution images of the subject. The quality of the reconstruction is highly influenced by the capture volume, which is usually much smaller for 3D reconstruction since the scene to capture is static. The colour response of the acquisition cameras depends on the capture environment, which is designed considering the objective of the capture. The system settings of the two reconstructions differ due to their dissimilar objectives, producing different colour responses that cause brighter colours in the static reconstruction. These factors lead to higher quality appearance of the reconstructed model in a static capture.

In this paper we introduce Super-Resolution Appearance Transfer (SRAT) from a human subject acquired with DSLR cameras in a small capture volume to dynamic video perfor-

mance capture of the same subject acquired with a sparse set of cameras in a larger capture volume. The objective of SRAT is to enhance the appearance of the 4D reconstruction of a human performance captured in a large volume by exploiting HR images of the same subject acquired with the DSLR cameras. More specifically, the approach improves the colour contrast of the appearance of the 4D reconstruction with a novel colour mapping approach and enhances its fine details by increasing the resolution of the texture maps through a Single-Image Super-Resolution (SISR) network. Figure 1 shows how, from the input low-resolution (LR) capture video (Figure 1a), the final appearance of the 3D model (Figure 1c) is obtained with appearance detail similar to the HR capture (Figure 1b). The contributions presented in this paper are:

- A novel pipeline for super-resolution appearance transfer to enhance the appearance of LR dynamic performances of people from HR images of the same subjects acquired with static DSLR cameras in a small volume.
- A new automatic approach for colour mapping between images acquired with different systems: after the selection of optimal image couples, the colours of the video frames are corrected with an extension of the colour transfer algorithm [19] to multi-view images.

2. Related Works

Colour transfer: colour transfer aims to modify the colours of a target image taking as reference the colours of another image. For some frameworks, there must be pixel correspondences between target and reference images. Examples of these are classic computer vision methods [17, 25, 39, 41] and deep learning approaches such as image-to-image translation networks [26, 51]. Due to the necessity of pixel correspondences, these methods cannot be applied in our case. Early techniques that do not require pixel correspondences define a parametric affine transfer function through statistical moments of colour distributions. These can model only certain types of distributions [42,45]. Methodologies that use Optimal Transport (OT) framework were then studied but they either introduce grainy artefacts in the gradient of the corrected image [16, 18, 44] or do not modify the image's luminance channel [9,43]. Recent techniques model the colour distribution as Gaussian Mixture models to define correspondences between Gaussian components of the target and reference distributions [19,27,52]. The colour transfer function is learned from a single target-reference image pair, failing in learning a complete map for all the colours of a 3D surface. Unsupervised cycle-in-cycle neural networks are used to enhance the colours of images as well [10, 22, 64]. The lack of a consistent amount of training data deteriorates their performances. This paper proposes a colour transfer algorithm that exploits the whole surface of a human model to learn the transfer function from multiple target and reference images without the need of pixel correspondences between them.

Single-image super-resolution: image super-resolution (SR) is an image processing techniques that aims to estimate a perceptually plausible HR image from a LR input image [54]. Recently, deep neural networks have shown their superior performance on the SISR task. Dong *et al*. [15] introduced the use of a convolutional neural network (CNN) to super-resolve an image for the first time. Two main topologies of network architecture were then applied for SR: residual networks [4, 13, 21, 36, 37, 38, 60] and generative adversarial networks [11, 32, 33, 48, 50]. While the former achieve higher values of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) with blurrier outputs, the second have lower figures of the mentioned metrics and unrealistic details in the SR images. Reference-based super-resolution (RefSR) is an alternative method that transfers HR textures from a given reference image to super-resolve the LR input [34]. CrossNet [62] uses optical flow to align input and reference. To be efficient, the reference and the LR images must have similar content and similar viewpoint. Other approaches [61], [55] apply a "patch-match" mechanism to swap the most similar features of the reference and the LR image. Even though they perform better when the similarity between reference and input image is low, they still introduce unpleasant artefacts. The patch-match mechanism consumes significant amount of GPU memory, making the usage of HR reference images impracticable.
To the best of our knowledge, there are only two deep learning works that aim to super-resolve 3D texture maps of objects. Li *et al*. [35] modified the architecture of EDSR [37] to exploit the information of both texture maps and normal maps of objects. Richard *et al*. [46] combined a redundancy-based part with a prior-based part in a network to create new texture maps. The first method requires the creation of normal maps, a process that introduces heavy computational cost for a high number of frames. The second approach is mainly oriented in the creation of texture maps. Following these two methods, we use a residual neural network in SRAT that, differently from 2D SISR networks, is trained with datasets of texture maps of people.

3. Methodology

SRAT aims to improve the appearance of LR capture video of human performance acquired with a sparse multi camera system by leveraging a collection of HR images of the same subject. The sparse multi camera system allows dynamic capture of human performance that covers a large volume at the cost of reduced resolution on the subject. In contrast, the static HR images are captured using an array of DSLR cameras resulting in increased resolution on the subject including fine details in the hair, skin and clothing.

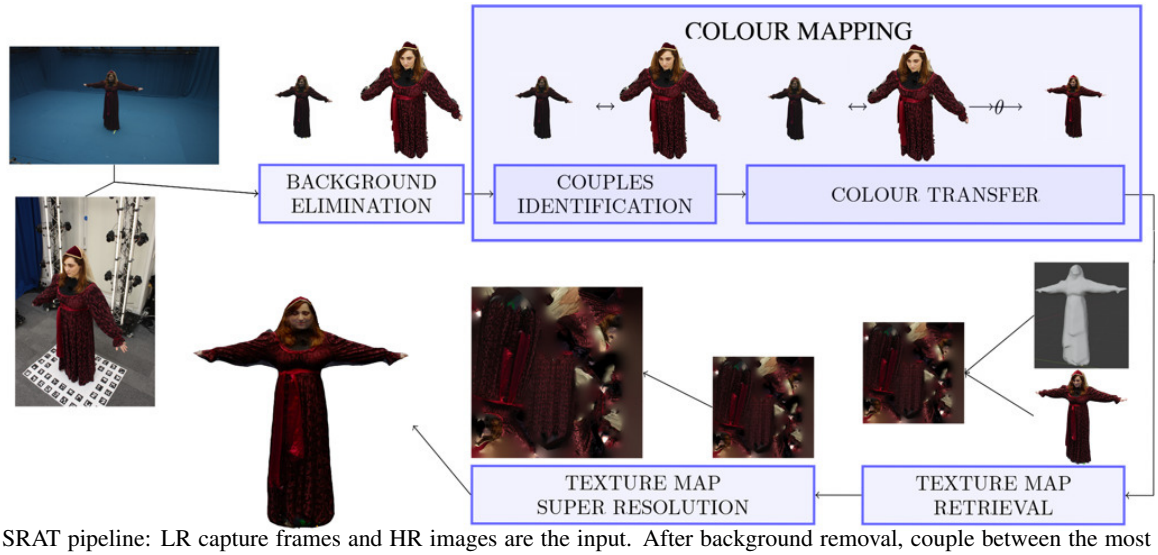


Figure 2: SRAT pipeline: LR capture frames and HR images are the input. After background removal, couple between the most similar frames and HR images are created. The colours of the frames are corrected. Their texture maps are retrieved and finally super-resolved.

To globally improve the appearance of the performance, we first tackle the problem that these systems will have a different capture settings and colour response by automatically learning a colour transfer function between the two systems (Section 3.2). The fine details of the dynamic LR texture maps are then enhanced by leveraging a learnt SR model and thus, the appearance is locally improved (Section 3.3).

3.1. Overview

The input data of the proposed approach are:

- LR capture video with N_{LR} camera. This consists of: RGB image set of a subject $\{\{I^i_{LR_l}(t)\}_{l=1}^{N_F}\}_{i=1}^{N_{LR}}$; N_F 3D meshes (one for every frame); N_F texture maps;
- HR image set of the same subject captured with a rig of N_{HR} DSLR cameras $\{I^j_{HR}\}_{j=1}^{N_{HR}}$

Without loss of generality, I_{LR} is a time instant of $I_{LR}(t)$. SRAT consists of 4 stages shown in Figure 2 and outlined below.

Background elimination: to ensure an accurate result of the colour mapping, the background of the frames and HR images must be removed. For the former, we use silhouettes computed via Chroma keying while for the latter the alpha matting method proposed by Hu and Clark [23] is applied.

Colour mapping: the second stage of the pipeline is a novel approach to map the colour features between images acquired with different systems. It consists of two steps. (i) *Couple identification:* each I_{LR} is paired with one I_{HR} to create couple of similar images by performing the similarity evaluation among partial texture maps. (ii) *Colour transfer:* a colour transfer function is learned from the multiple couples of the previous step to correct the colours of all I_{LR}.

Texture map retrieval: texture maps are retrieved by projecting the new corrected frames to the corresponding meshes reconstructed in a pre-processing step.

Texture map SR: the details of the texture maps are enhanced by super resolving them with an RCAN-style network [60] trained with human texture maps.

We make 3 important assumptions that affect the design of the pipeline and its evaluation: (i) geometric properties of the HR static model are not exploited to avoid estimation of geometric surface correspondence between the LR and HR reconstructions; (ii) the color response of the cameras within each capture system is the same; (iii) in the selection of a subset of $n < N_{LR}$ cameras for evaluation we assume evenly spaced cameras to maximise coverage of the subject.

3.2. Colour mapping of frame images

The camera responses of the two capture systems differ due to their environment settings that are defined based on the specific purposes of the captures. The contrast range of the DSLR cameras allows to acquire HR images with brighter colours and higher resolution. To obtain the same contrast in the LR capture video, the colours of the HR images are mapped to its frames with a novel approach that comprises 2 steps: pairs between the most similar images of the two systems are created according to surface visibility and viewing angle pairs (couples identification); colours of the LR capture frames are corrected with a colour transfer function estimated from multi-view images (colour transfer).

Couples identification: the algorithm that learns the colour transfer function to correct the colours of the frame images, requires a pair of target and reference images as input. In our case, the target images are the LR capture frames while the reference images are the HR images. Since a colour palette is defined from the reference images, the algorithm performs better if the content of I_{LR} is similar to the content of I_{HR}. If for instance, a reference image that shows the back of the person has been paired with a target frame representing the front part of the subject, the algo-

rithm may fail to transfer the colour of the face because it is not learned from the visible colour palette (Section 4.2). In the two systems, the camera settings are different, resulting in different representations of the same model. Creating the pairs by directly comparing the original images is ineffective and thus, we operate in the texture map domain. We propose a new automatic method to identify the couples between images of different systems. Since no geometric information of the HR model is available, we reconstruct partial texture maps of the frames $\{\{T(I^i_{LR_l})\}_{l=1}^{N_F}\}_{i=1}^{N_{LR}}$ and of the HR images $\{T(I^j_{HR})\}_{j=1}^{N_{HR}}$ by applying Densepose [6] and unwrapping the resulting UV maps. These partial texture maps are invariant to the camera orientation and position allowing a comparison between the two systems in a common domain. We evaluate the similarity between all the $T(I^i_{LR_l})$ and $T(I^j_{HR})$ with the SSIM metric [63]. Each I_{LR} is coupled with one I_{HR} whose partial texture map is the most similar defined by the SSIM metric as shown in Equation 1:

$$I^i_{LR_l} \leftrightarrow I^j_{HR} \; where \; \underset{I^j_{HR}}{\text{argmax}}\{SSIM(T(I^i_{LR_l}), T(I^j_{HR}))\} \quad (1)$$

where $I^i_{LR_l}$ is the l^{th} frame of the i^{th} camera and I^j_{HR} is the HR image of the j^{th} camera.

Colour transfer: the idea of the colour transfer algorithm applied is based on [19]. In their work, the parameters of the colour transfer function are learned from a single pair of target and reference images. In our case, the function must learn a map from all the colours of the static acquisition to the colours of I_{LR}: all the view-angles of the human model must be seen during the learning stage. We therefore extend the cited work [19] to more than two images as inputs. We model the colour transfer function $\phi_\theta(x)$ as a Thin Plate Splines function that depends on a set of parameters θ. This set is computed by minimising the following novel energy function with gradient descent algorithm:

$$\theta = \frac{1}{N}\sum_{l=1}^{N} \underset{\theta_l}{\text{argmin}}\{||p_f||^2 - 2 < p_f | p_I >\} \quad (2)$$

where N is the number of input couples, p_f is the distribution of I_{LR} with parameterised mean $\phi_\theta(\mu_f)$ and p_I is the distribution of I_{HR}. Equation 2 is further explained in the supplementary material. For each selected input couple, a set of parameters is computed by minimising the energy function. The parameters of $\phi_\theta(x)$ are obtained averaging these sets and used to correct the colours of all the frames.

3.3. Texture map super-resolution

After having retrieved the texture maps from the corrected frames, we treat them as 2D RGB images and we give them as input to the RCAN-style network [60]. The residual-in-residual RCAN network super-resolves the obtained texture maps through a channel attention mechanism that adaptively rescale each channel-wise feature by modelling the interdependencies across feature channels. In its original work, RCAN was trained augmenting 800 RGB images from DIV2K dataset [3]. Since we aim to super-resolve texture maps of a specific model, we first pre-train RCAN with a set of patches of human texture maps [2]. We then fine-tune it with the N_F original texture maps of the input model. Further details of its training are presented in the supplementary material. We finally assign the SR texture maps to the correspondent reconstructed meshes to obtain the enhanced 4D reconstruction of the performance.

4. Results evaluation

We capture two subjects, one male (*SingleM*) and one female (*SingleF*), using both LR capture video acquired with $N_{LR} = 16$ cameras for $N_F = 440,470$ frames respectively and a subset of $N_{HR} = 64$ DSLR cameras. We evaluate the proposed approach on the 4D performances of the models reconstructed with the method proposed by Starck and Hilton [47], applied in the pre-processing step to retrieve the input data. Additional visual results and evaluations are presented in the supplementary material.

4.1. Couples identification

The first study was conducted on the first part of the colour mapping stage. We evaluate the effect of using different similarity metrics to pair the partial texture maps of the two acquisition systems. We compute the similarity with the following metrics: SSIM [63], PSNR, feature based similarity index (FSIM [58]), spectral angle mapper (SAM [57]), signal to reconstruction error ratio (SRE [31]), root mean square error (RMSE). Figure 3 shows the partial texture maps of the HR images associated with the partial texture map of the LR capture frames and the correspondent original images for each of the studied metrics. In the case of *SingleF*, most of the metrics create a pair with two images where the pose of the subject is significantly different. SSIM associates the most similar HR image to the input frame. The orientation of *SingleM* in the HR images paired with other metrics do not match the one in the input frame. In these HR images, the nose of the subject is hardly visible oppositely from the HR image paired with SSIM.

4.2. Colour transfer input couples selection

We analyse the effect of using different numbers of input couples in the colour transfer step. We expect the algorithm to perform better if a complete coverage of the surface is exploited in the learning of the colour transfer function. The effect of having 1, 2, 4, 8, 12 and 16 couples as input is evaluated. The N_{LR} cameras are equally distributed in a studio, with 4 cameras on each side of a square surrounding the acquisition space. In the case of 12 couples, 3 cameras from each side are selected; in the case of 8, 2 cameras each side; in the case of 4, just 1 camera each side. With

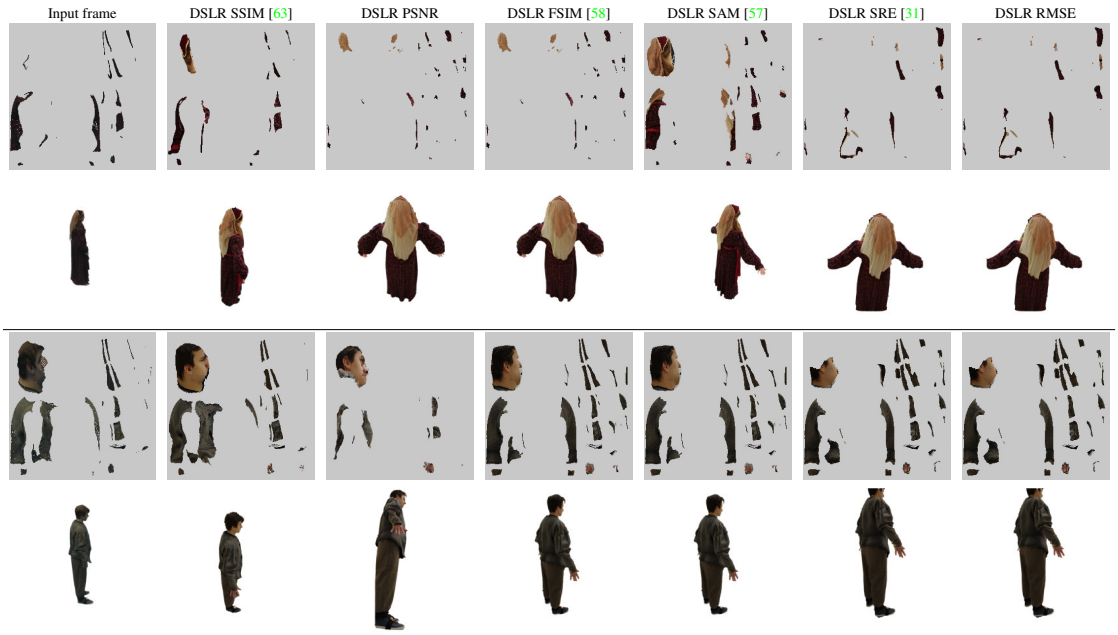


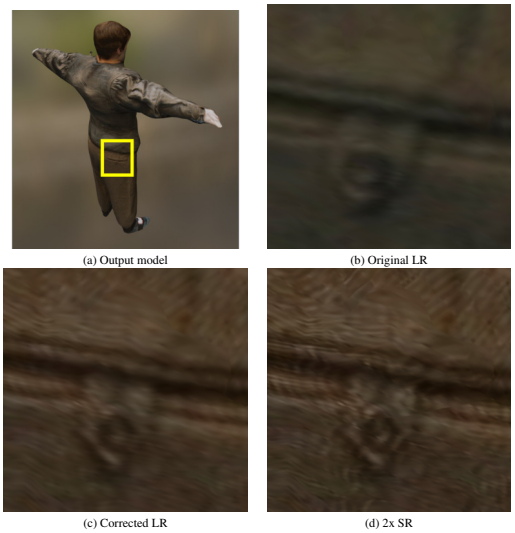
Figure 3: The first column shows the partial texture map and the LR capture frame for *SingleF* (first two rows) and *SingleM* (last two rows). The images and partial texture maps of the other columns are acquired with the DLSR system and paired with different similarity metrics.

Figure 4: Colour transfer with different numbers of input couples. In the top row *SingleF* model, in the bottom *SingleM*.

Input couples	**SingleF**		**SingleM**	
	JS ↓	χ^2 ↓	JS ↓	χ^2 ↓
1	0.5452	18.8659	0.4759	21.5098
2	0.5439	14.7575	0.4691	18.3891
4	0.5435	14.9826	0.4649	22.1790
8	0.5423	13.5905	0.4609	18.0991
12	**0.5321**	**12.9507**	**0.4542**	**15.5898**
16	0.5443	14.0894	0.4624	21.3688

Table 1: JS and χ^2 values for different numbers of input couples. ↓ indicates lower value is better.

less input couples (1 and 2), there is a lower surface coverage and the model is not completely captured. As quantitative evaluation, the dissimilarity between a specific corrected frame and its paired HR image is computed considering their colour histograms since the pixel-based metrics are not effective due to the different nature of the frame and the HR image. We compute the Jensen-Shannon (JS) divergence and the Chi-Squared (χ^2) distance between the normalized colour histograms of all the corrected frames and of the correspondent HR images. Even though these two metrics have been revealed to be the most efficient [59], they present some drawbacks. JS focuses only on the statistical distribution of the images while χ^2 only accounts for the difference between the corresponding bins and is hence sensitive to distortions and quantization [56]. The average of the distances for every corrected couple is presented in Table 1. Generally, the JS divergence decreases when the number of couples increases for both the datasets (except for the case with 16 couples). The χ^2 distance follows the same pattern except for the 4 couple case because of possible distortions of the corrected frames. As expected, the worst results are given by the 1 couple case. As shown in Figure 4, if 1 couple is the input, the face of the models have unnatural colours. The algorithm did not learn how to transfer the face colour as it was not seen in the input images during learning. For the same reason, the outputs of the 2 couple case seem blurry. Even though the lowest figures are given by having 12 couples as input, the results produced with 8 couples are quantitatively and qualitatively satisfactory and we use them throughout the presented results balancing visual quality with computational complexity of SRAT.



(a) Output model (b) Original LR (c) Corrected LR (d) 2x SR

Figure 5: Detail of *SingleM* model rendered with the original (b), the corrected (c) and the super-resolved (d) texture map.

Models	Scale	SingleF		SingleM	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
1a	2x	48.53	0.9910	50.31	0.9935
1b	2x	**48.64**	**0.9912**	**50.45**	**0.9937**
1c	2x	48.60	0.9910	50.41	0.9937
2a	2x	47.68	0.9892	48.91	0.9914
2b	2x	47.70	0.9893	48.94	0.9915
2c	2x	47.77	0.9895	40.01	0.9916
3a	2x	47.74	0.9893	48.92	0.9915
3b	2x	47.85	0.9895	49.22	0.9918
3c	2x	28.68	0.8737	49.33	0.0019

Table 2: PSNR and SSIM results of different training configurations of RCAN. ↑ indicates higher value is better.

4.3. Training models

We analyse the effect of using different trained models for RCAN. We train the network with three datasets:

1. Original texture maps of the LR input model.
2. Original HR images of the model acquired with the DSLR system.
3. Same as 2 but with a removal of the background.

and in three configurations:

(a) RCAN is trained with only our datasets.
(b) RCAN is trained with a dataset made of texture maps of 7 human models [2] cropped into patches and fine-tuned with our datasets.
(c) RCAN is trained with the DIV2K dataset as in the original paper and fine-tuned with our datasets.

The texture maps retrieved in the 3rd stage with the methods proposed by Allene *et al*. [5] are bicubic downscaled (2×) and then super-resolved with the different trained models of RCAN. PSNR and SSIM values are computed between SR and HR texture maps and presented in Table 2. For both *SingleM* and *SingleF*, the 1b model achieves the highest values. Visual results are presented in the supplementary material.

Figure 5 shows a detail of *SingleM* model with the corrected SR texture map. Compared to the original model, the button looks sharper and less blurry, showing the efficiency of SR application to enhance fine details of the appearance.

4.4. Configuration of pipeline

Another evaluation is performed by changing the order of the stages in SRAT. The first stage and the first step of the second stage are not modified. The configurations are:

1. (i) Texture map retrieval; (ii) Texture map colour transfer; (iii) Texture map SR.
2. (i) Texture map retrieval; (ii) Texture map SR; (iii) Texture map colour transfer.
3. (i) Input frames colour transfer; (ii) Corrected frames SR; (iii) Texture map retrieval.
4. (i) Input frames SR; (ii) Input frames colour transfer; (iii) Texture map retrieval.
5. (i) Input frames SR; (ii) Texture map retrieval; (iii) Texture map colour transfer.

We apply the output texture maps to the 3D models and only a qualitative evaluation is performed due to the lack of a ground-truth. Figure 6 shows *SingleF* and *SingleM* models for each configuration. Compared to the original model, the appearance is improved, with brighter colours and more visible details. If the colour transfer is done after the texture map retrieval (configurations 1 and 2), the dress presents artefacts (3rd and 6th rows). If the SR stage is applied before the texture map retrieval (configurations 3, 4 and 5), noise is introduced as seen on the face (2nd and 5th rows).

Performance evaluation: the SR stage requires more time ($\sim 128s/197s$ per texture map/frame) than the colour transfer for LR images ($\sim 8s/14s$) and for SR images ($\sim 35s/56s$). The 1st configuration is the fastest and least computationally expensive: the colour transfer and the SR stages are applied to the texture maps, which are less in number than the frames. The slowest and the most computationally expensive configuration is the 4th one because the two stages are applied to the input frame (16 images of the video cameras for every frame).

4.5. Comparison with related works

We compare the colour transfer algorithm and the SR network used in SRAT with related works.

Colour transfer approaches: the JS and χ^2 values of 5 colour transfer frameworks are shown in Table 3. Finlayson [17] and Vander [29] require pixel correspondences. Pitie [42] introduces a parametric colour transfer function and CycleGan [64] is an unsupervised deep learning technique trained with our frames and HR images. The last method is the framework TPS [19] without any modifications. As shown in Figure 7, Vander, Finlayson and CycleGan cannot transfer the colours from the HR image to

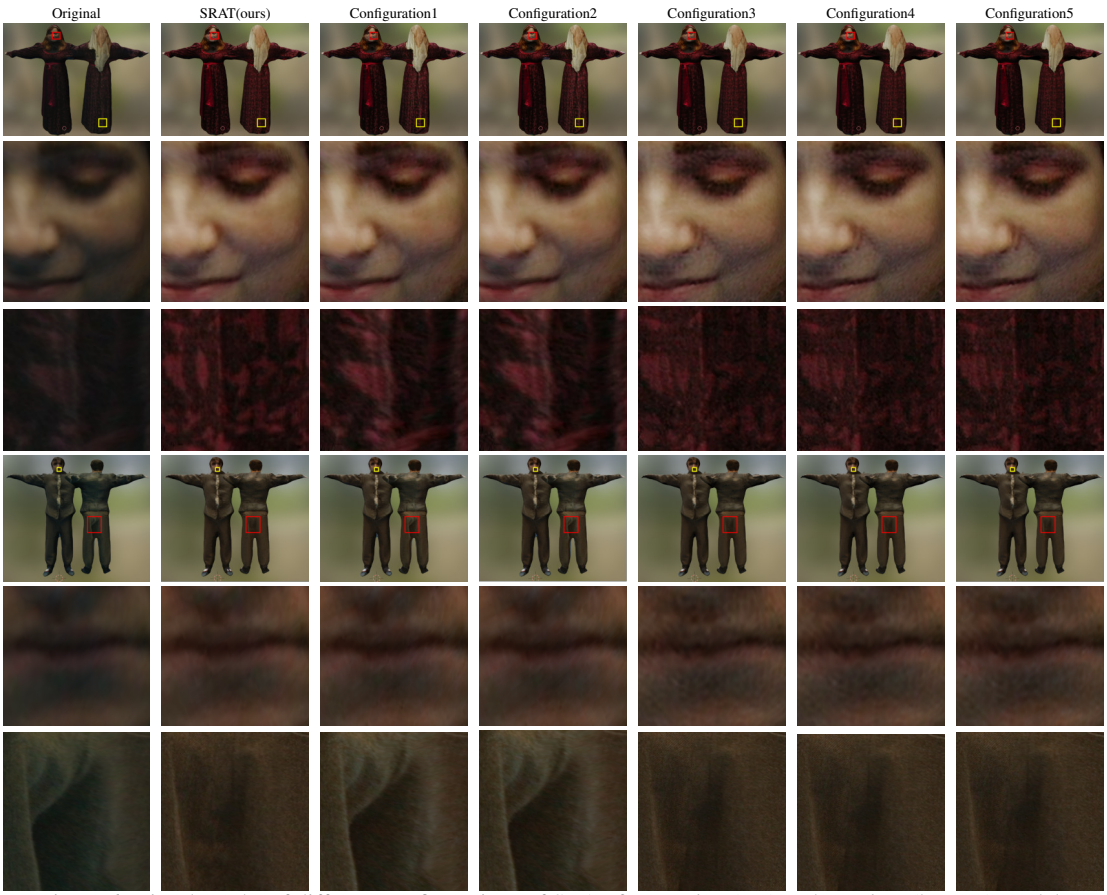


Figure 6: Visual results of different configurations of SRAT for *SingleF* (top) and *SingleM* (bottom) models.

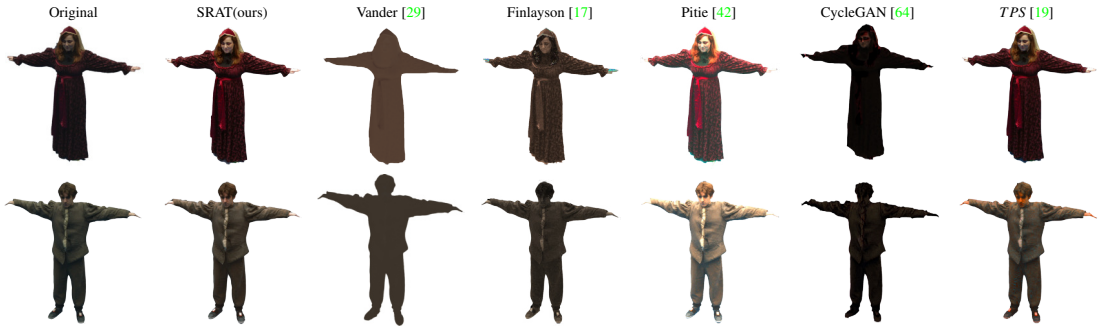


Figure 7: Outputs of different colour transfer approaches. In the top row *SingleF* model, in the bottom *SingleM*.

Methods	**SingleF**		**SingleM**	
	JS ↓	χ^2 ↓	JS ↓	χ^2 ↓
Vander [29]	0.8323	**1.90568**	0.8142	48.6026
Finlayson [17]	0.7949	50.4117	0.6375	121.647
Pitie [42]	0.7886	38.9494	0.7445	88.2087
CycleGan [64]	0.5776	584.419	0.6158	336.151
TPS [19]	**0.5029**	120.177	**0.4479**	191.874
SRAT(ours)	0.5435	13.5905	0.4609	**18.0991**

Table 3: JS and χ^2 values for different colour transfer algorithms.

the frames while Pitie outputs are too bright. If TPS is applied, a different function for every frame of each camera is learned. Therefore, the corrected frames of the same scene and the ones of two consecutive frames have different colours. W.r.t. the quantitative analysis, the JS divergence is the lowest when TPS is applied. TPS learns the colour transfer function by modelling the statistical distributions of the input images. The JS divergence focuses on the difference between statistical distributions and it is lower if these distributions are learned for every frame and not only for 8 selected couples as in our case. The χ^2 of TPS is the second highest. Its lowest figure is obtained when Vander is applied for *SingleF* and ours for *SingleM*. Vander colour transfer produces wrong outputs and the low χ^2 value is influenced by its sensitiveness to image distortion.

SR approaches: the top part of Table 4 shows quantitative comparisons for 2× SR with a classic computer vision approach (Bicubic) and 7 deep learning methods: DBPN [21], SRFBN [36], CSNLN [38], NLR and

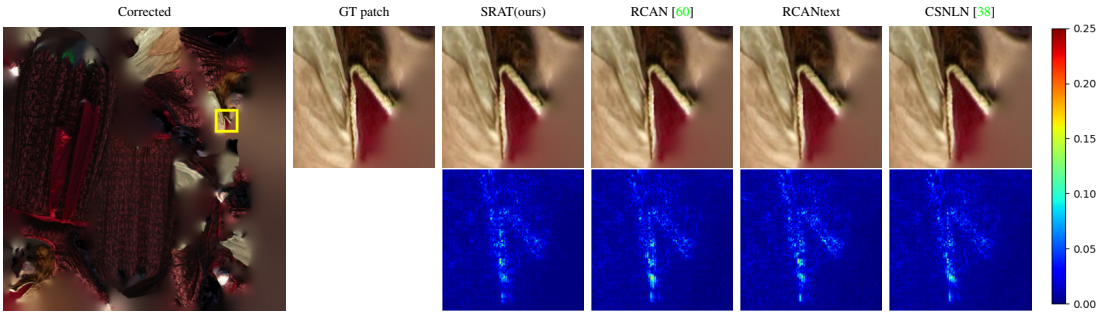


Figure 8: Visual comparison of 2× SR networks. At the bottom, heatmaps of the above *SingleF* texture map portion.

	Methods	SingleF		SingleM	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
***SISR* 2×**	Bicubic	45.96	0.9807	47.39	0.9885
	DBPN [21]	20.91	0.3000	19.73	0.1188
	SRFBN [36]	17.12	0.7583	17.67	0.7381
	CSNLN [38]	48.12	0.9900	49.36	0.9920
	NLR [35]	44.95	0.9856	46.73	0.9888
	NHR [35]	34.81	0.9788	42.48	0.9862
	RCAN [60]	48.14	0.9900	49.35	0.9919
	RCANtext	47.51	0.9886	48.50	0.9904
	SRAT(ours)	**48.64**	**0.9912**	**50.45**	**0.9937**
***RefSR* 4×**	CrossNet [62]	37.16	0.9155	38.96	0.9438
	TTSR [55]	41.29	0.9427	43.00	0.9648
	TTSR-l_2	42.86	0.9622	44.75	0.9759
	SRNTT [61]	42.92	0.9542	43.86	0.9677
	SRNTT-l_2	44.11	0.9626	45.26	0.9760
	SRAT(4×)	**44.42**	**0.9636**	**45.57**	**0.9770**

Table 4: Quantitative results of different SR approaches.

NHR [35], RCAN [60] and RCANtext. For testing with NLR and NHR, the normal maps were retrieved with Blender2.8 [1]. RCAN is the original version and RCANtext is RCAN trained with the human texture map dataset but not fine-tuned with the original texture maps as done in SRAT, which outperforms all the tested methods for both the datasets. The four best network outputs and the heatmaps of the difference with the ground-truth are shown in Figure 8: SRAT heatmap presents more blue pixels confirming its superiority. We then compare SRAT with 3 state-of-the-art RefSR methods (4× super-resolution). Since no training dataset of texture maps with relative references has been available online, we train Cross-net [62], TTSR [55] and SRNTT [61] with CUFED5 [61] dataset as done in their papers. During inference, we select a HR image for each model as reference and we downscale it (6×) for the problem of GPU memory consumption. For the same reason, the LR texture maps are cropped into patches (64x64 size). We train SRNTT and TTSR with only the reconstruction loss (indicated with the suffix l_2) for a fair comparison with SRAT. In the bottom part of Table 4, the PSNR and SSIM figures confirm the superiority of SRAT to RefSR methods.

4.6. More complex scenarios

Interaction scenario: we apply SRAT to an interaction scene where *SingleF* and *SingleM* perform at the same time. A colour transfer function is learned for each model using the same couples of SRAT as input. The colours of the frames are then corrected by applying the two functions to their correspondent models selected with a mask. For each model, its texture maps are retrieved for each frame and super-resolved with the two fine-tuned SR networks.

(a) Interaction: original (b) Interaction: SRAT

(c) Unseen pose: *SingleF* (d) Unseen pose: *SingleM*

Figure 9: Complex scenarios with multiple performers captured simultaneously (a,b) and unseen poses (c,d)

Unseen Poses: the proposed pipeline aims to enhance the appearance of a human perfomance. Therefore, it has to handle multiple poses of the performers. SRAT is effective also when it is applied to frames that were not used to learn either the colour transfer function or the SR model.
Figure 9 shows the results of these complex scenarios.

5. Conclusion

This paper proposes a novel pipeline to enhance the dynamic appearance of low-resolution capture video of human performance using a collection of static high-resolution images of the same subject. The pipeline enables multi-view performance capture systems to increase the capture volume without sacrificing the output reconstruction quality. A novel automatic colour mapping improves the global appearance by correcting the colours of LR capture frames while fine-scale surface details are transferred by an RCAN-style network from the high-resolution images to the super-resolved texture maps. A limitation of the proposed pipeline is that it does not enforce any temporal coherence between the super-resolved texture maps of consecutive frames. This as well as geometric detail transfer between the models of the two systems will be investigated as future works.

References

[1] Blender. https://www.blender.org/. Accessed: 2020-07-26.

[2] Renderpeople. https://renderpeople.com/. Accessed: 2020-07-26.

[3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[4] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Efficient deep neural network for photo-realistic image super-resolution. *arXiv preprint arXiv:1903.02240*, 2019.

[5] Cédric Allène, Jean-Philippe Pons, and Renaud Keriven. Seamless image-based texture atlases using multi-band blending. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[6] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[7] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5420–5430, 2019.

[8] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[10] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018.

[11] Zhiyong Chen, Jing Hu, Xuyang Zhang, and Xiangjun Li. Single image super-resolution based on enhanced deep residual gan. In *MIPPR 2019: Pattern Recognition and Computer Vision*, volume 11430, page 114301W. International Society for Optics and Photonics, 2020.

[12] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.

[13] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[14] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers*, pages 1–10. 2008.

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[16] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[17] Graham D Finlayson, Michal Mackiewicz, and Anya Hurlbert. Color correction using root-polynomial regression. *IEEE Transactions on Image Processing*, 24(5):1460–1470, 2015.

[18] Daniel Freedman and Pavel Kisilev. Object-to-object color transfer: Optimal flows and smsp transformations. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 287–294. IEEE, 2010.

[19] Mairead Grogan and Rozenn Dahyot. L2 divergence for robust colour transfer. *Computer Vision and Image Understanding*, 181:39–49, 2019.

[20] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)*, 38(6):1–19, 2019.

[21] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[22] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018.

[23] Guanqing Hu and James Clark. Instance segmentation based semantic matting for compositing applications. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 135–142. IEEE, 2019.

[24] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–354, 2018.

[25] Youngbae Hwang, Joon-Young Lee, In So Kweon, and Seon Joo Kim. Color transfer using probabilistic moving least squares. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3342–3349, 2014.

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[27] Kideog Jeong and Christopher Jaynes. Object matching in disjoint cameras using a color transfer approach. *Machine Vision and Applications*, 19(5-6):443–455, 2008.

[28] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[29] Allen Klinger. The vandermonde matrix. *The American Mathematical Monthly*, 74(5):571–574, 1967.

[30] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018.

[31] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018.

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[33] Wei-Yu Lee, Po-Yu Chuang, and Yu-Chiang Frank Wang. Perceptual quality preserving image super-resolution via channel attention. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1737–1741. IEEE, 2019.

[34] Kai Li, Shenghao Yang, Runting Dong, Xiaoying Wang, and Jianqiang Huang. Survey of single image super-resolution reconstruction. *IET Image Processing*, 14(11):2273–2290, 2020.

[35] Yawei Li, Vagia Tsiminaki, Radu Timofte, Marc Pollefeys, and Luc Van Gool. 3d appearance super-resolution with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9671–9680, 2019.

[36] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[37] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[38] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[39] Miguel Oliveira, Angel Domingo Sappa, and Vitor Santos. A probabilistic approach for color correction in image mosaicking applications. *IEEE Transactions on image Processing*, 24(2):508–523, 2014.

[40] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754, 2016.

[41] Jaesik Park, Yu-Wing Tai, Sudipta N Sinha, and In So Kweon. Efficient and robust color consistency for community photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 430–438, 2016.

[42] François Pitié and Anil Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. 2007.

[43] François Pitié, Anil C Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1-2):123–137, 2007.

[44] Tania Pouli and Erik Reinhard. Progressive color transfer for images of arbitrary dynamic range. *Computers & Graphics*, 35(1):67–80, 2011.

[45] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[46] Audrey Richard, Ian Cherabier, Martin R Oswald, Vagia Tsiminaki, Marc Pollefeys, and Konrad Schindler. Learned multi-view texture super-resolution. In *2019 International Conference on 3D Vision (3DV)*, pages 533–543. IEEE, 2019.

[47] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE computer graphics and applications*, 27(3):21–31, 2007.

[48] Subeesh Vasu, Nimisha Thekke Madam, and A. N. Rajagopalan. Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[49] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–11. 2009.

[50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[51] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018.

[52] Yao Xiang, Beiji Zou, and Hong Li. Selective color transfer with multi-source images. *Pattern Recognition Letters*, 30(7):682–689, 2009.

[53] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[54] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014.

[55] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[56] Wei Yang, Luhui Xu, Xiaopan Chen, Fengbin Zheng, and Yang Liu. Chi-squared distance metric learning for histogram data. *Mathematical Problems in Engineering*, 2015, 2015.

[57] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, volume 1, pages 147–149, 1992.

[58] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[59] Qianwen Zhang and Roxanne L Canosa. A comparison of histogram distance metrics for content-based image retrieval. In *Imaging and Multimedia Analytics in a Web and Mobile World 2014*, volume 9027, page 90270O. International Society for Optics and Photonics, 2014.

[60] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[61] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[62] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–104, 2018.

[63] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.