

Single-frame Regularization for Temporally Stable CNNs

Gabriel Eilertsen¹ Rafał K. Mantiuk² Jonas Unger¹

¹Dept. of Science and Technology, Linköping University, Sweden

²Dept. of Computer Science and Technology, University of Cambridge, UK

{gabriel.eilertsen, jonas.unger}@liu.se rafal.mantiuk@cl.cam.ac.uk

Abstract

Convolutional neural networks (CNNs) can model complicated non-linear relations between images. However, they are notoriously sensitive to small changes in the input. Most CNNs trained to describe image-to-image mappings generate temporally unstable results when applied to video sequences, leading to flickering artifacts and other inconsistencies over time. In order to use CNNs for video material, previous methods have relied on estimating dense frame-to-frame motion information (optical flow) in the training and/or the inference phase, or by exploring recurrent learning structures. We take a different approach to the problem, posing temporal stability as a regularization of the cost function. The regularization is formulated to account for different types of motion that can occur between frames, so that temporally stable CNNs can be trained without the need for video material or expensive motion estimation. The training can be performed as a fine-tuning operation, without architectural modifications of the CNN. Our evaluation shows that the training strategy leads to large improvements in temporal smoothness. Moreover, for small datasets the regularization can help in boosting the generalization performance to a much larger extent than what is possible with naïve augmentation strategies.

1. Introduction

Deep neural networks (DNNs) can represent complex non-linear functions but tend to be very sensitive to the input. For image data, this is manifested in sensitivity to small changes in pixel values. For example, techniques for generating adversarial examples have demonstrated that there exists images that are visually indistinguishable from each other, while generating widely different predictions [32]. It is also possible to find naturally occurring image operations that can cause a convolutional neural network (CNN) to fail in the learned task [2, 9, 39]. For image-to-image CNNs applied on video sequences, this sensitivity results in abrupt and incoherent changes from frame to

frame. Such changes are seen as flickering, or unnatural movements of local features. Previous methods for applying CNNs to video material most often use dense motion information between frames in order to enforce temporal coherence [29, 13, 6, 24]. This requires ground truth optical flow for training, modifications to the CNN architecture, and computationally expensive training and/or prediction. Moreover, there are many situations where reliable correspondences between frames cannot be estimated, *e.g.* due to occlusion or lack of texture.

Instead of relying on custom architectures, we take a simple, efficient, and general approach to the problem of CNN temporal stability. We pose the stability as a regularizing term in the loss function, which potentially can be applied to any CNN. We formulate two different regularizations based on observations of the expected behavior of temporally varying processing. The result is a light-weight method for stabilizing CNNs in the temporal domain. It can be applied through fine-tuning of pre-trained CNN weights and requires no special-purpose training data or CNN architecture. Through extensive experimentation for the application in colorization and single-exposure high dynamic range (HDR) reconstruction, we show the efficiency of the regularization strategies.

In summary, this paper explores regularization for the purpose of stabilizing CNNs in the temporal domain and presents the following main contributions:

- Two novel regularization formulations for temporal stabilization of CNNs, which both model the dynamics of consecutive frames in video sequences.
- A novel perceptually motivated smoothness metric for evaluation of the temporal stability.
- An evaluation showing that the proposed training technique improves temporal stability significantly while maintaining or even increasing the CNN performance.
- For scenarios with limited training data, the generalization performance of the regularization strategies is significantly better than traditional data augmentation.

2. Background and previous work

Adversarial examples: Adversarial examples introduce minor perturbations of an input image, which makes a DNN classifier to fail [32, 12], also without access to the particular model [23], and by performing natural image operations [2, 9, 39]. This points to the large sensitivity to the input of DNNs and, for image-to-image CNNs, it is manifested in inconsistent changes between frames when applied to a video sequence. Our goal is to train for robustness when it comes to the type of changes that can occur between frames in video sequences, so that video processed with CNNs can be expected to be well-behaved. This does not mean that the CNN will be robust to other types of changes, such as those created by certain adversarial example generation methods.

Regularization: While there exists a wide range of methods that classify as regularization [22], we are particularly interested in those that are designed to address the issue of neural networks input sensitivity. Depending on the context and different definitions, the terms invariance, robustness, insensitivity, stability, and contraction have been used interchangeably in the literature for describing the objective of such regularization.

The most straightforward method for increasing robustness and generalization is to employ data augmentation. However, augmentation alone cannot compensate for a CNN’s sensitivity to transformations of the input [2, 9] or degradation operations [39]. It would require too much training data to learn robustness for all transformations, and will most likely result in under-fitting. An explicit constraint needs to be enforced to learn a mapping that is smooth, so that small changes in input yield to small changes in output. This concept has been explored in a variety of formulations, *e.g.* by means of weight decay [21], weight smoothing [19], label smoothing [38], or penalizing the norm of the output derivative with respect to the network weights [14]. Of particular interest to our problem are methods that regularize by penalizing the norm of the Jacobian with respect to the input [28, 39]. For example, Zheng *et al.* [39] apply noise perturbations to the input images, and construct a regularization term that contracts the prediction of clean and noisy samples, resulting in an increased robustness to image degradation.

While the aforementioned works mostly deal with classification, we show that the same reasoning is true for image-to-image CNNs applied to video sequences — we cannot simply train a CNN on separate video frames, or transformed images by means of augmentation, and expect a robust behavior for temporal variations. Therefore, we formulate different regularization strategies particularly for training CNNs for video applications, and perform a study on which is most efficient for achieving temporal stability.

Temporal consistency: Methods for enforcing temporal consistency in image processing are mostly based on estimating dense motion, or optical flow, between frames [26, 4, 7, 37]. This is also the case for previous work in temporally consistent CNNs. For example, flow-based methods have been suggested for video style transfer [29, 13], video-to-video operations by means of generative adversarial networks (GANs) [34], and for imposing temporal consistency as a post-processing operation [24].

Another direction for video inference using neural networks is to employ recurrent learning structures, such as the long short-term memory (LSTM) networks [15]. For image data, CNNs have been constructed for recurrence using the ConvLSTM [36] and its variants [20], which have been used *e.g.* in video super-resolution [33] and video consistency methods [24]. However, mostly these structures have been explored in classification and understanding. There are also other recurrent or multi-frame based structures that have been used for image-to-image applications, *e.g.* for video super-resolution [16, 5], de-blurring [31], and different applications of GANs [34].

The flow-based and recurrent methods all suffer from one or more of the following problems: 1) high complexity and application specific architectural modifications, 2) need for special-purpose training data such as video frames and motion information, 3) a significant increase in computational complexity for training and/or inference, 4) failure in situations where motion estimation is difficult, such as image regions with occlusion or lack of texture. The strategy we propose handles all these limitations. It is lightweight, can be applied to any image-to-image CNN without changes, and does not require video material or motion estimation. At the same time, it offers great improvements in temporal stability without impeding the reconstruction performance.

3. Temporal regularization

We consider supervised training of image-to-image CNNs, with the total loss formulated as:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{rec} + \alpha\mathcal{L}_{reg}. \quad (1)$$

The first term is the main objective of the CNN, which promotes reconstruction of ground truth images y from the input images x . Given an arbitrary CNN that has been trained with the loss \mathcal{L}_{rec} , adding the term \mathcal{L}_{reg} is the only modification we make in order to adapt a CNN for video material. The scalar α is used to control the strength of the regularization objective.

This section presents three different regularization strategies, \mathcal{L}_{reg} in Equation 1, for improving temporal stability of CNNs. The first was introduced by Zheng *et al.* [39], while the two others are novel definitions that are specifically designed to account for frame-to-frame changes in video. All

three strategies rely on performing perturbations of the input image, and a key aspect is to model these as common transformations that occur in natural video sequences.

3.1. Stability regularization

The most similar to our work is the stability training presented by Zheng *et al.* [39]. Given an input image x , and a variant of it with a small perturbation $T(x) = x + \Delta x$, the regularization term is formulated to make the prediction of both images possibly similar. For an image-to-image mapping f , we can apply the term directly on the output image,

$$\mathcal{L}_{stability} = \|f(x) - f(T(x))\|_2. \quad (2)$$

While different distance measures can be used, we only consider the ℓ_2 norm for simplicity. The perturbation Δx is described as per-pixel independent normally distributed noise, $\Delta x \sim \mathcal{N}(0, \Sigma)$, with $\Sigma = \sigma^2 I$.

3.2. Transform invariance regularization

The typical measure of temporal incoherence [26, 4] is formulated using two consecutive frames y_{t-1} and y_t ,

$$E = \|y_t - W(y_{t-1})\|_2, \quad (3)$$

where W describes a warping operation from frame $t-1$ to t using the optical flow field between the two frames. If there are frame-to-frame changes that cannot be explained by the flow field motion, these are registered as inconsistencies.

In order to use this measure for regularization, without requiring video data or optical flow information, we introduce within-frame warping with a geometric transformation, $W(x) = T(x)$ (the transformation is described in more detail in Section 3.4). Then, x and $T(x)$ mimic two consecutive frames, which are used to infer $f(x)$ and $f(T(x))$. If these are temporally consistent, performing the warping to register the two frames should yield the same result, either comparing $f(x)$ to $T^{-1}(f(T(x)))$ or comparing $T(f(x))$ to $f(T(x))$. This results in the regularization term:

$$\mathcal{L}_{trans-inv} = \|f(T(x)) - T(f(x))\|_2. \quad (4)$$

Note that this loss is fundamentally different from the standard reconstruction loss for an augmented patch:

$$\mathcal{L}_{augment} = \|f(T(x)) - T(y)\|_2. \quad (5)$$

While $\mathcal{L}_{augment}$ promotes an accurate reconstruction with respect to an augmented (transformed) sample, $\mathcal{L}_{trans-inv}$ promotes the reconstruction that is consistent with a transformation, but not necessarily accurate. If there is an error in the reconstruction, $\mathcal{L}_{augment}$ will minimize that error in the transformed (augmented) patch, potentially at the cost of consistency, while $\mathcal{L}_{trans-inv}$ will ensure that any error is consistent between the original and the transformed patches.

3.3. Sparse Jacobian regularization

Supervised learning typically relies on fitting a function to a number of training points without considering what is the function behavior in the neighborhood of those points. It would be arguably more desirable to provide to the training not only the function values, but also the information about partial derivatives in a form of a Jacobian of that function at a given point. However, for typical image-to-image CNNs, using a full Jacobian matrix would be impractical: if 32×32 patches are used, we need to train $f : \mathbb{R}^{1024} \rightarrow \mathbb{R}^{1024}$ and the Jacobian has over a million elements. However, we are going to demonstrate that even if we use a sparse estimate of the Jacobian and sample just a few random directions of our input space, we can much improve stability and accuracy of the predictions.

By providing sparse information on the Jacobian, we can also infuse domain expertise into our training. In the case of image-to-image mapping, we know that an input patch transformed by translation, rotation and scaling, should result in a transformed output patch. Each of those transformations maps to a vector change in the input and output space, for which we can numerically estimate partial derivatives. That is, we want the partial derivatives of the trained function f to be possibly close to those of the ground truth output patches:

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} \approx \frac{y(x + \Delta x) - y(x)}{\Delta x}, \quad (6)$$

where Δx represents the effect of one of the transformations on the input space, $y(x)$ is the output patch from the training set corresponding to x , and $y(x + \Delta x)$ is the transformed output patch. For the consistency of notation, we define $T(x) = x + \Delta x$ and $T(y) = y(x + \Delta x)$, so that we can formulate a regularization term as:

$$\mathcal{L}_{jacobian} = \|(f(T(x)) - f(x)) - (T(y) - y)\|_2 \quad (7)$$

$$= \|(f(T(x)) - T(y)) - (f(x) - y)\|_2. \quad (8)$$

Although the term may look similar to $\mathcal{L}_{augment}$ from Equation 5, $\mathcal{L}_{jacobian}$ promotes consistency rather than accuracy: the loss is minimized when the prediction error for the transformed patches is similar to the prediction error for the original patches.

3.4. Transformation specification

The perturbation function $T(\cdot)$ in all of the introduced regularization terms rely on a transformation of the input image. For our purpose, this should capture the possible motion that can occur between frames in a video sequence. We make use of simple geometric transformations in order to accomplish this. These include translation, rotation, zooming, and shearing, which all can be described in a 2×3

Table 1. Ranges of transformation parameters.

Parameter	Min	Max
Translation	-2 px	2 px
Rotation	-1°	1°
Zoom	$0.97\times$	$1.03\times$
Shearing	-1°	1°

transformation matrix that transforms the indices of the image x (see supplementary material for an exact formulation). The matrix is randomly specified for each image, with transformation parameters drawn from uniform distributions in a selected range of values as specified in Table 1.

Although motion can occur on a more local level in real videos, we argue that the transformations can make for good regularization. It is important to note that we do not train the network to predict the transformation or transformed patch, which would arguably require training on local transformations. Instead, we train to produce an image in which pixels do not shift and which is consistent with the input in the presence of any transformation, both local or global. We argue that the type of motion (global/local) is in this case less relevant as long as the regularization term pushes the trained model towards predicting consistent results.

3.5. Implementation

While it is possible to train for a loss function with one of the regularization terms from scratch, we instead start with a pre-trained network and include the regularization in a second training stage for fine-tuning. We found that fine-tuning makes training convergence more stable while providing the same gain in temporal consistency as training from scratch. Another very important advantage is that fine-tuning can be applied to already optimized large-scale CNNs, which take long time to train.

For each regularization method, we follow the exact same loss evaluation scheme. The perturbed sample’s coordinates are transformed as described in Section 3.4, with randomly selected transformation parameters. Both the original and the transformed sample, x and $T(x)$, respectively, are taken through the CNN by the means of a weight-sharing (siamese) architecture. This gives us $f(x)$ and $f(T(x))$, which can be used with the three different regularization definitions, Equation 2, 4, and 8, by complementing with the transformations $T(f(x))$ and $T(y)$.

4. Experiments

We evaluate the novel temporal CNN stabilization/regularization techniques using two different applications: colorization of grayscale images and HDR reconstruction from single-exposure images. These tasks were selected as they are different in nature, and rely on different CNN architectures. While colorization attempts to infer colors over the complete image, the HDR reconstruction tries to recover

Table 2. CNN training setups used in the evaluation experiments.

	Colorization	HDR reconstruction
Architecture	Autoencoder [17]	Autoencoder [8]
Down-sampling	Strided conv.	Max-pooling
Up-sampling	Resize + conv.	Transposed conv.
Skip-connections	No	Yes
Weights	1,568,698	1,289,653
Training data	CelebA [27]	Procedural images
Resolution	128×128	128×128
Training size	20,000	10,000
Epochs	50	50
Training time	$\approx 35\text{m}$	$\approx 20\text{m}$

local pixel information that have been lost due to sensor saturation. The colorization CNN uses the same design as described by Iizuka *et al.* [17], but without the global features network and with fewer weights. It implements an autoencoder architecture, with strided convolution for down-sampling, and nearest neighbor resizing followed by convolution for up-sampling. The HDR reconstruction CNN uses the same design as described by Eilertsen *et al.* [8], but with fewer weights. This is also an autoencoder architecture, but implemented using max-pooling and transposed convolution, and it has skip-connections between encoder and decoder networks. More details on the CNNs and training setups are listed in Table 2.

In order to be able to explore a broad range of hyperparameters, we use datasets that are restricted to specific problems. For colorization, we only learn the task for close-up face shots. For the HDR reconstruction, we restrict the task to a simple procedural HDR animation.

Training data for the colorization task is 20,000 images from the CelebA dataset [27]. For testing, we use 72 video sequences from the YouTube Faces dataset [35]. These have been selected to show close-up faces in order to be more similar to the training data, and are cut to be between 50 – 200 frames long. Figure 1 shows an example of a test frame.

Training data for the HDR reconstruction task is 10,000 frames that have been generated in a completely procedural manner. These contain a random selection of image features with different amount of saturated pixels. The features move in random patterns and are sometimes occluded by randomly placed beams. For the training data we only use static images, with no movement, and for the test data we include motion to evaluate the temporal behavior. The test set consists of 50 sequences, 200 frames each. Figure 2 shows an example of a test video frame.

4.1. Performance measures

The goal of the proposed regularization strategies is to achieve temporally stable results while maintaining the reconstruction performance. In order to evaluate whether both goals are achieved, we measure reconstruction performance by means of PSNR and introduce a new measure of

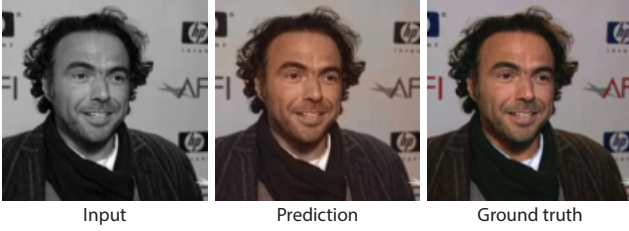


Figure 1. Colorization test sample, from YouTube Faces [35].

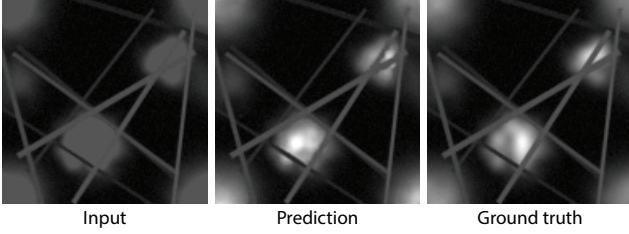


Figure 2. HDR reconstruction test sample, from procedural HDR video dataset. The image is displayed at a reduced exposure in order to show differences in saturated regions.

smoothness over time. Our measure computes the ratio of high temporal frequencies between the reference and reconstructed video sequences. We first extract the energy of the high temporal frequency component from both sequences,

$$D(f(x))_{i,j,t} = |f(x)_{i,j,t} - (G_\sigma * f(x))_{i,j,t}|^2, \quad (9)$$

where the convolution with the Gaussian filter G_σ is performed in the temporal dimension t . The parameter σ is selected to eliminate the low frequency components that the eye is insensitive to, but which carry high energy. Figure 3 shows the spatio-temporal contrast sensitivity function of the visual system and the high-pass filter we use with $\sigma = 0.15$ seconds. The smoothness is computed as the ratio of the sum of the ground truth and the reconstruction video energies,

$$S = \sqrt{\frac{\sum_{i,j,t} D(y)_{i,j,t}}{\sum_{i,j,t} D(f(x))_{i,j,t}}}. \quad (10)$$

If $S < 1$, the reconstructed video is less smooth than the ground truth video and the opposite can be said for $S > 1$.

4.2. Experimental setup

We fine-tune the CNNs in Table 2 for the two applications, and run a large number of trainings in order to sample the performance at different settings. For the total loss in Equation 1, we compare the three different regularization formulations: stability (2), sparse Jacobian (8), and transform invariance (4). These are evaluated using the transformation described in Section 3.4. For the stability regularization we also include a setting with noise perturbations, $T(x) = x + \Delta x$, with $\Delta x \sim \mathcal{N}(0, \sigma^2 I)$, in order to compare to previous work. We choose different σ for each image, drawn from a uniform distribution, $\sigma \sim \mathcal{U}(0.01, 0.04)$.

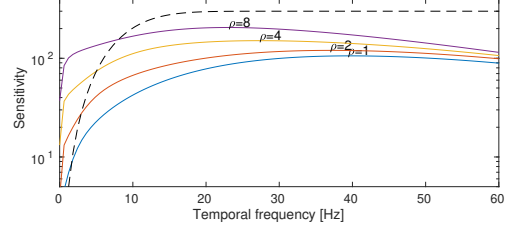


Figure 3. Colored lines: Spatio-temporal contrast sensitivity function for different spatial frequencies ρ in cycles per visual degree, based on the model from [25]. Dashed-black line: the high-pass filter used for our smoothness measure.

Finally, we also include trainings that use traditional augmentation by means of the transformation. For each of the aforementioned setups, we then run 10 individual trainings in order to estimate the mean and standard deviation of each datapoint.

We also experimented with incorporating the reconstruction loss of the transformed sample, Equation 5, but mostly this degraded the performance, possibly due to underfitting.

4.3. Results

The results of the experiments can be found in Figure 4 for colorization and in Figure 5 for HDR reconstruction. The baseline condition uses the pre-trained model before fine-tuning and without regularization. The PSNR and smoothness measures have been calculated on the a and b channels of the CIE Lab color space for the colorization application and only in saturated pixels for the HDR reconstruction application. Such modified measures can better capture small differences.

In both experiments we can observe significant improvements in both PSNR and smoothness for all regularization strategies. However, the *stability* that relies on noise performs visibly worse in both experiments than the same regularization but based on transformations. *Transform invariance* and *sparse Jacobian* regularizations result in higher PSNR and visually better reconstruction than the *stability* regularization (refer to the video material). Although the *stability* formulation can generate smoother video for HDR reconstruction, this is at the cost of very high reconstruction error, and for $\alpha > 0.99$ it most often learns the identity mapping, $f = x$. The performance of the two novel formulations are comparable. The sparse Jacobian results in a slightly higher PSNR for HDR reconstruction and transform invariance results in higher smoothness. The sparse Jacobian also seems to be more robust to the choice of the regularization strength. The traditional augmentation using the transformations (the blue-dashed line) can improve smoothness and PSNR but the improvement is much smaller than the other regularization strategies.

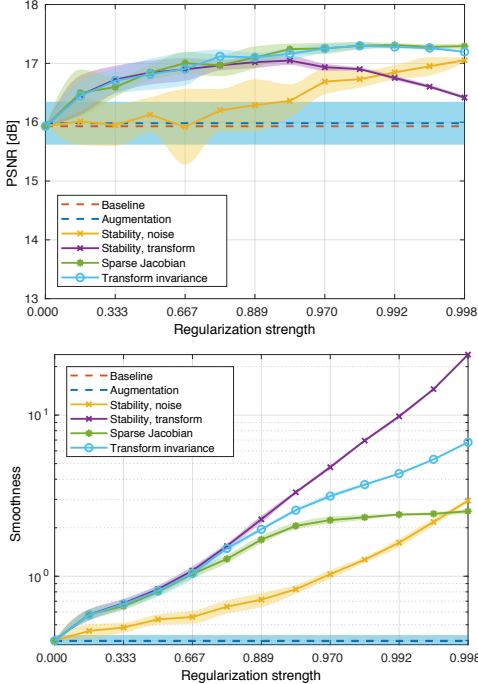


Figure 4. Colorization performance, evaluated using PSNR (top) and smoothness (bottom). The datapoints are estimated as the mean over 10 individual trainings, and the shaded regions illustrate the standard deviation. The Baseline condition is hidden under the dashed-blue Augmentation line in the bottom plot.

In summary, the experiments give us a good indication of the large improvements in temporal stability for widely different applications that can be achieved from explicitly regularizing for this objective. However, differentiating between the two proposed formulations is more difficult, and could potentially be application dependent. Finally, we have large improvements in PSNR for our scenarios with limited training data, indicating that the proposed regularization strategies can improve generalization performance.

5. Example applications

In this section we demonstrate that the proposed regularization terms improve the results not only for the limited scenarios in Section 4, but also for large-scale problems trained on large amounts of data.

5.1. Colorization

For this application, we start from the architecture used by Iizuka *et al.* [17]. However, we skip the global features network and replace the encoder part of the CNN with the convolutional layers from VGG16 [30]. In this way, we can initialize the encoder using pre-trained weights for classification. This setup resulted in a significant improvement in the performance as compared to using the original encoder design. In total, the network is specified from

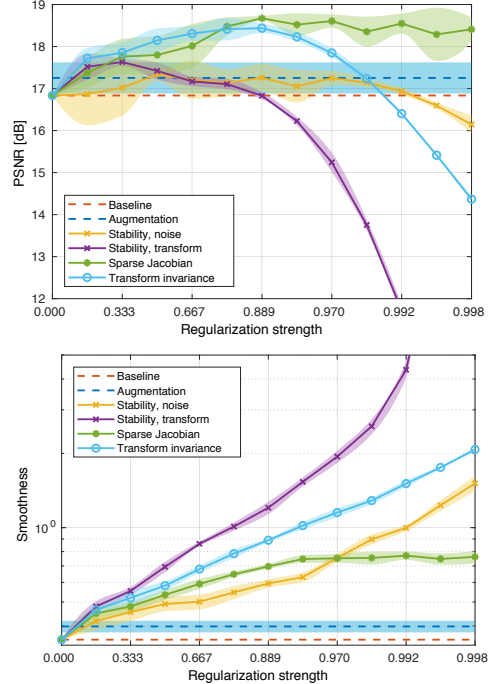


Figure 5. HDR reconstruction performance, evaluated using PSNR (top) and smoothness (bottom). The notation is the same as in Figure 4.

$\sim 19\text{M}$ weights. We train it on the Places dataset [40], and use weights pre-trained for classification on the same dataset. We remove from training around 5% of the images that showed the least color saturation. The CNN was then trained for ~ 15 epochs on the remaining $\sim 2.1\text{M}$ images, at a resolution of 224×224 pixels.

We fine-tune the colorization CNN using two proposed regularization strategies. The effect of the fine-tuning is measured in terms of PSNR and the smoothness measure, see Table 3. The table also includes a fine-tuning without regularization for comparison, and processing the baseline output using the method by Lai *et al.* [24]. Overall, the regularizations offer slight improvements in PSNR (around 0.3–0.5dB) while increasing smoothness substantially. This also goes for comparison to the flow-based post-processing network by Lai *et al.* The transform invariance formulation with $\alpha = 0.95$ gives the best smoothness, and with a PSNR close to the other regularization settings.

Examples of the impact of the regularization techniques are demonstrated in Figure 6. The baseline CNN can exhibit large frame to frame differences, which is much less likely after performing the regularized training. Also, there is an overall increase in the reconstruction performance — whereas the baseline has a tendency to fail in many of the frames, this is less likely to happen when accounting for the differences between frames in the loss evaluation. For example, in the bottom example of Figure 6 the pixel values

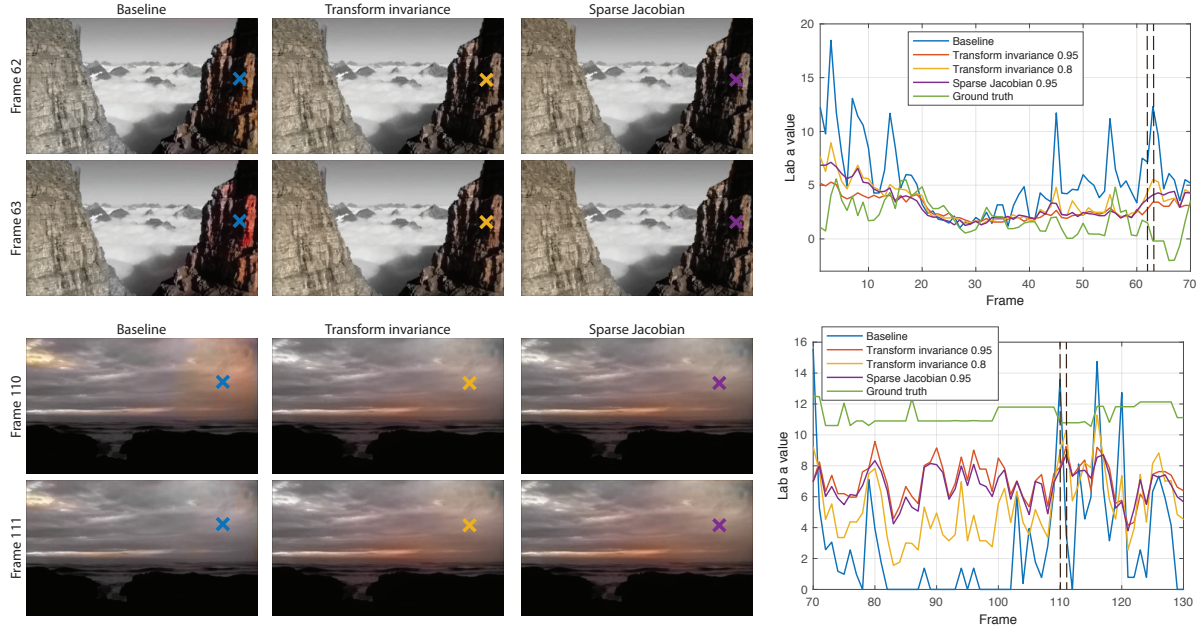


Figure 6. Two video colorization examples from the YouTube-8M dataset [1]. On the left there are two consecutive frames displayed for each sequence, comparing baseline to the two video regularization techniques. The plots on the right show the pixel values of the locations marked in the frames, over a larger range of frames. The values are taken from the a channel of the Lab color space. The vertical dashed lines indicate where the displayed frames are. The transform invariance regularization has been performed at two strengths, α .

Table 3. Performance after fine-tuning of the colorization CNN. The measures have been evaluated and averaged over the a and b channels in the Lab color encoding. Test data are 23 sequences from the YouTube-8M dataset [1].

Training strategy	PSNR	Smoothness
Baseline	18.5805	0.7243
Fine-tuning (no regularization)	18.4315	0.6348
Transform invariance, $\alpha = 0.95$	18.8880	2.8934
Transform invariance, $\alpha = 0.8$	18.9437	1.9074
Sparse Jacobian, $\alpha = 0.95$	18.8852	2.5079
Blind video consistency [24]	18.6086	1.0287

plotted for the baseline CNN are in many cases close to 0, and occasionally spike to high values. This problem is alleviated by the regularization, resulting in both overall better reconstruction and smoother changes between frames.

5.2. HDR reconstruction

In this application we employ the CNN that was used by Eilertsen *et al.* [8] and initialize it with the trained weights provided by the authors. The CNN contains in total $\sim 29M$ weights. We perform fine-tuning on a gathered set of $\sim 2.7K$ HDR images from different online resources, which are used to create a dataset of $\sim 125K$ 320×320 pixel training images by means of random cropping and augmentation.

The fine-tuning result is measured by PSNR and smoothness in Table 4, demonstrating a significant increase in smoothness at the cost of a small decrease in PSNR. Com-

Table 4. Performance after fine-tuning of the HDR reconstruction CNN. The measures have been evaluated and averaged over the saturated pixels only. Test data are 10 HDR video sequences from two different sources [10, 3]. The blind video consistency has been performed on gamma corrected HDR images.

Training strategy	PSNR	Smoothness
Baseline	25.5131	5.9951
Fine-tuning (no regularization)	25.9865	5.8538
Transform invariance, $\alpha = 0.95$	24.1678	10.6435
Transform invariance, $\alpha = 0.8$	25.4374	8.0798
Sparse Jacobian, $\alpha = 0.95$	24.7287	7.3048
Blind video consistency [24]	25.3702	7.2035

pared to the colorization application, regularization of the HDR reconstruction should be selected at a slightly lower α in order to not degrade reconstruction performance. The transform invariance formulation at $\alpha = 0.8$ only reduces the reconstruction performance by $\sim 0.1dB$ while providing better smoothness than the sparse Jacobian formulation. This setting also shows better performance as compared to the blind video consistency method by Lai *et al.* [24], both in terms of PSNR and smoothness.

Figure 7 shows an example of the difference in performance for one HDR video sequence. In contrast to the colorization application it is difficult to clearly see the differences between consecutive frames in a side-by-side comparison. However, in the video material the differences in the temporal robustness around saturated image regions are evident. This can be seen in the pixel plots in Figure 7,

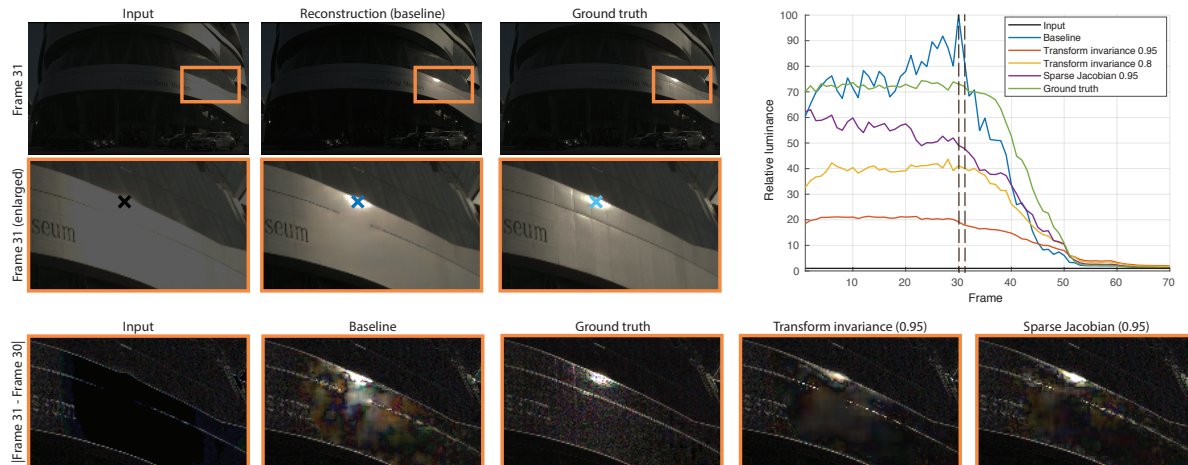


Figure 7. HDR video reconstruction example from the HdM-HDR dataset [10]. On the top left we have an example of the reconstruction compared to input and ground truth, displayed at a reduced exposure (-3 stops) in order to demonstrate the differences in saturated pixels. The bottom row displays the absolute difference between two consecutive frames, for an enlarged region of the image and for different training strategies. The plot on the right shows the HDR luminance values of the pixels marked in the frames, over a larger range of frames. The vertical dashed lines indicate the frames used for the difference evaluations.

where the regularized results are more stable over time for the selected saturated pixel. The figure also shows the absolute difference between two frames for an enlarged image region, highlighting the improvements achieved from regularization when comparing to the ground truth difference.

6. Limitations and future work

Striking the right balance between reconstruction performance and smoothness is still an open problem. A small regularization strength leaves video with temporal artifacts, whereas a too large strength may risk degrading the reconstruction performance. Also, the tendency to impair reconstruction performance with strong regularization could be in some respect analogous to the reduced sharpness when L2 norm is used as the loss function in reconstruction problems (denoising, deconvolution, etc.). We do not address this problem in our current work, but believe that this can be alleviated by exploring other regularization loss functions, such as L1, perceptual loss (for color), or by means of a GAN architecture. The method could also benefit from combining the reconstruction error and smoothness, for a better measure of perceived quality. Moreover, although the transform invariance formulation in some situations can give a better trade-off between PSNR and smoothness, the sparse Jacobian formulation tends to be more robust to large regularization strengths, see *e.g.* Figure 5.

Our approach optimizes towards short-term temporal stability without a guarantee for the long-term temporal consistency. For example, even if colors are consistent in consecutive frames for the colorization application, they may change inconsistently over a longer sequence of frames. An interesting area for future work is therefore to

investigate how long-term temporal coherence can be enforced upon the solution. Finally, it would also be interesting to explore regularization of more complicated loss functions, such as those based on GANs [11], *e.g.* the pix2pix [18] CNN or cycle-GANs [41].

7. Conclusion

This paper explored how regularization using models of the problem dynamics can be used to improve the temporal stability of pixel-to-pixel CNNs in video reconstruction tasks. We proposed two formulations for temporal regularization, which can be used when training a network from scratch, or for fine-tuning pre-trained networks. The strategy is light-weight, it can be used without architectural modifications of the CNN, and it does not require video or motion information for training. It avoids the costly and often inaccurate estimation of optical flow, inherent to previous stabilization methods. Our experiments showed that the proposed approach leads to substantial improvements in temporal stability while maintaining the reconstruction performance. Moreover, for some situations, and especially when training data is limited, the regularization can also improve the reconstruction performance of the CNN, and to a much larger extent than what is possible with traditional augmentation techniques.

Acknowledgments This project was supported by the Wallenberg Autonomous Systems and Software Program (WASP), the strategic research environment ELLIIT, and has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 725253–EyeCode).

References

- [1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 7
- [2] A. Azulay and Y. Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 1, 2
- [3] A. Banitalebi-Dehkordi, M. Azimi, M. T. Pourazad, and P. Nasiopoulos. Compression of high dynamic range video using the HEVC and H. 264/AVC standards. In *Proceedings of International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine 2014)*, pages 8–12. IEEE, 2014. 7
- [4] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister. Blind video temporal consistency. *ACM Transactions on Graphics*, 34(6):196:1–196:9, 2015. 2, 3
- [5] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017. 2
- [6] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua. Coherent online video style transfer. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2017)*, 2017. 1
- [7] X. Dong, B. Bonev, Y. Zhu, and A. L. Yuille. Region-based temporally consistent video post-processing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 2015. 2
- [8] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017. 4, 7
- [9] L. Engstrom, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017. 1, 2
- [10] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Proceedings of SPIE, Digital Photography X*, volume 9023, 2014. 7, 8
- [11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of International Conference on Neural Information Processing Systems (NIPS 2014)*, pages 2672–2680, 2014. 8
- [12] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2
- [13] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei. Characterizing and improving stability in neural style transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 4067–4076, 2017. 1, 2
- [14] S. Hochreiter and J. Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems (NIPS 1995)*, pages 529–536, 1995. 2
- [15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2
- [16] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 235–243, 2015. 2
- [17] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics*, 35(4):110:1–110:11, 2016. 4, 6
- [18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017. 8
- [19] J. S. Jean and J. Wang. Weight smoothing to improve network generalization. *IEEE Transactions on neural networks*, 5(5):752–763, 1994. 2
- [20] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. In *Proceedings of International Conference on Machine Learning (ICML 2017)*, volume 70, pages 1771–1779, 2017. 2
- [21] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems (NIPS 1992)*, pages 950–957, 1992. 2
- [22] J. Kukačka, V. Golkov, and D. Cremers. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017. 2
- [23] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 2
- [24] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision (ECCV 2018)*, 2018. 1, 2, 6, 7
- [25] J. Laird, M. Rosen, J. Pelz, E. Montag, and S. Daly. Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli. In *Human Vision and Electronic Imaging*, volume 6057, page 605705, 2006. 5
- [26] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics*, 31(4):34:1–34:8, 2012. 2, 3
- [27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2015)*, 2015. 4
- [28] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of International Conference on Machine Learning (ICML 2011)*, pages 833–840, 2011. 2
- [29] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. In *German Conference on Pattern Recognition*, pages 26–36. Springer, 2016. 1, 2

- [30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [31] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017. 2
- [32] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2
- [33] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, pages 22–29, 2017. 2
- [34] X. Wei, J. Zhu, S. Feng, and H. Su. Video-to-video translation with global temporal consistency. In *Proceedings of ACM International Conference on Multimedia (MM 2018)*, pages 18–25, 2018. 2
- [35] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR 2011)*, pages 529–534. IEEE, 2011. 4, 5
- [36] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 802–810, 2015. 2
- [37] C.-H. Yao, C.-Y. Chang, and S.-Y. Chien. Occlusion-aware video temporal consistency. In *Proceedings of ACM International Conference on Multimedia (MM 2017)*, pages 777–785. ACM, 2017. 2
- [38] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR 2018)*, 2018. 2
- [39] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 4480–4488, 2016. 1, 2, 3
- [40] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using Places database. In *Proceedings of International Conference on Neural Information Processing Systems (NIPS 2014)*, pages 487–495, 2014. 6
- [41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, 2017. 8

Supplementary material

S1: Transformations

The image perturbations $T(\cdot)$ are performed by means of a linear transformation of the pixel indices i and j ,

$$\begin{bmatrix} i' \\ j' \end{bmatrix} = \begin{bmatrix} T_{1,1} & T_{1,2} & T_{1,3} \\ T_{2,1} & T_{2,2} & T_{2,3} \end{bmatrix} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix}, \quad (11)$$

where (i', j') are the transformed indices, such that $T(x)_{i,j} = x_{i',j'}$. The transformation matrix elements are defined as follows:

$$\begin{aligned} T_{1,1} &= \frac{z \cos a}{\cos(h_x)}, \\ T_{1,2} &= \frac{z \sin(a)}{\cos(h_x)}, \\ T_{1,3} &= \frac{s_x \cos(h_x) - s_x z \cos a}{2 \cos(h_x)} \\ &\quad + \frac{2t_x z \cos a - s_y z \sin a + 2t_y z \sin a}{2 \cos(h_x)}, \\ T_{2,1} &= \frac{z \sin b}{\cos(h_y)}, \\ T_{2,2} &= \frac{z \cos b}{\cos(h_y)}, \\ T_{2,3} &= \frac{s_y \cos(h_y) - s_y z \cos b}{2 \cos(h_y)} \\ &\quad + \frac{2t_y z \cos b - s_x z \sin b + 2t_x z \sin b}{2 \cos(h_y)}. \end{aligned} \quad (12)$$

Here, we have $a = h_x - r$ and $b = h_y + r$, and (s_x, s_y) is the image size. The formulation assumes that the image origin is in the corner of the image, thus incorporating a translation of the image center to the origin before performing the image transformations and translating back afterwards. $(t_x, t_y), r, z$, and (h_x, h_y) are translation offset, rotation angle, zoom factor, and shearing angles, respectively. All the transformation parameters are drawn from uniform distributions, in a selected range of values as specified in Table 1.

S2: Implementation

The transformations and the loss formulations in Section 3 can be implemented with little modification of an existing CNN training script. An example implementation is provided in Listing 1, using Tensorflow. It evaluates the CNN on the input image x and the transformed image $T(x)$ by means of a weight-sharing network.

S3: Training time

The regularized losses take approximately 2 times longer to evaluate as compared to training with only the loss

```

1 import numpy as np
2 import tensorflow as tf
3
4 # Other initialization stuff
5 ...
6
7 # The ground truth (bs is batch size)
8 y = tf.placeholder(tf.float32, [bs, sx, sy])
9
10 # Random transformations
11 ang = np.deg2rad(1.0)
12 tx = tf.random_uniform(shape=[bs,1], minval=-2.0, maxval=2.0, dtype=tf.float32)
13 ty = tf.random_uniform(shape=[bs,1], minval=-2.0, maxval=2.0, dtype=tf.float32)
14 r = tf.random_uniform(shape=[bs,1], minval=-ang, maxval=ang, dtype=tf.float32)
15 z = tf.random_uniform(shape=[bs,1], minval=0.97, maxval=1.03, dtype=tf.float32)
16 hx = tf.random_uniform(shape=[bs,1], minval=-ang, maxval=ang, dtype=tf.float32)
17 hy = tf.random_uniform(shape=[bs,1], minval=-ang, maxval=ang, dtype=tf.float32)
18
19 # Transformation matrix
20 a = hx - r
21 b = hy + r
22 T1 = tf.divide(z*tf.cos(a), tf.cos(hx))
23 T2 = tf.divide(z*tf.sin(a), tf.cos(hx))
24 T3 = tf.divide(sx*tf.cos(hx)-sx*z*tf.cos(a)+2*tx*z*tf.cos(a)-sy*z*tf.sin(a)+2*
    ty*z*tf.sin(a), 2*tf.cos(hx))
25 T4 = tf.divide(z*tf.sin(b), tf.cos(hy))
26 T5 = tf.divide(z*tf.cos(b), tf.cos(hy))
27 T6 = tf.divide(sy*tf.cos(hy)-sy*z*tf.cos(b)+2*ty*z*tf.cos(b)-sx*z*tf.sin(b)+2*
    tx*z*tf.sin(b), 2*tf.cos(hy))
28 T7 = tf.zeros([bs,2], 'float32')
29 T = tf.concat([T1, T2, T3, T4, T5, T6, T7], 1)
30
31 # Perform transformation
32 Ty = tf.contrib.image.transform(y, T, interpolation='BILINEAR')
33
34 # Prepare input x from ground truth y
35 x = prepare_data(y)
36 Tx = prepare_data(Ty)
37
38 # Model
39 with tf.variable_scope("siamese") as scope:
40     fx = cnn_model(x)
41
42 # Weight-sharing
43 scope.reuse_variables()
44 fTx = cnn_model(Tx)
45
46 # Transformation on prediction
47 fTx = tf.contrib.image.transform(fx, T, interpolation='BILINEAR')
48
49 # Reconstruction loss
50 loss = (1.0-alpha)*tf.reduce_mean(tf.square(fx-y))
51
52 # Regularization loss
53 if stability:
54     loss += alpha*tf.reduce_mean(tf.square(fx-fTx))
55 elif transform_invariance:
56     loss += alpha*tf.reduce_mean(tf.square(fTx-fTx))
57 elif sparse_jacobian:
58     loss += alpha*tf.reduce_mean(tf.square((fTx-fx)-(Ty-y)))
59
60 # Train model using the regularized loss
61 ...

```

Listing 1. Tensorflow example for formulating regularized loss.

$\mathcal{L}_{rec} = \|f(x) - y\|_2$. For the HDR reconstruction application, the Sparse Jacobian formulation took on average 1.92 times longer, whereas the transform invariance took 1.99 times longer. The latter is slightly slower since it requires running the transformation $T(f(x))$ on the reconstructed image $f(x)$.

S4: Experimental setup

The two different applications used for the experiments are evaluated in the following way:

- In the total loss in Equation 1, we use three different formulations of \mathcal{L}_{reg} : stability (2), transform invariance (4), and sparse Jacobian (8) regularization.
- The regularization strength is sampled at 12 locations, $\alpha_i = \frac{l_i}{l_i+1}$, $i = 1, \dots, 12$, where $l_i = 2^{i-3}$. This means that the relative regularization strength, or ratio $\frac{\mathcal{L}_{reg}}{\mathcal{L}_{rec}}$,

will double for each point.

- For the perturbed sample $T(x)$, we use the geometric transformation specifying the warping from coordinate transformations according to Equation 11. For the stability regularization we also add one setting with noise perturbations, $T(x) = x + \Delta x$, where $\Delta x \sim \mathcal{N}(0, \sigma)$, and σ is randomly selected for each image, $\sigma \sim \mathcal{U}(0.01, 0.04)$.
- We complement with a training run using $T(x)$ for specifying naïve augmentation, increasing the training dataset size from N to $2N$.
- For each combination of the above, we run 10 individual trainings, in order to estimate a proper mean and standard deviation of each datapoint.

In total, the combinations and repeated runs means that for each of the two applications we perform 500 optimization runs.