Supplemental Material for End-to-End Learning of Video Super-Resolution with Motion Compensation

Osama Makansi, Eddy Ilg, and Thomas Brox

Department of Computer Science, University of Freiburg

1 Computation of PSNR values

For all our evaluation results, the reported PSNR values are computed using only the Y channel of the estimated YCbCr image. In case of RGB images, we first convert to YCbCr color space and then compute on the Y channel. For all experiments using the SRCNN [1] or VSR [5] architecture, we follow [5] and for technical reasons crop away 12 pixels of the boundary from the estimated high-resolution images before computing PSNR values.

2 Displacement magnitudes

We have noted that improvements using motion compensation are generally smaller on Myanmar than on Videoset4. In Table 1, we compute the average motion magnitudes of the datasets and note that the displacements are also generally smaller in the Myanmar validation set.

Dataset	Avg. Mag.
Myanmar training	1.50px
Myanmar validation	0.43px
Videoset4	1.29px

Table 1: Average motion magnitudes computed using FlowNet2 [4]. The numbers show that the Myanmar validation set has the smallest displacements.

3 Video super-resolution with patch-based training

Using patch-based training, we retrain and evaluate SRCNN [1] and VSR [5] using different kind of motion compensations. However, the resulting PSNR scores in Table 2 are all similar and we conclude that motion compensation on Myanmar has no effect. We also evaluate on Videoset4 (Table 3) and there see a small increment of 0.18 for FlowNet2 [4] and FlowNet2-SD [4].

Arch.	Tested on / Trained on	only center	no warp	Drulea [2]	FlowNet2-SD [4]	FlowNet2 [4]
SRCNN	only center	**31.62**	-	-	-	-
VSR	only center	**31.76**	-	-	-	-
	no warp	31.80	**31.83**	-	-	-
	Drulea [2]	31.77	31.74	**31.81**	-	-
	FlowNet2-SD [4]	31.75	31.75	31.79	**31.77**	-
	FlowNet2 [4]	31.76	31.76	31.80	31.78	**31.79**

Table 2: PSNR scores for patch-based video superresolution on the Myanmar validation set. We retrained the architecture of [5] using only the center frames (replicated three times), original images, and motion compensated frames. One can observe that all scores are nearly the same and motion compensation on the Myanmar validation set has no effect over providing original images or even only the center image.

Motion compensation during training and testing	Videoset4 PSNR
only center	24.60
no warp	24.59
Drulea [2]	24.69
FlowNet2-SD [4]	24.77
FlowNet2 [4]	24.77

Table 3: Evaluation of the different retrained VSR models from Table 2 on Videoset4. Motion compensation shows a small performance improvement.

Setting	Patch-based	Image-based
Learning rate	$1e-05$	$1e-05$
Learning rate policy	fixed	multistep†
Momentum	0.9	0.9
Weight decay	0.0005	0.0004
Batch size	240	2
Input resolution	36×36	960×540
Image pixels in batch	311k	1M
Training iterations	200k	300k
Training time	7 hours	32 hours

Table 4: Different settings of patch- and image-based traing. Settings are very similar, except that the number of pixels and trainig time in image-based training are larger. Note that the number of pixels is also further boosted much more by sliding the convolutions from the VSR architecture over the entire images with a stride of one. †multiplied by 0.5 every 50k iterations.

4 Video super-resolution with image-based training

We perform the same set of experiments as in the last section for image-based training. Comparing Table 2 to Table 5, we find that PSNRs are generally around 1 point higher. We also provide all the training settings in Table 4. Image-based training in general processes more training data and sees a lot of similar data during training by sliding the convolutions over an entire image. Motion compensation on Myanmar (Table 5) still seams to have little effect, while motion compensation on Videoset4 does show better PSNR values (Table 6).

Arch.	Trained on \ Tested on	only center	no warp	Drulea [2]	FlowNet2-SD [4]	FlowNet2 [4]
VSR	only center	**32.41**	-	-	-	-
	no warp	32.38	**32.55**	-	-	-
	Drulea [2]	32.37	32.26	**32.60**	-	-
	FlowNet2-SD [4]	32.35	32.37	32.58	**32.62**	-
	FlowNet2 [4]	32.37	32.36	32.61	32.61	**32.63**

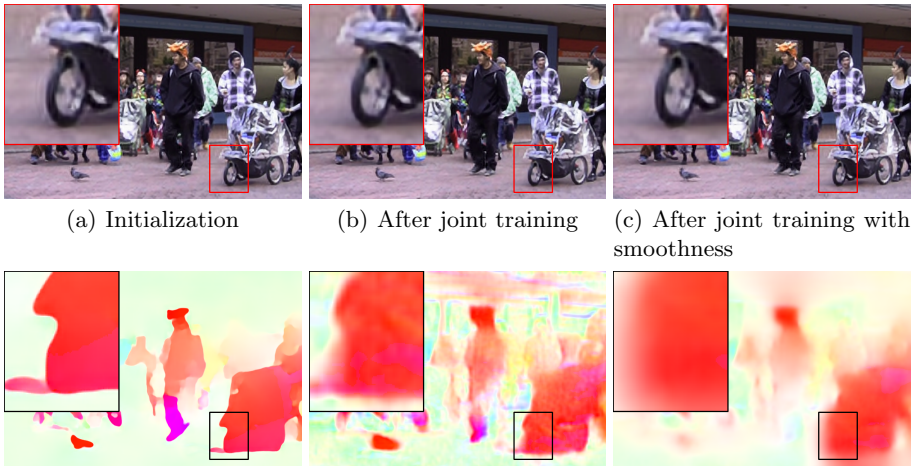
Table 5: PSNR scores from Myanmar validation (ours). We now train the architecture of [5] by applying it as a convolution over the complete images. We again evaluate using only the center frame, original images and differently motion compensated frames. One can observe that scores are significantly better compared to the patch-based training, but motion compensation on the Myanmar validation set still has negligible effect compared to training on original frames.

Training input	PSNR
only center	24.66
no warp	24.79
Drulea [2]	24.91
FlowNet2-SD [4]	25.12
FlowNet2 [4]	25.13

Table 6: Evaluation of the different image-based models on Videoset4. Motion compensation in this case also shows a performance improvement.

5 Joint Training

Since the FlowNet2-SD [4] is completely trainable, we can refine the optical flow for the task of video super-resolution by training the whole network end-to-end with the super-resolution loss. This potentially allows the optical flow estimation to focus on aspects that are most relevant for the super-resolution task. As an initialization we took the VSR network trained on FlowNet2-SD [4] from the last section and used the same settings from Table 4, but now cropped the images to a resolution of 256×256 to enable a batch size of 8. We then trained for 100k more iterations. The result is given in Table 7 and Figures 1(b) and 1(e). We cannot see the flow itself improve, but we see a small improvement in the PSNR value on Videoset4 and from the images one can observe that the ringing artifacts disappear.



(a) Initialization (b) After joint training (c) After joint training with smoothness

(d) Initialization (e) After joint training (f) After joint training with smoothness

Fig. 1: Example super-resolved image after training FlowNet2-SD [4] with VSR (a and d) jointly (b and e) and including a smoothness constraint (c and f).

Test set	After initialization	After joint training	After joint training with smoothness
Myanmar validation	32.62	32.63	32.61
VideoSet4	25.12	25.21	25.19

Table 7: Evaluation of refining FlowNet2-SD [4] on the super-resolution task.

In Figure 1(e), one can observe that many image details become flow artifacts. This is due to nature of the gradient through the warping operation; it corrects the flow vector to the best directly neighboring pixel, which is in most cases a local minimum. Following [3], we add a regularization loss that penalizes deviations from smoothness in the optical flow field, weighted with an image edge-aware term:

$$\mathcal{L}_R = \sum_{i,j} \left(e^{-||\partial_x I_{i,j}||} \left(|\partial_x u| + |\partial_x v|\right) + e^{-||\partial_y I_{i,j}||} \left(|\partial_y u| + |\partial_y v|\right) \right) \tag{1}$$

where I is the first image. i, j is a pixel location and u, v are the x, y components of the flow vector. The results of training with this additional smoothness term are given in Table 7 and Figures 1(c) and 1(f). The flow shows less artifacts than Figure 1(e), but compared to Figure 1(b) some very slight ringing artifacts still remain.

6 Evaluating Architectures and Datasets

In first part of the paper, the architecture from Dong et al. [1] adapted to video super-resolution by Kappeler et al. [5] and the Myanmar training dataset were used. Here, we investigate the effect of architectures and training datasets. We extended the Myanmar training set by more high-resolution videos that we downloaded from Youtube. The resulting dataset has 162k frames of resolution 960×540 and we named it MYT. We evaluate and compare the SRCNN [1], the FlowNet2-SD [4] (here used for super-resolution, not flow) and the encoder-/decoder part of the architecture from Tao et al. [6] (SPMC-ED) for single image super-resolution on the old and new datasets. The results are given in Table 8.

	SRCNN [1] trained on Myanmar training (ours)	SRCNN [1] trained on MYT	FlowNet2-SD [4] trained on MYT	SPMC-ED [6] trained on MYT
Myanmar validation (ours)	32.42	31.98	31.47	32.63
Videoset4	24.63	24.70	24.93	25.07
Number of parameters	57K	57K	14M	491K

Table 8: PSNR values for different architectures and training datasets tested for single-image super-resolution.

One can observe that SRCNN [1] tends to overfit on the Myanmar dataset. The much deeper FlowNet2-SD [4] architecture performs worse on Myanmar, but can generalize better to Videoset4. The size of SPMC-ED [6] is between the former two and we observe that it performs best on Myanmar and also for generalization to Videoset4. It clearly gives better results than SRCNN [1] and for this reason we also use it for the final network in the paper.

References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38(2), 295–307 (Feb 2016)
2. Drulea, M., Nedevschi, S.: Total variation regularization of local-global optical flow. In: IEEE Conference on Intelligent Transportation Systems (ITSC). pp. 318–323 (Oct 2011)
3. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. CoRR abs/1609.03677 (2016), http://arxiv.org/abs/1609.03677
4. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
5. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging (TCI) 2(2), 109–122 (June 2016)
6. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. CoRR abs/1704.02738 (2017), http://arxiv.org/abs/1704.02738