A report on Multi-frame Motion Coupling for Video Super Resolution

Chirag Bhuvaneshwara, *MSc Student, Universität des Saarlandes*

***Abstract*—In today's world of high-resolution displays, video super-resolution algorithms are necessary to obtain sharp image sequences that fully utilize these displays. While many variational video super-resolution algorithms have been proposed for this problem, they all had a general structure which required too many optic flow computations and they might also introduce flickering in the high-resolution solution. Here, we discuss a new variational model which requires significantly fewer optic flows, solves the flickering problem and produces very high-quality results which are better than the existing variational models. But more importantly, the results show that the new model discussed is better in most cases compared to even the state-of-the-art deep learning models. This is obtained by introducing a convex energy functional which couples the unknown high-resolution frames directly and involves infimal convolution regularization to compensate for the absence of accurate optic flows using a new spatiotemporal scaling term. The results from this new multi-frame motion coupling model are thoroughly evaluated against other models that are at the forefront of video super-resolution.**

***Index Terms*—Video Super-Resolution, Variational Methods, Energy Minimization, High Resolution**

I. Introduction

Moving from a low-resolution(LR) image space to its corresponding high-resolution(HR) space introduces new pixel coordinates in the HR space which also need to be filled in. In this case, spatial information from the input LR image can be used. This method is known as interpolation. There are many different approaches in interpolation as seen in [2] and two of the simpler approaches are the nearest neighbour and bicubic interpolation. Bicubic interpolation is an extension of the nearest neighbour approach. It fills in the brightness value at an unknown pixel coordinate in the HR space as a Gaussian weighting of the nearest 4*4 neighbours from the LR image space.

But when the input is a sequence of images, there exists a relationship between the different images in the sequence. Optic flow(OF) gives the sequential relationship between the images and highly accurate estimations can be obtained using [6]. Applying simple interpolation per frame for an LR video is not exploiting the OF information. As a result, interpolation on a sequence of images produces a video which consists of frames, each of which is of the desired HR but is not as sharp as it could have been if the sequential relationship, obtained from the OF field, were to be exploited. Making use of this additional information in sequential images is exactly what super-resolution(SR) algorithms do. As a result, they produce HR image sequences containing high level of details i.e. SR takes the input sequence from LR to HR by filling in the brightness values for the new coordinates in HR space and also adds high-frequency image details obtained by using the OF fields in the different frames as stated in [3].

SR algorithms are implemented using deep learning or variational energy minimization schemes, both of which require both the LR image sequence and the OFs as input. This report focuses on the latter and evaluates the performance of a newly proposed variational scheme against the different state-of-the-art SR models.

Reference [1] states that the classical variational SR models as in [5], have a general structure which estimates the motion from the current frame to the neighbouring frames and models the data formation process via warping, blur and downsampling, and uses a suitable regularization to avoid artefacts in the computed HR image frame. This classical model explained in Section II, requires quadratic number of OF estimations with respect to the number of LR frames and as each HR frame is computed separately, the model does not restrict flickering in the HR solution being computed.

These drawbacks are solved in [1] by introducing the Multi-frame Motion Coupling(MMC) variational model which is explained in Section III. MMC solves for the entire HR image sequence by directly coupling each pair of the unknown neighbouring HR frames, using infimal convolution regularization to choose between when to smooth in the spatial direction and when to smooth in the spatiotemporal direction, and introduces a novel way of estimating the spatiotemporal scaling.

This report discusses each of the key aspects of the MMC model. The required OF field for the model is computed from the model explained in Section IV. Some practical optimizations are presented in Section V. Finally, the performance of the MMC variational SR model is evaluated against the state-of-the-art deep learning and variational SR models in Section VI.

II. Classical Variational SR Models

All previously implemented variational SR models have a common general structure where each HR frame is computed by imposing temporal similarity between the downcasted HR frame and each of the frames in the LR space. This particular type of coupling is shown in Figure 1.

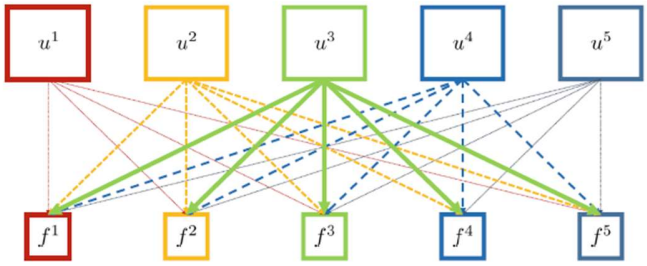


Figure 1:shows classical coupling as seen in [1]

To impose such a temporal similarity, a quadratic number of OF estimations are required w.r.t the number of frames in the image sequence.

$$u^{i^*} = argmin_{u^i}\left(\left\| D(b * u^i) - f^i \right\|_{H^{\in d}} + \lambda \left\| \nabla u^i \right\|_{H^{\in r}} + \sum_{j \neq i} \left\| D(b * W^{j,i} u^i) - f^i \right\|_{H^{\in d}} \right) \quad (1)$$

Equation (1) gives the classical energy model, w.r.t the Hueber Norm H, to be minimized to obtain one HR frame. Here, D is the down-sampling operator that takes HR frames to LR space and b is a Gaussian blurring operator to remove high frequency components in the HR frame to avoid artifacts in the LR space which is motivated from the sampling theorem explained in [7] by Weickert. So, the 3 terms in (1) are the spatial consistency data term, spatial consistency smoothness term and the temporal consistency data term which is motivated by the coupling in Figure 1. This clearly shows that the model does not explicitly enforce temporal smoothness which might lead to flickering in the HR image sequence solution.

III. MMC Model

The Multi-frame Motion Coupling(MMC) variational SR model introduced in [1] by Geiping et al computes each HR frame by considering the OF directly between two unknown neighbouring HR frames as indicated in Figure 2 which makes the number of OF computations linear in the number of frames. This coupling is based on the brightness constancy assumption.

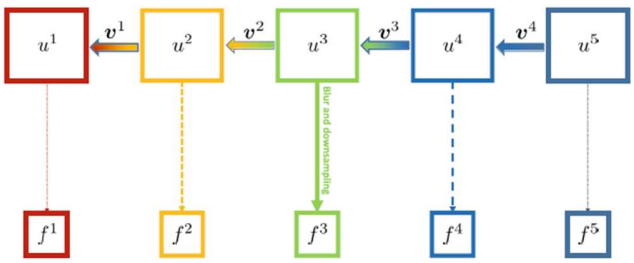
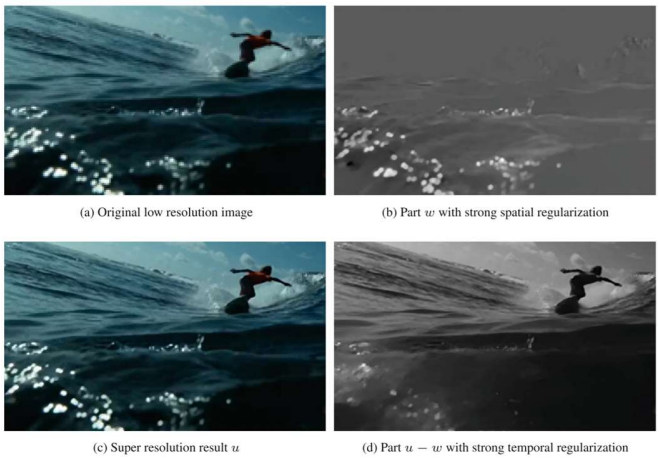


Figure 2: Shows MMC coupling as seen in [1]

Equation (2) gives the energy to be minimized to obtain the entire HR image sequence solution. The first term in (2) is a spatial consistency date term which is also indicated in Figure 2. The second term in the energy is an application of infimal convolution regularization from [4] by Holler et al. It depends on splitting u into 2 parts which is explained further in sub-section II A.

$$u^* = argmin_u \left(\sum_{i=1}^{n} \left\| D(b * u^i) - f^i \right\|_1 + \alpha \, inf_{u=w+z}\{R_{temp}(w) + R_{spat}(z)\} \right) \quad (2)$$

A. Spatiotemporal infimal convolution



(a) Original low resolution image (b) Part w with strong spatial regularization

(c) Super resolution result u (d) Part $u - w$ with strong temporal regularization

Figure 3: Illustrates the behaviour of infimal convolution regularization as seen in [1]

Geiping et al state in [1] that infimal convolution can be regarded as a convex approximation to a logical OR connection. Equation (3) presents infimal convolution regularization as the infimum obtained from splitting u into w and z.

$$(R_{temp} * R_{spat})(u) := inf_{u=w+z}\{R_{temp}(w) + R_{spat}(z)\} \quad (3)$$

So, the goal is to find a split of u in a way that w captures the part of u for which OF fields can be calculated accurately and z is calculated trivially as $(u - w)$ which allows it to capture the information of high frequency objects in u for which the OF cannot be calculated accurately. This results in w containing spatiotemporal information and z containing spatial information from each frame of the objects that moved too fast or were hidden behind occlusions. Figure 3 illustrates the information in w and z for one particular LR frame f.

Equations (4) and (5) show that both $R_{temp}(w)$ and $R_{spat}(z)$ have the same structure where the summation is for all frames n, first 2 terms are spatial derivatives indicated by the subscripts and the third term is the warping operator which computes the spatiotemporal derivative.

$$R_{temp}(w) = \sum_{i=1}^{n} \left\| \sqrt{(\kappa w_x^i)^2 + (\kappa w_y^i)^2 + W((w^i, w^{i+1}))^2} \right\|_1 \quad (4)$$

$$R_{spat}(z) = \sum_{i=1}^{n} \left\| \sqrt{(z_x^i)^2 + (z_y^i)^2 + \kappa W((z^i, z^{i+1}))^2} \right\|_1 \quad (5)$$

The scaling factor κ selected to enforce that $R_{temp}(w)$ contains only the trustworthy spatiotemporal component in (4) and to enforce that (5) contains only the spatial components.

Substituting (4) and (5) in (3), it is easy to see that infimal convolution regularization smoothens in the spatiotemporal direction when OF can be computed accurately or smoothens in the spatial direction when objects don't meet the OF assumptions.

B. Warping Operator

The warping operator used in (4) and (5) computes the spatiotemporal derivative by considering the optic flow field $\boldsymbol{v^i}$ between the i^{th} and the $(i+1)^{th}$ frames. Equation 6 shows how this warping operator is computed. Here, h is the spatiotemporal scaling and it is often set to 1.

$$W((u^i, u^{i+1})) \;=\; \frac{u^i(x) \;-\; u^{i+1}(x \;+\; \boldsymbol{v^i}(x))}{h} \tag{6}$$

C. Spatiotemporal scaling

Reference [1] introduces a novel way of automatically estimating the spatiotemporal scaling h as it would be incorrect to assume that objects moving from frame to frame would move by just 1 unit along the OF $\boldsymbol{v}$.

To compute this new scaling estimate, bicubic interpolation must be applied to each LR frame f^i to obtain a sequence of HR frames u_0. Then, the new spatiotemporal scaling is given by (7). Here, W computes the spatiotemporal derivative for all the frames in u_0 by using the warping operator given by (6). The denominator sums up the L1 norms of the spatial derivatives of u_0.

$$h = \frac{\|Wu_0\|_1}{\left\|\frac{d}{dx}u_0\right\|_1 + \left\|\frac{d}{dy}u_0\right\|_1} \tag{7}$$

In [1], Geiping et al have stated that since the warp operator is multiplied with h^{-1} it provides an image adaptive way to make sure that the spatial and temporal regularity terms are in the same order of magnitude. This term also makes sense in terms of the units as h is supposed to be in terms of time units which is exactly what the unit of (7) is.

D. Compact Notation

By putting all the flattened image frames of an image sequence in a single vector, (2) can be re-written as in (8). In the energy, the first term corresponds to the spatial consistency data term, the second term corresponds to $R_{temp}(w)$ which performs smoothing in the spatiotemporal direction when the OF field $\boldsymbol{v}$ is accurate and the third term corresponds to $R_{spat}(z)$ which performs smoothing in the spatial direction when the OF assumptions are not met leading to inaccurate estimates of $\boldsymbol{v}$.

The matrix A in (8) is a block diagonal matrix containing a combination of the down-sampling and blurring operators of each frame along the diagonal. The vector u contains all the flattened HR frames $u^1, u^2, \dots, u^n$ and similarly, the vector f contains all the flattened LR frames. W is a matrix that applies the warping operator from (6) to all the frames in w and z. The notation $\|.\|_{2,1}$ stands for applying L2 norm first and then applying L1 norm.

$$(u^*, w^*) = argmin_{(u,w)} \begin{pmatrix} \|Au - f\|_1 \\ +\alpha \left\| \begin{pmatrix} \nabla w \\ \kappa W \mathrm{w} \end{pmatrix} \right\|_{2,1} \\ +\alpha \left\| \begin{pmatrix} \kappa\nabla(u-w) \\ W(u-w) \end{pmatrix} \right\|_{2,1} \end{pmatrix} \tag{8}$$

IV. Optic Flow Computation

The MMC variational SR model requires OF $\boldsymbol{v}$ as an input which also must be estimated for a given image sequence f. In [1], Geiping et al used a variational energy minimization scheme given by (9) to compute the OF field between each pair of frames in an image sequence. This scheme contains two data terms as it is based on the 2 assumptions that the gradient and the brightness do not change from frame to frame in an image sequence. It also contains one smoothness term to enforce regularity of the OF field.

The summation term in (9) is over all the frames in the image sequence which results in the optimal solution containing all the flow fields in the image sequence. The first term models the gradient consistency data term, the second term models the brightness consistency data term and the last term models the smoothness of the flow field in both the spatial directions.

$$\begin{aligned} \boldsymbol{v}^* = argmin_{\boldsymbol{v}} \Bigg(& \sum_{i=1}^{n-1} \int_{\Omega} \left\| \nabla f^i(x) - \nabla f^{i+1}\left(x + \boldsymbol{v}^i(x)\right) \right\|_1 dx \\ & + \int_{\Omega} \left| f^i(x) - f^{i+1}(x + \boldsymbol{v}^i(x)) \right| dx \\ & + \beta \sum_{j=1}^{2} \left\| \nabla \boldsymbol{v}_j^i \right\|_{H^{\epsilon}} \Bigg) \end{aligned} \tag{9}$$

In [1], Geiping et al have first linearized the brightness and gradient constancy terms using first-order Taylor expansions with respect to the current estimate $\tilde{\boldsymbol{v}}^i$ of the OF field which results in a convex energy minimization problem for each linearization. It has been stated that they use the well-known iterative coarse-to-fine approach to obtain accurate OF $\boldsymbol{v}$.

V. Optimization

To obtain the MMC SR solution for an LR image sequence f, the optimization procedure would be two-fold as the OF $\boldsymbol{v}$ has to be computed first by (9) and then the SR solution u can be computed using (8). Reference [1] states that many different variations of this two-fold procedure were applied. They tried an alternating scheme which first finds OF $\boldsymbol{v}$ on LR frames f, then solves for MMC SR solution u and then again computes the OF on the SR solution to see if the more accurate OF betters the performance of MMC model. But it was found in [1] that the effective resolution increase from such a compute intensive alternating scheme is marginal. Another scheme that was considered in [1] was up sampling the LR frames f to the

desired resolution by bicubic interpolation, computing the OF $\boldsymbol{v}$ on this bicubic estimate, then performing MMC SR.

However, it was found in [1] that computing OF $\boldsymbol{v}$ on LR frames f, up sampling the obtained $\boldsymbol{v}$ to the desired resolution by bicubic interpolation and then solving the MMC super-resolution problem is sufficient. This scheme used by Geiping et al is significantly less compute intensive than the previously described scheme which computed OF fields in the HR space. In spite of significantly less computations, the simple two-fold scheme employed in [1] provides results which are as precise as the other schemes considered.

VI. Numerical Results

Reference [1] finds that a good and robust trade-off for arbitrary video sequences for a magnification factor of 4 is obtained when the regularization weights are defined as static parameters. The best regularization weight values were $\beta = 0.2$ for OF computation, $\kappa = 0.25$ for infimal convolution and $\alpha = 0.01$. for the MMC super-resolution model.

In order to super-resolve colour videos, it is possible to apply the MMC variational method on each of the colour channels. But a simpler approach is to use the well-known advantage provided by the YC_bC_r colour space. Here, the luminance channel Y contains almost all the detail information. Geiping et al found in [1] that super resolving just the Y channel using MMC and performing bicubic interpolation on the chroma channels Cb and Cr results in almost exactly the same peak signal-to-noise ratio(PSNR) as super resolving each channel separately.

To super-resolve long image sequences, Geiping et al found in [1] that it is suitable to divide the image sequence into frame batches of a fixed length and super resolve each frame batch separately. Temporal consistency can be provided between the different frame batches by using the last super-resolved frame from the previous batch as a boundary value for the current batch.

Reference [1] performs a thorough analysis of the MMC variational model against other super-resolution methods. In this evaluation, the baseline is the Nearest Neighbour (NN) Interpolation method. In addition to NN, the evaluation considers the Bicubic (BIC) interpolation method, 3 variational models using the coupling described in Section II and 3 deep learning-based models.

The different algorithms are tested on several scenes with very different complexity and resolution. The resulting super-resolution result for one simple synthetic scene consisting of a planar motion of the London subway map (tube) is shown in Figure 4. This particular sequence has 13 frames and the super-resolution performed is by a factor of 4. In [1], Geiping et al state that due to their idea of jointly super resolving multiple frames, the result in Figure 4 for the MMC model is superior to the competing variational approaches. This particular result for the London tube scene also outperforms the deep learning methods of Deep Draft and VSRnet. Reference [1] states that the higher performance of the MMC model compared to the Deep Learning(DL) based models is due to the DL models not having seen similar data during the training phase.

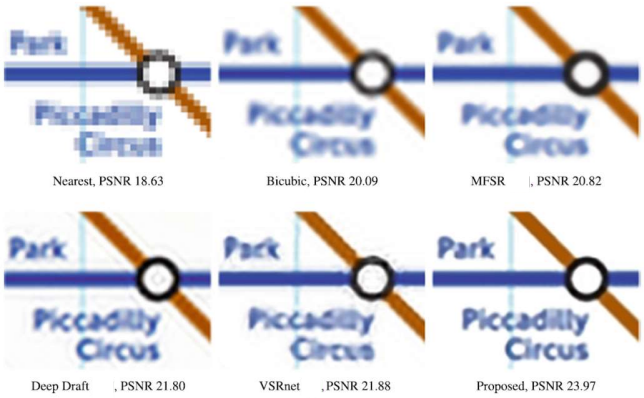


Figure 4: Results for super-resolving a set of 13 images of a London tube map by a factor of 4 as seen in [1].

A. Evaluation of changes introduced by the MMC model

Reference [1] which presents the MMC model introduces several new concepts which enable variational SR results to be comparable with state-of-the-art DL models. To quantify how each of these new changes improves the performance, Geiping et al introduce the Additive Regularizer model in (10). This model has just the new coupling and does not include infimal convolution. Experiments were conducted to compare the average PSNR from the SR process for Bicubic interpolation, some classical variational models and the models capturing each of the new changes introduced in the MMC model. Figure 5 shows that the additive regularizer model with $h = 1$ is already better than the Classical model presented in Section II. Adding the new spatiotemporal scaling to the additive regularizer model also improves the performance. The next highest PSNR result is obtained for the infimal convolution model from (8). But the highest PSNR value is obtained for the full MMC model with infimal convolution and the new spatiotemporal scaling which is given by (7).

$$u^* = argmin_u(\|Au - f\|_1 + \alpha\|(W\mathrm{u})\|_1 + \alpha\|\nabla u\|_1) \qquad (10)$$

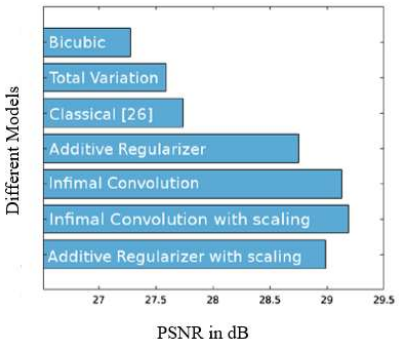


Figure 5: Shows the effect of each of the proposed changes. Obtained from [1].

B. Comparison with other methods

Reference [1] conducts an in-depth analysis of the state-of-the-art SR algorithms on 12 datasets of varying complexity. Some of these datasets chosen are simple while other datasets are more realistic in terms of the brightness changes and the number of occlusions over the sequence. Geiping et al compare

the performance of the MMC model against 9 other SR models. In such a scenario, it is found that the MMC SR model better than all other SR algorithms on 7 out of the 12 datasets with an average PSNR of 29.19 dB over all the datasets. The remaining 5 cases, the DL based VDSR model performs better with an average PSNR of 29.13 dB over all the datasets. These 5 cases are datasets that have lower frame rates or contain a large number of occlusions leading to the deterioration in the performance of the MMC model. But it is to be noted that in these 5 cases, the MMC model is still the second best model and comes very close to the PSNR values of the VDSR model. Figure 6 shows the PSNR performance of multiple SR schemes on the walk dataset which also consists of a significant number of occlusions. Here, the MMC model is still able to produce the best result.

Figure 6: Shows the middle super-resolved frame from a sequence for the walk dataset as seen in [1].

C. Numerical Analysis

In [1], Geiping et al performed another experiment to see how the MMC super-resolution model performs when the Ground Truth (GT) OF is used and how it performs when the OF is computed from (9). This analysis is performed on the Sintel MPI dataset which is also a synthetic dataset which is why the GT flow can be computed. This dataset contains different image sequences with increasing level of realism which means that the frame rate and the number of occlusions are varied. Due to this, the OF from (9) cannot be computed accurately as its assumptions are not met.

Table 1 summarizes the results from the experiment on the Sintel MPI dataset. Albedo, Clean and Final are arranged in increasing level of realism. It is clear from the table that the MMC model performs better with the GT flow for Albedo and Clean sequences. But when it comes to the Final sequence, there is a surprising result as the MMC model performs better with the OF from (9) rather than the GT flow. Geiping et al explain this behaviour as the MMC model penalizing the GT flow more severely as it does not satisfy brightness assumptions. This same brightness assumption is made by both the MMC model and the OF model. Due to this reason, the MMC model prefers to use the OF from (9).

Table 1: Presents Avg. PSNR for MMC SR model on the Sintel MPI dataset when using GT flow v/s OF from (9) as seen in [1]

Rendering	GT flow	Our OF
Albedo	32.53	31.91
Clean	27.88	27.68
Final	33.31	34.65

VII. Conclusions

The MMC variational super-resolution technique introduces a new method of directly coupling the unknown HR frames. This enforces temporal consistency of the super-resolved video directly avoiding any flickering while reducing the number of OF computations to a linear scale in the number of frames. The addition of infimal convolution regularization enables the MMC model to utilize the OF information whenever it is correct while still providing some smoothing using spatial components when the OF is not correct. From the provided extensive numerical analysis of several state-of-the-art SR models, it is found that the MMC model is better than competing variational models. More importantly, Geiping et al show that variational models can be developed to be competitive even with the top-of-the-line DL models. It is shown that the MMC model performs better than even the DL models when there is a high frame rate and the image sequence contains few occlusions. Even when these assumptions are not met, the results of the MMC model is still significantly close to the DL models.

References

[1] Geiping, Jonas & Dirks, Hendrik & Cremers, Daniel & Moeller, Michael. (2018). Multiframe Motion Coupling for Video Super Resolution. 10.1007/978-3-319-78199-0_9., associated website: http://www.vsa.informatik.uni-siegen.de/en/superResolution

[2] Parsania, Pankaj & Paresh V. Virparia, Dr. (2016). A Comparative Analysis of Image Interpolation Algorithms. IJARCCE. 5. 29-34. 10.17148/IJARCCE.2016.5107.

[3] Nasrollahi, Kamal and Thomas B. Moeslund. "Super-resolution: a comprehensive survey." *Machine Vision and Applications* 25 (2014): 1423-1468.

[4] Holler, M., Kunisch, K.: On infimal convolution of TV-type functionals and applications to video and image reconstruction. SIAM J. Imaging Sci. 7(4), 2258–2300 (2014)

[5] Unger, M., Pock, T., Werlberger, M., Bischof, H.: A convex approach for variational super-resolution. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) DAGM 2010. LNCS, vol. 6376, pp. 313–322. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15986-2 32

[6] Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004). https://doi.org/10.1007/ 978-3-540-24673-2 3

[7] J. Weickert, Image Processing & Computer Vision lecture series, Image Transforms II: Sampling Theorem and Discrete Fourier Transform, 2018