# Super-Resolution Appearance Transfer for 4D Human Performance Capture

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

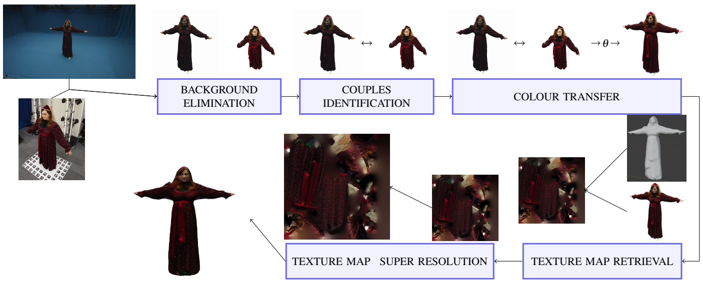


Figure 1: Super-resolution and colour transfer pipeline: The input is a low-resolution (LR) video and high-resolution (HR) static images of a person; LR-HR couples are identified with similar views for colour transfer to match the HR colour distribution; super-resolution is then applied to the LR texture map to obtain a HR texture for the 4D model.

## 1 Introduction

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. Unless the acquisition system has a large number of high-resolution (HR) video cameras ($> 100$), the quality of the appearance is much lower compared to a static acquisition of the same person, which requires digital stills cameras ($> 8$k) to capture the person in a small volume ($< 8m^3$). The 3D model reconstructed with these static HR images captures fine details in both shape and appearance. In this paper we propose a practical solution that enhances the appearance of a low-resolution (LR) dynamic video performance capture acquired with a sparse set of cameras through super-resolution (SR) appearance transfer from the same human subject acquired with DSLR cameras used for static 3D reconstruction.

## 2 Method

We propose FSTD, 'From Static to Dynamic', a pipeline (shown in Figure 1) that performs local super-resolution and global colour correction of a 4D dynamic performance reconstruction from multi-view video using HR static capture from multiple DSLR cameras. The input of our pipeline is a 4D video sequence of a person as well as a set of $K$ HR images $\{I^q_{HR}\}^K_{q=1}$ of the same person acquired with static cameras. The 4D video sequence consists of $M$ LR frames $\{I^i_f\}^M_{i=1}$, reconstructed meshes and texture maps. No geometric information of the HR capture is used.

**LR-HR image couple identification:** After foreground-background separation, pairs of reference images ($I_{HR}$) and target frames ($I_f$) with similar content are identified for the colour transfer. To identify the couples, we propose a new automatic method that computes the similarity between them in the texture map domain instead of using the image domain due to the different acquisition systems. We first apply Densepose [1] to reconstruct partial texture maps of the frames ($T_p(I_f)$) and of the HR images ($T_p(I_{HR})$). These partial texture maps are invariant to the camera orientation and position. Similarity between partial texture maps $T_p(I_f)$ and $T_p(I_{HR})$ is evaluated using the SSIM metric. A LR frame $I_f$ is coupled with the HR image $I_{HR}$ whose partial texture map is the most similar defined by the SSIM metric:

$$I^i_{f_j} \leftrightarrow I^q_{HR} \; where \; \underset{I^q_{HR}}{\operatorname{argmax}}\{SSIM(T_p(I^i_{f_j}), T_p(I^q_{HR}))\} \tag{1}$$

where $I^i_{f_j}$ is the i-th frame of the video camera $j$ and $I^q_{HR}$ is the HR image of the camera $q$.

**Colour transfer:** Once the couples are identified, we select those whose $I_f$ corresponds to the first frame of LR cameras where the models are in a T-pose. The image colour transfer approach [2] is extended to multi-view images as input to learn a colour transfer function $\phi(x)$ from HR to LR images. The function $\phi(x)$ is modelled as a Thin Plate Spline that depends on a set of parameters $\theta$. This set is computed by minimizing the

(a) Original model (b) FSTD output (c) HR image

Figure 2: a) Input frame image; b) FSTD output; c) HR image.

following energy function with a gradient descent algorithm:

$$\theta = \frac{1}{N}\sum_{l=1}^{N} \underset{\theta_l}{\operatorname{argmin}}\{||p_f||^2 - 2 < p_f|p_I >\} \tag{2}$$

where $N$ is the number of input couples, $p_f$ is the Gaussian distribution of the colours of $I_f$ with parameterised mean $\phi_\theta(\mu_f)$, $p_I$ is the Gaussian distribution of the colours of $I_{HR}$ and $< p_f|p_I >$ is their scalar product [2]. The colours of all the input LR frames are corrected with the computed function. The dynamic perfomance is then reconstructed with the method proposed by Starck and Hilton [3] and the LR texture maps of the model are retrieved for each frame by projecting the colour corrected images to the shapes.

**Texture map super-resolution** To further improve the appearance of the model, fine details are enhanced by super-resolving the retrieved texture maps (treated as 2D RGB images) with the RCAN-style network presented by Dai et al. [4]. In [4], RCAN was trained augmenting 800 general images. Since we aim to super-resolve texture maps of a specific model, we first train RCAN with a set of patches of human models texture maps. We then fine-tune it with a dataset of texture maps of the input model retrieved in a pre-processing stage. For every input model of FSTD, RCAN is fine-tuned with the texture maps of that model.

## 3 Results

(a) Final model (b) LR detail (c) 2x SR detail

Figure 3: Effect of SR stage in the reconstructed 3D models.

| Methods | SingleF | | SingleM | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| RCAN [4] | 48.14 | 0.9900 | 49.35 | 0.9919 |
| FSTD | **48.64** | **0.9912** | **50.45** | **0.9937** |

Table 1: Quantitative results of RCAN and FSTD.

The qualitative results obtained applying FSTD show improvements in the quality of the appearance of dynamic performances. In particular, Figure 2 shows that the colour transfer stage enhances the global appearance of the reconstructed model by colour correction to match the HR image colour distribution. Figure 3 illustrates how the SR enhances the fine detail appearance of the low-resolution 4D model. Table 1 shows that our training approach achieves higher values of PSNR and SSIM for two performers compared to the original RCAN.

[1] 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.

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

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

[4] 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.