ETH
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 6, 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

PSNR Peak signal-to-noise ratio

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 IC in both the image and the video domain have received significant attention [25]. 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. [24], [23], [9], [22]) 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 [12]. Despite of the large progress in SR in the last years ([25]) 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. ([12]) for the image domain only. However, as shown in Fig. 2 the TAD implementation purposed in [12] 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 [12]: Perturbation (left), Runtime (right)

Therefore the goals of this work are the following:

- reimplement and evaluate the TAD framework purposed in ([12])
- 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

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 [6] 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.

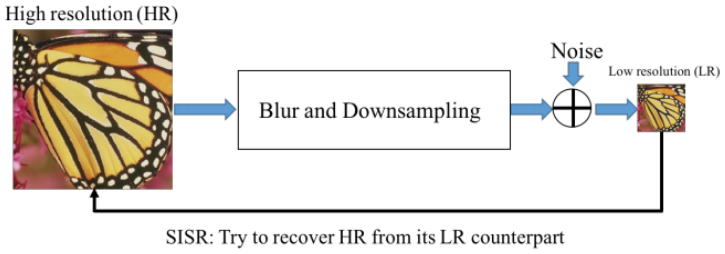

Figure 3: General SISR problem according to [27].

Sparsity-based techniques assumes the LR image to be transformable in another domain (usually a dictionary of image atoms [7]) and tries to find correspondences between the LR and HR patches in the transformed space, as implemented in [5]. However, these techniques usually are very computationally expensive. Among other learning based approaches such as the use of random forests [18], in-place example regression models [26] or adjusted anchored neighborhood regression [21], in terms of accuracy applying CNN based approaches have shown the largest success. [1] Dong et al. [3] 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. [13] (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. [14].

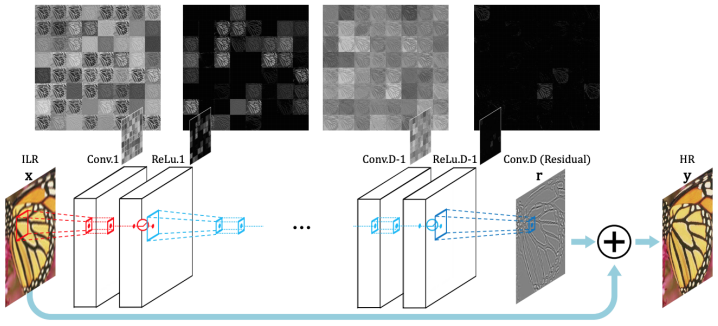

Figure 4: Overview of VDSR network design [13].

[1] An overview of various other deep learning based approaches for SISR can be found in [27].

2.2 Super-Resolution in Video Domain

VSR combines information from multiple adjacent LR frames to take temporal information into account, leading to higher quality results. Takeda et al. [19] apply a 3D kernel regression on a patch of adjacent LR frames to implicitly encounter temporal information. Since purposed by Caballero et al. [4] end-to-end approaches including motion compensation such as the CNN framework from [4] have large success in the VSR area. Liu et al. [15] 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. [17] purposed a frame-recurrent architecture iteratively using the previously inferred HR frames for the subsequent prediction. Wang et al. [22] (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. 23).[2]

2.3 Colorization

Image colorization methods can be categorized in two categories: Non-parametric approaches, such as [8], 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. [29] (CIC) purpose posing colorization as a classification task and use class-rebalancing at training time to increase the diversity of colors in the result, not requiring any user-interaction.

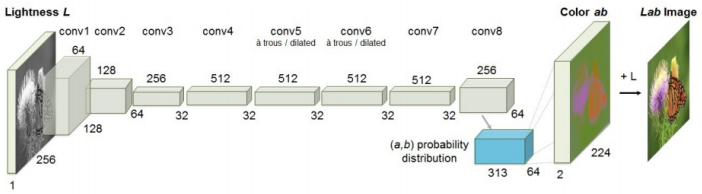

Figure 5: Overview of CIC network design [29].

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. [12] (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 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.

[2]Since SOFVSR was used in the project an overview of SOFVSR baselines can be found in the appendix.

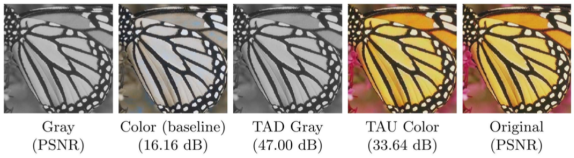

Figure 6: Qualitative results of TAID colorization for Set14 dataset.

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.

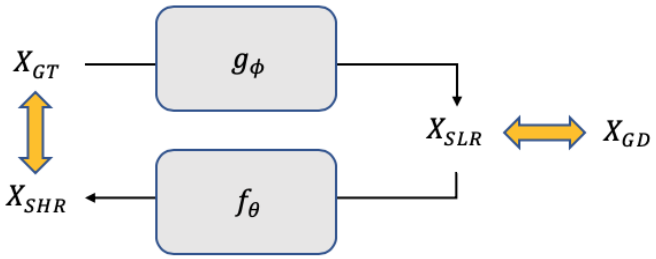

Figure 7: General TAD problem formulation.

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 Architecture

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.

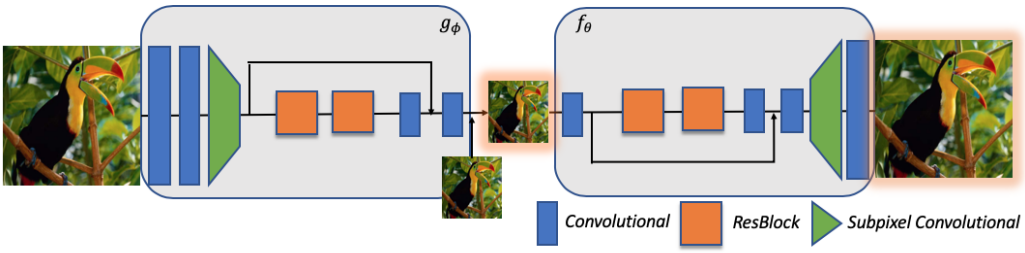

Figure 8: Example architecture of the TAD autoencoder network design for SISR task.

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 ([10]) 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 ϵ-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 ϵ-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.

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

3.5 Task-Specific Design

While the general approach derived in Section 3.1 can be trivially applied on the SISR and on the IC problem the extension to the VSR task is more advanced. In the following the specification of all three problems are presented:

Single-Image-Super-Resolution Problem

An example of the overall design of the TAD pipeline applied on the SISR problem is shown in Fig. 8. The high-dimensional space here is a high-resolution image ($X_{GT} = X_{HR}$), while the low-resolution guidance image is the bilinearly downscaled image ($X_{GD} = X_{LR}^{B}$). As shown for an example of scaling factor 4 in Fig. 10 the model can be trained more efficiently as well as performs better when iteratively scaling (i.e. scaling two times with factor 2, instead of one time with factor 4 directly).

Figure 10: Comparison between iterative and direct upscaling for an example scaling factor 4: Loss curve (left), PSNR curve (right)

Video-Super-Resolution

As pointed out in Chapter 2 the challenge of VSR compared to SISR is to not only take one frame into account but subsequent frames in order to reconstruct opaqued objects, reflect motions, etc. As shown in Chapter 2 most approaches tackling the VSR problem follow a two-step approach, firstly reconstructing the optical flow using the current and previous low-resolution, using the result to warp the image and finally upscale it. An improvement in any of these building blocks would improve the overall reconstruction, therefore there are multiplie possibilities for finding a more optimal low-dimensional representation. Several approaches were tried including a direct approach which encodes the reconstruction capability of a given VSR model into the loss function and tries to shape the low-dimensional representation such that the model's performance improves as well as an approach directly targeting the optical flow calculation (more details in the appendix).

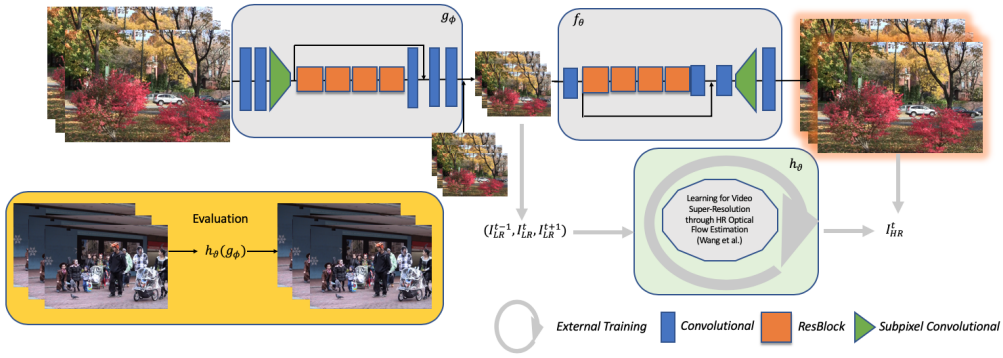

Figure 11: Design for TAD video super-resolution task (example network architecture).

The most promising approach is displayed in Fig. 11 and involves a training in multiple steps. In the first step, the autoencoder model is trained on incoherent images (similar to SISR) to learn a good general down- and upscaling transformation. In a second step the pretrained model is explicitly trained on a video dataset, in order to produce the training dataset for a VSR model, which is trained on this SLR frames specifically in the third step. After training to validate the model subsequent frames are downscaled first using the video pretrained TAD model, the downscaled images are then fed to the trained VSR model, upscaling them. While this approach basically is appliable to all VSR

frameworks in the scope of this project the SOFVSR ([22]) model was used. [3]

Image Colorization

While in a grayscale image information about intensity, color value and saturation are mingled over all channels, other color spaces split these information in seperate channels. In order to contain as much information about the original colors a non-uniformly-weighted (while grayscale would be a uniformly weighted sum, destroying color contrast information) and static (i.e. non periodic like hue in HSV color space) sum of original color values would be optimal, which is e.g. the Y channel of the YCbCr channel. Therefore it is used as guidance image ($X_{GD} = X^{Y}_{GRY}$), while the original colored image can be used as groundtruth ($X_{GT} = X_{COL}$).

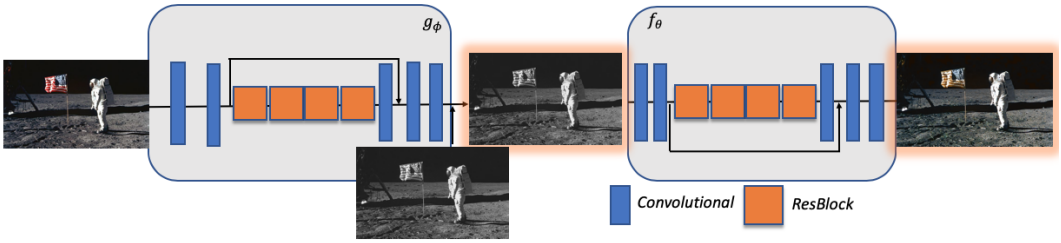

Figure 12: Design for TAD colorization task (example network architecture).

3.6 Implementation

The project was implemented in Python 3, using the PyTorch deep learning framework. Although some ideas from Kim et al. [12] 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. Since most of the open-source models for the VSR problem merely are available in Tensorflow, as is the used model, it was reimplemented and adapted to the TAD pipeline.
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. The full code stack can be found on `https://github.com/simon-schaefer/tar`.

[3] Further details about the selection of SOFVSR and a description of the approach itself can be found in the appendix.

4 Experiments and Results

In order to test the previously described TAD approach several experiments were performed, to find the optimal model for both the SISR and IC tasks, to show the impact of the L1 ball on the model's robustness against perturbations as well as evincing the feasibility of applying the TAD method in the video domain.

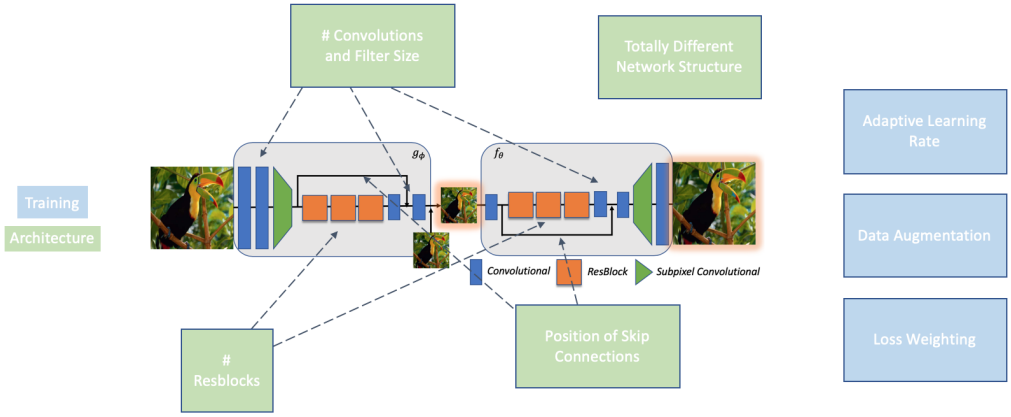

Figure 13: Model and Training adaptions during experiments in comparison to the baseline model by [12].

As shown in Fig. 13 a bunch of adjustments to the baseline model (by [12]) were tried for analysing the coherence between the model complexity and its reconstruction performance. Thereby very small architectures (6 layers) as well as comparable large architectures (19 layers) were tested (baseline model has 10 layers), spanning from 375.926 to $1.299.126$ parameters. Since the model already is quite shallow the removal of each layer had an impact on the resulting performance, therefore especially the number of *Resblocks* (two convolutional layers and ReLU) has a huge impact on the reconstruction accuracy. Next to adaptions to the baseline architecture completly new architectures have been tested, such as networks without any residual pass (so no Resblocks) but convolutional layers with either constant or varying (first increasing then decreasing) number of filters (up to 256) have been used.
Next to several architectures the baseline model was improved by advancing the training procedure. Next to enhancing the loss function as discussed in Section 3.3 instead of a constant an linearly annealing learning rate was used, starting at $4 * 10^{-4}$ and annealing by factor $\gamma = 0.25$ after 20, 100 and again after 200 training epochs. Adam optimizer was used with $\beta = (0.9, 0.999)$, $\epsilon_{ADAM} = 10^{-8}$, gradient clipping and zero weight decay.
To guarantee comparability to other super-resolution and colorization paper in the image domain the model was trained using the DIV2K training dataset ([20]), while validated on the SET5 ([2]), SET14 ([28]), BSDS100 ([16]), URBAN100 ([11]) and VDIV2K ([20]) dataset. For similar reason for the video domain the model was pretrained using DIV2K, actually trained on video clips from CDVL Database (Ntiaspen) and validated on the widely known Vid4 dataset (Calendar, Foliage, Walk, City). For improving generalization capabilities of the model and avoid overfitting the image training data were also augmented (rotated, mirrored).
Similiarly, as widely used in the field of image reconstruction as a performance measurement the PSNR will be used.

A complete list of the most important testing configurations and their results as well as a list of architectures can be found in the appendix.

4.1 Impact of L1 Ball

As already seen in Chapter 3 introducing an ϵ-ball to the loss term pervents the model from overfitting on the low-dimensional image, $X_{SLR} = X_{GD}$ (i.e. trivial solution $g_\theta = 0$ in the beginning of the training).[4] However, the main contribution of the ϵ-ball consists in the increasing robustness against perturbations of X_{SLR}, which are modeled as white Gaussian noise with standard deviation σ within this project. While a model trained with $\epsilon = 0$ is highly vulnerable to perturbations, dropping PSNR by 42% by adding noise with $\sigma = 0.11$ ($X_{SLR} \in [0, 1]^n$), a model trained with $\epsilon = 10$ is more stable dropping only about 10% in the same scenario (scale = 4, dataset = SET14).

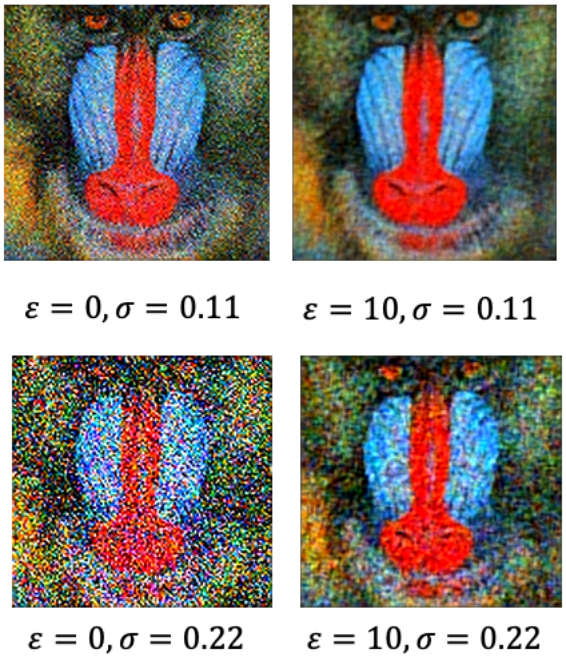

Figure 14: Qualitative comparison between the impact of perturbation on a model trained without and with ϵ-ball (scale = 4, dataset = SET14).

In the following experiment the same model (AETAD)[5] was trained with different radii of the ϵ-ball. It turns out that the right choice of ϵ is a trade-off between the reconstruction performance and robustness aginst perturbations, comp. Fig. 15.

Also an increasing ϵ improves the convergence rate during training, as shown in Fig. 16, the model otherwise first overfits to X_{GD} and then eventually finds a more optimal trade-off between fitting the low- and high-dimensional image (with increasing $\frac{\alpha}{\beta}$ ratio). However, as Fig. 15 the radius of the ϵ-ball around X_{GD} cannot be choosen infinitely large, since the overall performance worsens as the impact of the guidance image on the model convergence decreases (e.g. $\epsilon = 100$ in Fig. 15).

[4] As discussed in Chapter 3 the training is splitted into two parts, first learning only the difference between the guidance image and a more optimal representation and later learning the low-dimensional representation independent from the

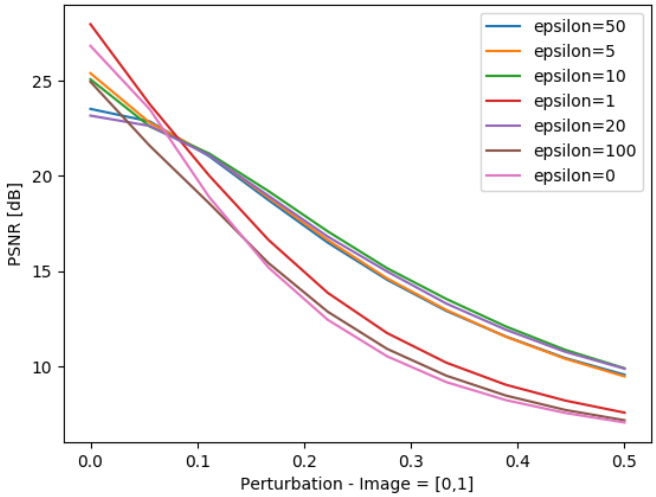

Figure 15: Reconstruction performance over several values of ϵ and σ (scale = 4).

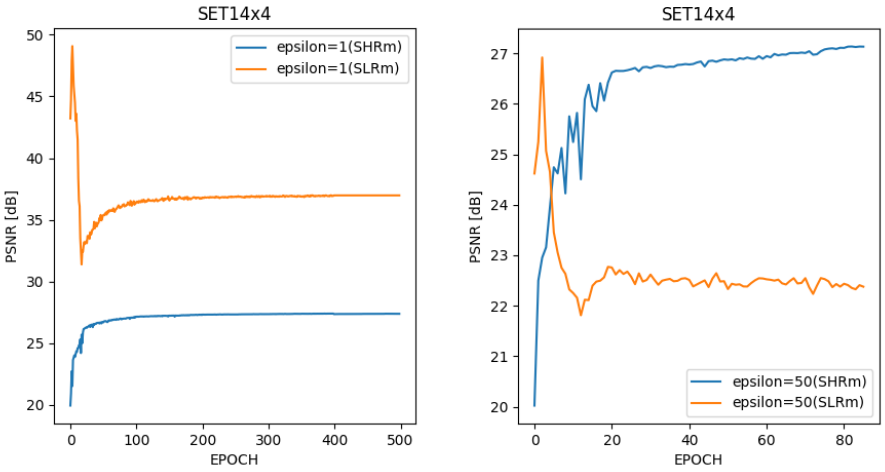

Figure 16: Closeness between X_{SLR} and guidance image (blue) and X_{SHR} and groudtruth image (orange) for different values of ϵ (left: $\epsilon = 1$, right $\epsilon = 50$).

Table 1 displays the impact of models trained with different values of ϵ on the reconstruction capabilities on non-trained scales. In general the model overfits on a scale it is trained on, however allowing to deviate from the guidance image by increasing ϵ amplifies the effect of overfitting the learnt transformation to a the scaling factor the model is trained on. Hence, as shown below especially for large deviations in scaling factor the accuracy worsens with increasing ϵ.

epsilon	x2	x4	x8	x16
0	8.446	24.406	8.555	18.938
20	6.244	22.188	6.492	17.308
50	6.771	22.459	6.925	16.992
100	10.901	23.784	11.870	13.225

Table 1: Impact of ϵ on performance on non-trained scales (trained scale = 4, dataset = SET14).

Overall the choice of ϵ depends very much on the application and its external conditions that should be solved, e.g. whether perturbations are probable for example during the storing, downloading etc. process. In case only the performance without any disturbances matters, $\epsilon < 10$ is a good choice, as it fastens convergence of the model, prevents overfitting on the guidance image but also does not affect the accuracy without any disturbances much. For the further experiments a value $\epsilon = 1$ was used.
Although described for the SISR problem here the described impact of the ϵ-ball is similar over all other problems, which omitted here for the matter of compactness of this report.

4.2 Single-Image Super-Resolution

Next to improving the robustness of the TAD model several improvements were made on both the way of training as well as on the architecture itself. Fig. 17 shows the correlation on both the SET5 and the SET14 dataset for different architectures which were trained using the same parameters (learning parameters as described above, $\epsilon = 10$) for the sake of a comparability.

Several conclusion can be drawn from Fig. 17, which could be confirmed also for the other validation datasets used:

- The model *no_tad* is the baseline model which is trained to only upscale based on the guidance image, i.e. the standard approach to solve the SISR problem. Although the reconstruction performance of the *no_tad* model is worse than state-of-the-art methods (e.g. VDSR [13] shown in Fig. 4), clearly the task-aware downscaling approach could largely improve the performance, compared the equivalent *aetad* model which is trained using task aware downscaling. In fact for similar performance more than about 25% of the model parameters can be omitted, still resulting in a better accuracy (comparison *no_tad* vs *aetad_small*).

- The *Resblocks* have a large impact on the models performance, purely convolutional models with neither a skip connection nor a ReLU layer (as occurring in *Resblocks*), such as the models *conv_only* or *conv_only_very_large*, have an overall worse reconstruction performance with a similar number of parameters.

guidance image.

[5] An overview over all model architectures can be found in the appendix.

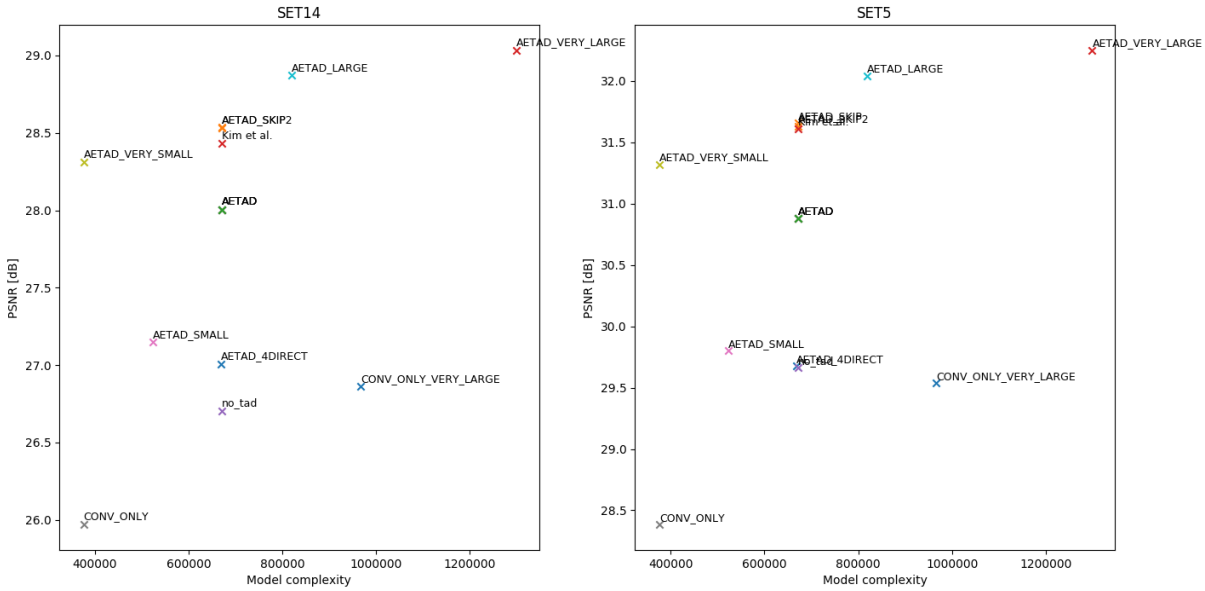

Figure 17: Model complexity vs reconstruction performance for SISR problem (scale = 4, dataset = SET14 and SET5).

- In a direct comparison the model *aetad_direct4* performs worse than the iteratively scaling structured, but otherwise equivalent model *aetad*.
- In both displayed validation datasets the reconstruction performance stagnates with increasing number of parameters, as the difference in PSNR between the *aetad_large* and the *aetad_very_large* model does not improve much anymore.

While more complex architectures than the baseline (*Kim et al.*-model) do not gain a lot of accuracy, for less complex models with similar architecture there is a large drop in accurcy. Therefore, the baseline architecture already is very reasonable.

Fig. 18 shows the PSNR curves for a model trained on a scaling factor of 2, while being validated on both of the scaling factor 2 and 4.

model	x2	x4	x8	x16
no_tad	7.457	27.204	7.459	19.488
aetad_very_small	8.527	28.334	7.780	18.799
aetad_very_large	5.882	29.617	5.928	18.673

Table 2: Comparison of the reconstruction performance on non-trained scales between task aware and non task aware trained models on SET14.

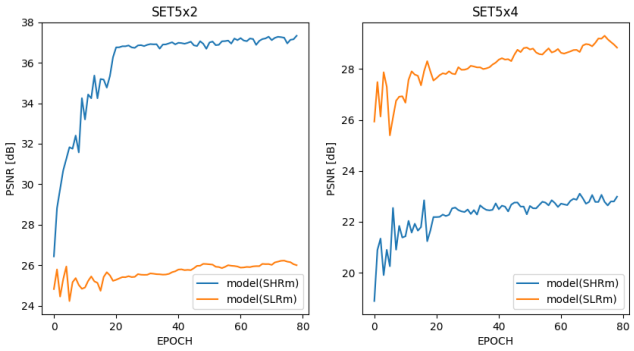

Figure 18: PSNR curve on SET5 validation dataset for different scales (trained scaling factor = 2).

scale	dataset	PSNR (Kim et al.)	PSNR (*aetad_skip2*)	PSNR (*aetad_very_small*)
x4	SET5	31.81	31.814	30.302
x4	SET14	28.63	28.665	28.334
x4	URBAN100	26.63	24.156	23.084
x4	BSDS100	28.51	28.601	25.719

Table 3: Comparison of reconstruction accuracy between Kim et al. [12], *aetad_skip2* and *aetad_very_small* model for scaling factor 4 on several validation datasets.

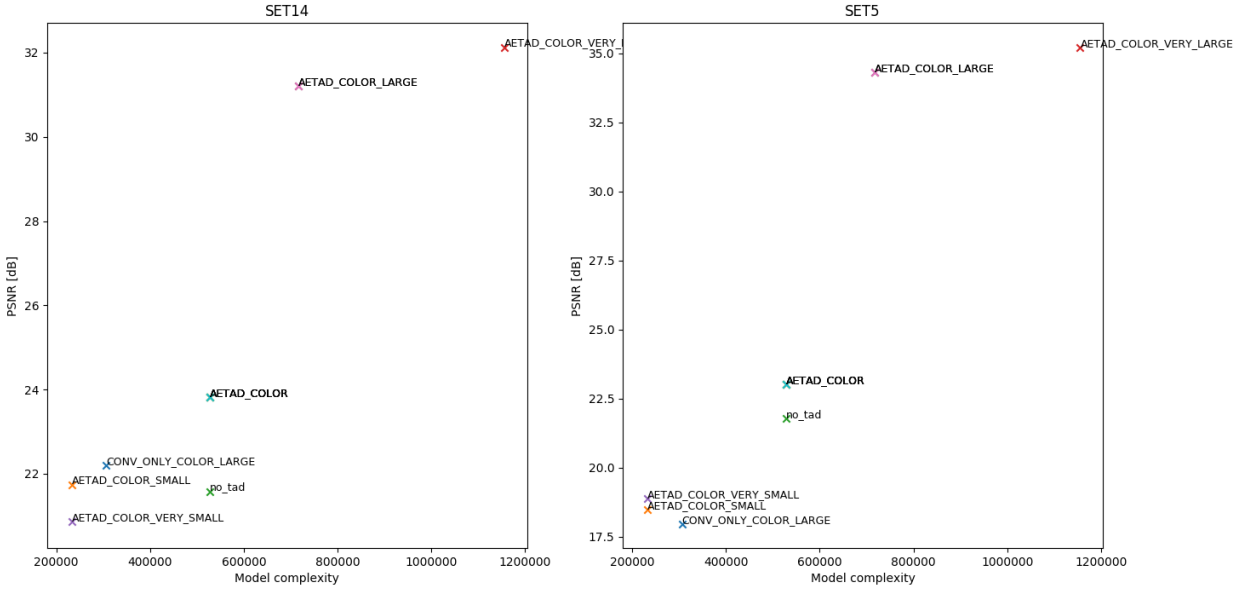

Figure 19: Model complexity vs reconstruction performance for IC problem (dataset = SET14 and SET5).

scale	dataset	PSNR (Kim et al.)	PSNR (*aetad_color_large*)
x4	SET5	None	34.416
x4	SET14	None	31.262
x4	URBAN100	33.68	33.604
x4	BSDS100	36.14	36.786

Table 4: Comparison of reconstruction accuracy between Kim et al. [12], *aetad_color_large* model on several validation datasets.

4.3 Image Colorization

4.4 Video Super-Resolution

scale	dataset	SISR model	non task aware SOFVSR	task aware SOFVSR
x4	CALENDAR	21.297	19.190	18.573
x4	CITY	25.332	24.677	24.191

Table 5: Comparison of reconstruction accuracy of SISR model and the rebuild SOFVSR model without and with TAD.

1. super-large scale for videos

4.5 Qualitative Improvements

As demonstrated above TAD is able to improve the performance of image reconstruction models quantitatively, but the TAD approach also improves the results in a qualitative manner, in sense of that images can be restored that would be able to be restored from the trivially downscaled image, in the following shown using the example of a IC task.
Consider Fig. 20 displaying objects with equivalent shape but different colors (gummibears). While by using trivial downscaling methods like averaging over the colors (grayscale) some color information are unrecoverably lost, e.g. the yellow and orange gummibear looking nearly equivalent in grayscale, a task aware approach learns to keep basic color information despite of downscaling, figuratively speaking. Hence, even the color of shapes which are (exactly) similar in grayscale can be restored.

Figure 20: Image colorization of similar shapes (from upper left to lower right: grayscale, groundtruth, colorized without TAD, colorized with TAD)

5 Discussion and Conclusion

A Appendix

Video-Super-Resolution using Optical Flow Reconstruction

One approach that was tried in order to apply TAD on VSR was to find a low-dimensional representation that is optimal to reconstruct the optical flow. Fig. 21 indicates the design for this approach, which was trained using the UCL optical flow dataset ([1]) and the OpenCV Farneback dense optical flow function (since easily available for a conceptual test). Unfortunately, the approach did not converge to useful representation. Since other approaches could successfully solve the VSR task this approach was discarded.

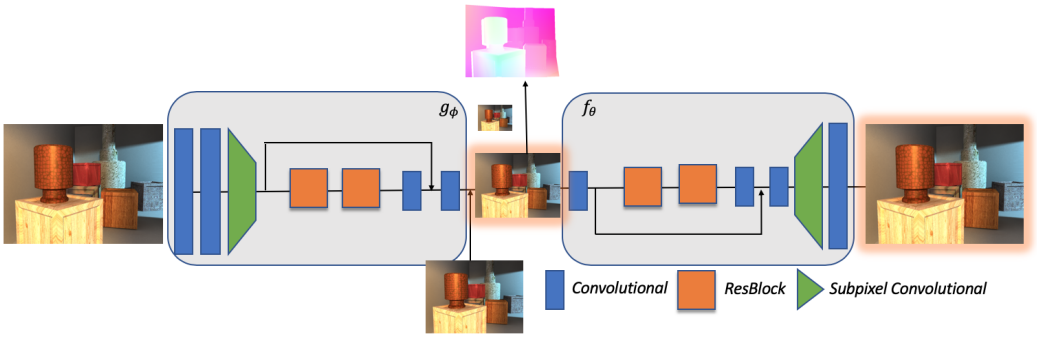

Figure 21: Design for TAD video super-resolution task (example network architecture) for optical flow reconstruction.

Figure 22: Optical flow reconstruction: Groundtruth (left) and reconstructed image (right).

SOFVSR model selection and working principle

Wang et al. [22] (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. 23).

The SOFVSR model was selected based on a combination of criteria such as the stated reconstruction performance (PSNR) and runtime, the closeness to state-of-the-art and the availability of a PyTorch open-source implementation.

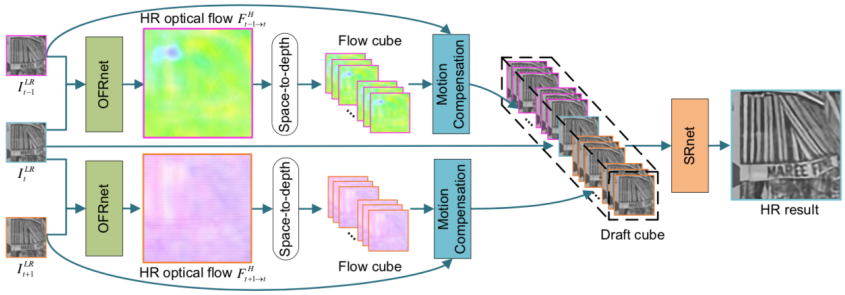

Figure 23: Overview of SOFVSR pipeline [22].

Rank	Method	PSNR	SSIM	MOVIE	Paper Title	Year	Paper	Code
1	VSR-DUF	27.31	0.832		Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation	2018		
2	RBPN/6-PF	27.12	0.818		Recurrent Back-Projection Network for Video Super-Resolution	2019		
3	FRVSR	26.69	0.822		Frame-Recurrent Video Super-Resolution	2018		
4	SOF-VSR	26.01	0.771	4.32	Learning for Video Super-Resolution through HR Optical Flow Estimation	2018		
5	DRDVSR	25.88	0.774		Detail-revealing Deep Video Super-resolution	2017		
6	DBPN	25.37	0.737		Deep Back-Projection Networks For Super-Resolution	2018		
7	VESPCN	25.35	0.7557	5.82	Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation	2016		
8	ESPCN	25.06	0.7394	6.54	Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network	2016		
9	SRCNN	24.68	0.7158	6.90	Image Super-Resolution Using Deep Convolutional Networks	2014		
10	BRCN	24.43	0.662		Bidirectional Recurrent Convolutional Networks for Multi-Frame Super-Resolution	2015		

Figure 24: Comparison of several VSR methods according to PapersWithCode.

List of Experiments

List of Architectures

Figure 25: List of important (non-complete) experiments.

Figure 26: List of important (non-complete) architectures used.

References

[1] Simon Baker, Daniel Scharstein, J.P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 2011.

[2] M. Bevilacqua, A. Roumy, C. Guillemot, and M.L Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *BMVC*, 2012.

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

[4] 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.

[5] 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.

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

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

[8] 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.

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

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

[11] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. *CVPR*, 2015.

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

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

[14] 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.

[15] 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.

[16] D.R. Martin, C.C. Fowlkes, D. Tal, and J Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2001.

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

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

[19] 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.

[20] Radu Timofte, Shuhang Gu, Jiqing Wu, Luc Van Gool, Lei Zhang, Ming-Hsuan Yang, Muhammad Haris, et al. Ntire 2018 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

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

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

[28] R. Zeyde, M. Elad, and M Protter. On single image scale-up using sparse-representations. *Proceedings of the International Conference on Curves and Surfaces*, 2010.

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