Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Efficient Task Aware Super-Resolution and Colorization

For Image and Video Domain

Semester Project

Simon Schaefer

Advisor: Dr. Radu Timofte, Shuhang Gu
Supervisor: Prof. Dr. Luc van Gool
Computer Vision Laboratory
Department of Information Technology and Electrical Engineering

June 1, 2019

Abstract

The abstract gives a concise overview of the work you have done. The reader shall be able to decide whether the work which has been done is interesting for him by reading the abstract. Provide a brief account on the following questions:

- What is the problem you worked on? (Introduction)
- How did you tackle the problem? (Materials and Methods)
- What were your results and findings? (Results)
- Why are your findings significant? (Conclusion)

The abstract should approximately cover half of a page, and does generally not contain citations.

Acknowledgements

I would like to thank a number of people who have encouraged and helped us in writing this Semester Project. I am proud of the achieved results and I appreciate the support I received from all sides.

I am much obliged to Prof. Dr. Luc Van Gool for his support, for the confidence in my project and for providing the opportunity to execute it.

Special thanks to my advisors Radu Timofte and Shuhang Gu. Their broad knowledge and experiences in machine learning had a wide influence on the results and also on my education.

Most importantly, I am very grateful to all my friends and family members who supported us during the whole project and helped me to achieve my aims.

Zürich in June 2019,
Simon Schaefer

Contents

List of Figures

List of Tables

Abbreviations

COL colored image

GRY grayscale image

HR high-resolution image

IC Image-Colorization

LR low-resolution image

SISR Single-Image-Super-Resolution

SLR task aware-low-resolution image

SR Super-Resolution

TAD Task-Aware-Downscaling

VSR Video Super-Resolution

1 Introduction

With the rise of deep learning in image processing Super-Resolution (SR) and Image-Colorization (IC) in both the image and the video domain have received significant attention [21]. While SR aims to reconstruct a high-resolution image (HR) from a low-resolution image (LR), image colorization deals with the transformation from an uncolored, grayscale image (GRY) to a RGB colored image (COL). However, in most of the recent works (e.g. [20], [19], [8], [18]) the problem of downscaling and upscaling or decolorization and colorization are regarded as seperate problems although upscaling often is preceded by downscaling, leading to a loss of information from the downscaling process which makes the inverse problem of SR highly ill-posed [10]. Despite of the large progress in SR in the last years ([21]) very specific details therefore often cannot be reconstructed, when interpolation is used for downsampling. However, as shown in Fig. 1 the downsampling method has a large impact on the performance of the subsequent upscaling task.

Figure 1: Comparison between an upscaled image based on bicubic downsampled (left) and task aware downsampled (right) LR image applied on the same model with upscaling factor 4.

As can be seen above a task aware approach can dramatically improve the performance of existing super-resolution models in terms of reconstruction quality while keeping the compression rate constant. However, the research on task aware downscaling methods is a very new field and therefore there still are a lot of unresolved issues such as the effect of perturbation or the feasibility of applying it to tasks other than SISR and IC.

1.1 Focus of this Work

For this reason this work focuses on TAD for several standard computer vision problems such as super-resolution or colorization in both the image and video domain, as recently purposed by Heewon Kim et. alt. ([10]) for the image domain only. However, as shown in Fig. 2 the TAD implementation purposed in [10] suffers from vulnerability against perturbation of the downscaled image. Although the purposed model is quite shallow having

10 convolutions for each scaling process only, there still is potential for improvement, which especially gains importance when TAD is applied to the video domain (for real-time capabilites).

Figure 2: Problems of task aware downscaling as purposed by [10]: Perturbation (left), Runtime (right)

Therefore the goals of this work are the following:

- reimplement and evaluate the TAD framework purposed in ([10])
- improve the TAD framework especially with regards on the trade-off between model-complexity (runtime) and restoring quality (PSNR) as well as with regards on robustness against perturbations
- extend the TAD framework to the video domain

By that to the best of our knowledge this work is the first one using deep learning for downscaling in the video domain.

1.2 Thesis Organization

After the problem statement Chapter 1 related works are introduced for both the image and video domain Chapter 2. Chapter 3 explains the methods that are used in order to achieve the goals described above and which are evaluated in Chapter 4. A final discussion of the results as well as an outlook on further work can be found in Chapter 5. Further visualization and experiments are shown in the abstract.

2 Related Work

In the following previous work in super-resolution, colorization and task aware-downscaling are presented. At the end of each section the models used for comparison and evaluation of the the underlying approach are further explained in detail. Thereby the models were selected based on several criterias performance compared to the state-of-the-art, the use as benchmark in related papers and availability of (pretrained) models.

2.1 Super-Resolution in Image Domain

The problem of SR in the image domain is called SISR and is shown in Fig. 3. A lot of approaches have been tried in order to cope with the SISR problem. While early approaches such as bicubic and Lanczos [5] tackle the problem using simple deterministic filters which are computational cheap but produce blurry results and lack in high frequency details, more recent approaches approach the problem using example-based methods such as sparse encoding or deep learning methods.

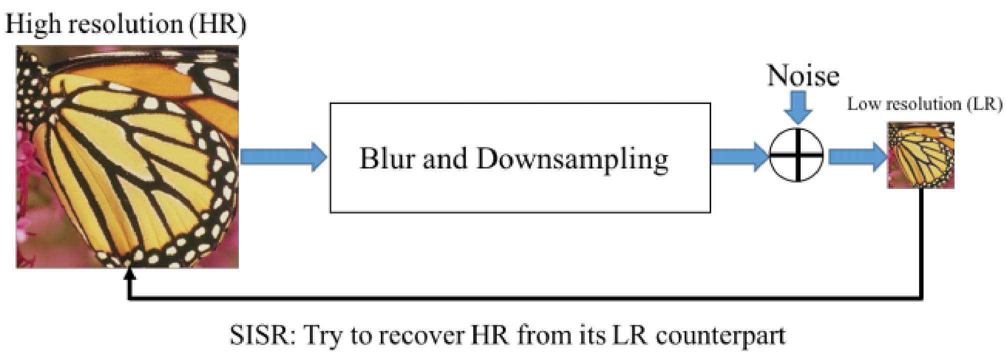

Figure 3: General SISR problem according to [23].

Sparsity-based techniques assumes the LR image to be transformable in another domain (usually a dictionary of image atoms [6]) and tries to find correspondences between the LR and HR patches in the transformed space, as implemented in [4]. However, these techniques usually are very computationally expensive. Among other learning based approaches such as the use of random forests [15], in-place example regression models [22] or adjusted anchored neighborhood regression [17], in terms of accuracy applying CNN based approaches have shown the largest success. [1] Dong et al. [2] trained a shallow CNN end-to-end to build the HR image based on a bicubicly upscaled LR image. This approach was improved by Kim et al. [11] (VDSR) using a deeper network (20 layers) and cascading small filters many times in a deep network structure to exploit contextual information over large image regions in an efficient way. By advancing the network model VDSR was further improved by Lim et al. [12] which got the best results in the NTIRE2017 Super-Resolution Challenge [1].

2.2 Super-Resolution in Video Domain

Video Super-Resolution (VSR) combines information from multiple adjacent LR frames to take temporal information into account, leading to higher quality results. Takeda et al. [16] apply a 3D kernel regression on a patch of adjacent LR frames to implicitly encounter temporal information. Since purposed by Caballero et al. [3] end-to-end approaches including motion compensation such as the CNN framework from [3] have large success in the VSR area. Liu et al. [13] added temporal addaptivity to the framework to be able to aggregate the resulting HR frame based on a weighted sum of several estimates as well as a varying number of input LR frames. Sajjadi et al. [14] purposed a frame-recurrent architecture iteratively using the previously inferred HR frames for the subsequent

[1] An overview of various other deep learning based approaches for SISR can be found in [23].

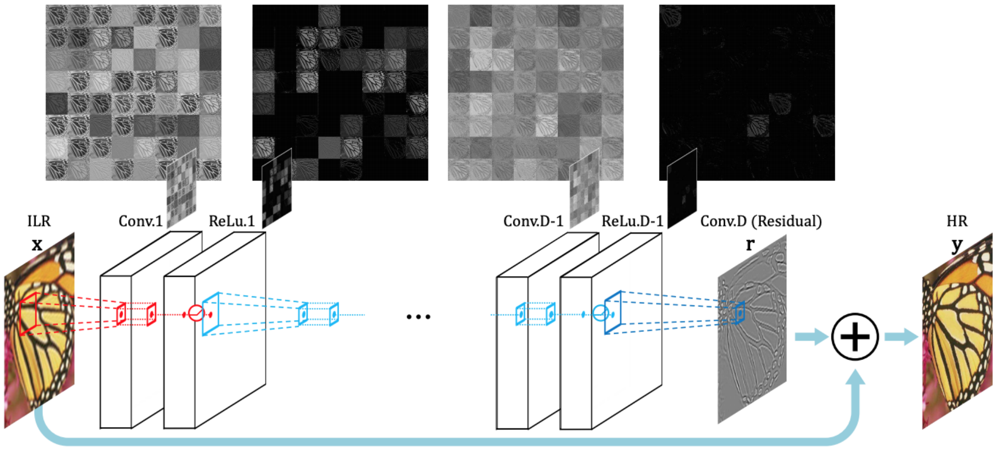

Figure 4: Overview of VDSR network design [11].

prediction. Wang et al. [18] (SOFVSR) implemented an end-to-end trainable approach to predict both, the HR frame as well as the HR optical flow. Therefore, first the HR optical flow is inferred in a coarse-to-fine manner, then motion compensation is performed according to the HR optical flows and finally, the compensated LR inputs are fed to a super-resolution network to generate the HR frame estimate (comp. Fig. 5).

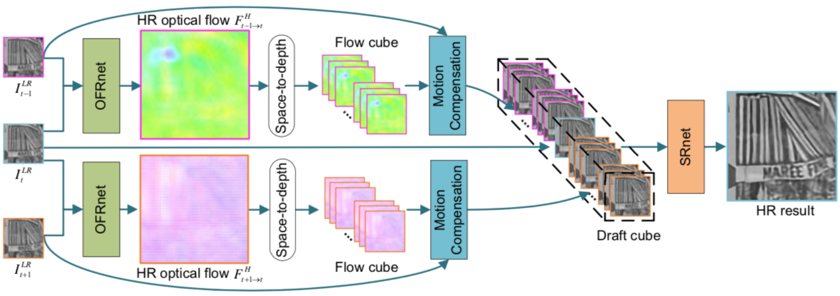

Figure 5: Overview of SOFVSR pipeline [18].

2.3 Colorization

Image colorization methods can be categorized in two categories: Non-parametric approaches, such as [7], model the correspondence between the grayscale and the colored image by finding analogeous regions in reference image(s), while parameteric models learns this correspondence from large datasets, transforming the colorization problem into a regression problem. Zhang et al. [24] (CIC) purpose posing colorization as a classification task and use class-rebalancing at training time to increase the diversity of colors in the result, using the CNN shown in Fig. 6 and not requiring any user-interaction.

2.4 Task-Aware-Downscaling

Over all of the problems stated above most of the approaches merely take into account one side of the process, e.g. by fixing the transformation HR to LR to bicubic interpolation in order to large amount of training data and focusing on estimating the inverse transformation. Kim et al. [10] (TAID) purpose taking into account the downscaling method in order to improve the upscaling performance, by training an autoencoder in an end-to-end manner while the latent space representation again is an image of same size as the LR image. The loss

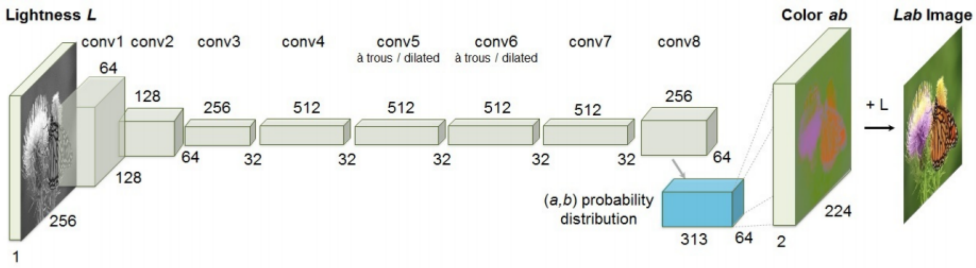

Figure 6: Overview of CIC network design [24].

function thereby contains both the difference between the decoded SHR and the original HR image as well as the difference between the encoded SLR and the bicubic interpolated LR image, such that the SLR image is a humanly understandable representation. Next to SISR the approach is shown to be applicable for large scale factor up to 128 as well as for colorization.

3 Materials and Methods

3.1 General Problem Formulation

The general idea behind TAD is that a high-dimensional input (e.g. a high-resoluted or colored image) is transformed in a low-dimensional space so that it first can be reconstructed as good as possible and second still is human-understandable in lower dimensional space. Besides, both transformations should be computationally efficient, so that an optimal trade-off between network complexity (efficiency) and reconstruction capabailites is met.

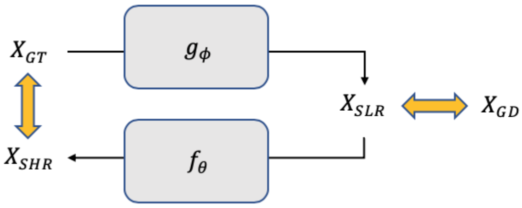

Figure 7: General TAD problem formulation.

With g_ϕ the downscaling and f_θ the upscaling function, X_{GT} the groundtruth (input) as well as X_{SLR}, X_{SHR} its low- and high- dimensional representation, the TAD problem can be formulated as combined optimization problem constraining both the low-dimensional representation (readibility) as well as the high-dimensional reconstruction (accuracy). hile the second constraint can be easily formulated using the input image X_{GT} as groundtruth the first constraint is more vague and hard to quantify. Therefore, it is assumed that the optimal latent space encoding is similar to a trivial low-dimensional representation like a (bilinearly) interpolated or grayscale image. As further described in Section 3.4 X_{SLR} is thereby not derived from scratch but builds up on the guidance image in the training procedure so that both optimization problems can be solved more independently than learning both X_{SLR} and X_{SHR} from scratch and typically the first optimization problem (readibility of X_{SLR}) is easier to solve for the model.

3.2 Autoencoder Network Design

As no groundtruth for the low resolution image is available, since TAD poses requirements for both the down- and upscaling and because it has proven to work for the TAD problem in previous works an autoencoder design is used.

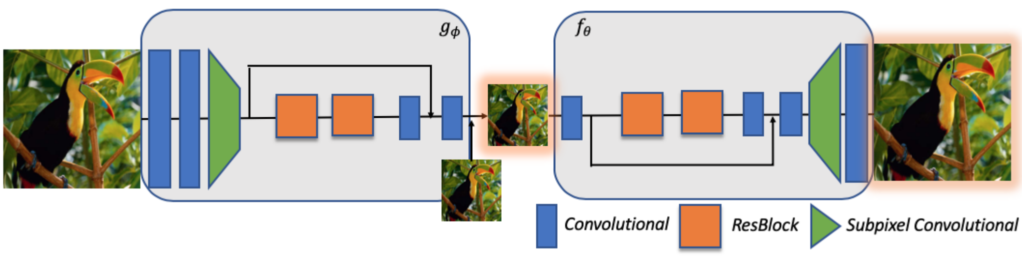

Figure 8: Example architecture of the TAD autoencoder network design for SISR task.

The autoencoder should be able to handle an input image of general size, it should be runtime-efficient, store as much information as possible while downscaling as well as end-to-end and efficiently trainable. Therefore a convolutional-only, reasonable shallow network design is used. To avoid the loss of information during downscaling instead of pooling operations subpixel convolutional layers are employed. Furthermore, in order to enable efficient training and circumvent vanishing gradient problems (especially for larger networks that were tested) next to direct forward passes ResNet ([9]) like *Resblocks* are used, which are structured as

$$Resblock(x) = x + Conv2D(ReLU(Conv2D(x)))$$

Since this network design does not continuously downscale the input but applies pixel shuffeling to downscale while all other layers do not alter their inputs shape, the networks also is easily adaptable to design changes, which simplifies the architecture optimization process.

3.3 Loss Function

The loss function L consists of two parts, representing both optimization problems introduced in Section 3.1. The first one, L_{TASK}, is task-dependent and states the difference between the decoders output X_{SHR} and the desired output X_{GT}, e.g. the original HR in the SISR task.

$$L_{TASK} = L1(X_{GT}, X_{SHR})$$

The second part, L_{LATENT}, encodes the human-readibility of the low-dimensional representation. So L_{LATENT} is the distance between the interpolated guidance image X_{GD} and the actual encoding X_{SLR}:

$$L_{LATENT} = \begin{cases} L1(X_{GD}, X_{SLR}) & \text{if } ||L1/d_{max}|| \geq \epsilon \\ 0.0 & \text{otherwise} \end{cases}$$

with $||L1/d_{max}||$ being the $L1(X_{GD}, X_{SLR})$ loss normalized by the maximal deviation between X_{GD} and X_{SLR}. Hence, L_{LATENT} is zero in an $l1$-ball around the guidance image, ensuring that task aware-low-resolution image (SLR) is close to the guidance image but also helps to prevent overfitting to the trivial solution $X_{GD} = X_{SLR} \Leftrightarrow g_\phi = 0$. As shown in Chapter 4 introducing an $l1$-ball also improves the model's robustness against perturbations.
The overall loss function is a weighted sum of both of the loss function introduced above. The relative weight (α, β) is of large importance for the trade-off between the readibility requirement and the performance of the model's upscaling part (super resolution, colorization). However, as described above the readibility requirement is less strict so that typically $\alpha >> \beta$.

$$L = \alpha L_{TASK} + \beta L_{LATENT}$$

3.4 Training Specifications

Even if a guidance image is part of the loss function learning both the low- and high-dimensional representation from scratch poses a combined optimization problem which usually is very hard to solve. To ensure (faster) convergence, therefore in the beginning of the training procedure the guidance image is added to the encoder's output. This improves both the convergence rate of X_{SLR} and X_{SHR}, especially in the beginning of the training procedure, since merely a difference between the interpolated representation and the more optimal encoding has the be derived and the down- and upscaling can be learnt more independently since the lower dimensional representation is always guaranteed to be useful for upscaling.

3.5 Implementation

The project was implemented in Python 3, using the PyTorch deep learning framework. Although some ideas from Kim et al. [10] were adopted as described above the pipeline had to be re-implemented from scratch and re-validated since neither code nor any pretrained model have been available publically (nor upon request). As PyTorch merely supports subpixel convolutional layers, their inverse transformation was implemented as well. During program development it was paid attention to generality and commutability in order to efficiently test a variety of different models and datasets as well as guarantee comparability of different approaches.

Figure 9: Loss curve without adding guidance image (left) and with adding guidance image (right) while training.

4 Experiments and Results

- PSNR for different scales
- Noise vs performance
- different lambdas
- time vs network size
- approaches of video scaling, i.e. flow, direct external
- noise resistence normal downscaling vs tar downscaling
- large scales (training directly, performance of other non-trained scales)
- scale and colorization for x4
- super large scale for videos
- torch - implementation of inverse pixel shuffeling

5 Discussion and Conclusion

A Appendix

In the appendix, list the following material:

- Data (evaluation tables, graphs etc.)
- Program code
- Further material

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017.

[2] Dong C., Loy C.C., He K., and Tang X. Learning a deep convolutional network for image super resolution. *ECCV 2014*, 2014.

[3] Jose Caballero, Christian Ledig, Andrew P. Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. *CoRR*, abs/1611.05250, 2016.

[4] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, July 2011.

[5] Claude E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8):1016–1022, 1979.

[6] M. Elad. Sparse and redundant representations: From theory to applications in signal and image processing. *Springer Publishing Company*, 2010.

[7] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. Image colorization using similar images. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 369–378, New York, NY, USA, 2012. ACM.

[8] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. *CoRR*, abs/1903.10128, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[10] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee. Task-aware image downscaling. In *ECCV*, 2018.

[11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *CoRR*, abs/1511.04587, 2015.

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *CoRR*, abs/1707.02921, 2017.

[13] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang. Robust video super-resolution with learned temporal dynamics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2526–2534, Oct 2017.

[14] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. *CoRR*, abs/1801.04590, 2018.

[15] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. pages 3791–3799, June 2015.

[16] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *IEEE Transactions on Image Processing*, 18(9):1958–1975, Sep. 2009.

[17] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *ACCV*, 2014.

[18] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through HR optical flow estimation. *CoRR*, abs/1809.08573, 2018.

[19] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. ESRGAN: enhanced super-resolution generative adversarial networks. *CoRR*, abs/1809.00219, 2018.

[20] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. *CoRR*, abs/1804.02900, 2018.

[21] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *CoRR*, abs/1902.06068, 2019.

[22] J. Yang, Z. Lin, and S. Cohen. Fast image super-resolution based on in-place example regression. pages 1059–1066, June 2013.

[23] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, and Jing-Hao Xue. Deep learning for single image super-resolution: A brief review. *CoRR*, abs/1808.03344, 2018.

[24] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. *CoRR*, abs/1603.08511, 2016.