doi:10.11887/j.cn.201605022 http://journal.nudt.edu.cn

用于视频图像帧间运动补偿的深度卷积神经网络*

龙古灿[1,2],张小虎[1,2],于起峰[1,2]
(1.国防科技大学 航天科学与工程学院,湖南 长沙 410073;
2.国防科技大学 湖南省图像测量与视觉导航重点实验室,湖南 长沙 410073)

摘 要:为探索深度学习理论在视频图像帧间运动补偿问题中的应用,提出一种用于视频图像帧间运动补偿的深度卷积神经网络。该网络由卷积模块和反卷积模块构成,可以处理不同分辨率输入图像并具备保持较完整图像细节的能力。利用具有时序一致性的视频图像序列构造训练样本,采用随机梯度下降法对设计的深度卷积神经网络进行训练。视觉效果和数值评估实验表明,训练得到的网络较传统方法能更有效地进行视频图像帧间运动补偿。

关键词:深度学习;卷积神经网络;时序一致性;运动补偿帧插值
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2016)05-143-06

Deep convolutional neural network for motion compensated frame interpolation

LONG Gucan[1,2], *ZHANG Xiaohu*[1,2], *YU Qifeng*[1,2]
(1. College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China;
2. Hunan Key Laboratory of Videometrics and Vision Navigation, National University of Defense Technology, Changsha 410073, China)

Abstract: In order to explore the application of deep learning theory in the problem of motion compensated frame interpolation, a DCNN (deep convolutional neural network) built with convolutional blocks and deconvolutional blocks was proposed. The proposed DCNN is capable of processing input images with different resolutions and preserving fine-grained image details. The temporal coherent image sequences were used to construct the training sample and the stochastic gradient descent method was adopted to train the designed DCNN. Qualitative and quantitative experiments show that the trained DCNN obtains better interpolated images than the traditional approach in two testing images sequences.

Key words: deep learning; convolutional neural network; temporal coherence; motion compensated frame interpolation

视频图像帧间运动补偿又称运动补偿帧插值[1-2](motion compensated frame interpolation),指利用视频序列中连续两帧图像进行运动插值以合成中间帧图像的过程。作为图像与视频处理领域的经典问题之一,其在视频帧率提升、慢速视频制作以及虚拟视图合成等场合具有广泛应用。目前常用的视频图像帧间运动补偿方法首先基于光流场估计算法对输入图像对进行密集匹配,然后利用获得的密集匹配信息逐像素对输入图像进行内插以合成中间帧图像。由于光流场估计本身是一个病态问题,尤其在图像纹理较弱或存在遮挡等情况下易效果不佳,现有方法在实际应用中常面临困难。

近年来,基于深度学习的方法得到了广泛关注,在如目标分类[3-4]、人脸识别[5]等众多计算机视觉问题中取得了显著优于传统方法的效果。注意到该类方法成功的关键在于利用海量训练样本对合适的深度神经网络进行训练。对于本文关注的视频图像帧间运动补偿问题,由于可以利用现有的海量视频数据构造训练样本,适合采用基于深度学习的相关方法进行求解,但文献中尚未见相关报道。本文对深度学习在视频图像帧间运动补偿问题中的应用进行探索,利用具有时序一致性的视频图像序列构造训练样本,对设计的深度卷积神经网络进行训练以实现运动插值图像的合成,在两组包含大量弱纹理区域的测试图像序列中取得了优于传统方法的效果。

以下将首先介绍本文设计的用于视频图像帧

* **收稿日期**:2016-04-27
基金项目:国家重点基础研究发展计划资助项目(2013CB733100)
作者简介:龙古灿(1988—),男,湖南浏阳人,博士研究生,E-mail:longgucan@163.com;
张小虎(通信作者),男,研究员,博士,博士生导师,E-mail:zxh1302@hotmail.com

间运动补偿的卷积神经网络结构;然后从训练数据、目标函数设计以及训练过程等方面介绍对设计的神经网络进行训练的情况;最后将基于本文深度卷积神经网络的方法与传统采用逐像素插值策略的运动补偿帧插值方法进行比较。

1 深度卷积神经网络设计

相较于设计传统面向视觉目标分类的深度神经网络[3-4],设计用于视频图像帧间运动补偿的卷积神经网络需要针对以下三个问题进行特别考虑:

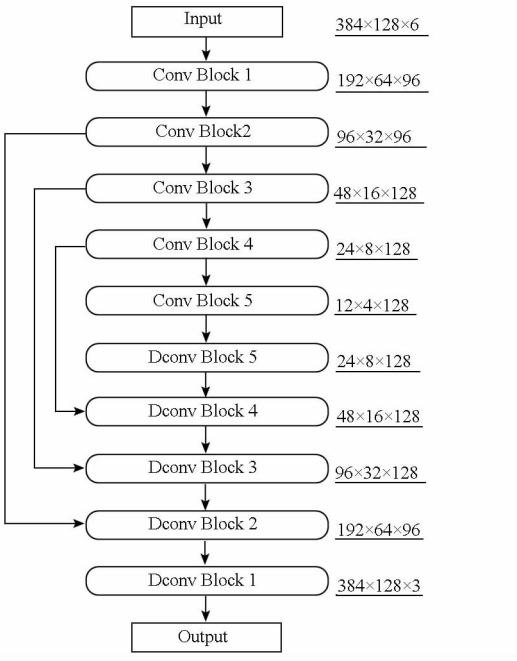
1)不同于用于目标分类的网络输出仅为目标分类概率,面向本文任务的网络输出应为与输入图像分辨率相同的一幅完整图像;

2)考虑到不同应用场合中采用的图像长宽比通常各不相同的实际情况,用于运动补偿帧插值的神经网络应具备处理不同长宽比图像的能力;

3)设计的网络应当具备保持良好图像细节的能力;同时应当考虑在网络层数较多的情况下如何通过优化网络结构以降低梯度弥散现象(gradient vanishing)的影响,使得可以采用随机梯度下降法对其进行有效训练。

本文设计的用于视频图像帧间运动补偿的卷积神经网络总体结构如图 1 所示。构成该网络的基本组件为卷积模块(conv block)和反卷积模块(dconv block),其具体组成如图 2 所示。卷积模块参考标准的卷积神经网络进行设计,由卷积层(convolution layer)和激活函数层(activation layer)交替重复排列三次,并在最后加上池化层(pooling layer)组成。对于卷积层,本文采用 VGG-Net[6]的建议,将感受野(receptive field)尺寸、跨步(stride)和内边距(padding)依次设为 3, 1, 1。激活函数层则采用参数化修正线性单元(Parametric Rectified Linear Unit, PReLU)[7]作为激活函数;池化层的感受野尺寸为 3,跨步为 1。

从图 1 各模块右侧表示其输出数据维数的数字可见,数据每经过一次卷积模块处理,空间尺寸减半。考虑本节开头提出的第一个问题,为使整个网络的输出与输入图像保持同一空间分辨率,本文网络的后半部分采用反卷积模块进行构建。如图 2 所示,每个反卷积模块包含一个卷积转秩层(CONVolution Transpose layer, CONVT)和两个卷积层。其中卷积层的参数与卷积模块中的卷积层参数一致,卷积转秩层的感受野尺寸为 4,跨步为 1,内边距为 1,其具体组成形式请参见文献[8-9]。数据每经过一次反卷积模块处理,空间尺寸增加一倍。如此,输入数据经过本文网络的 5 个卷积模块和 5 个反卷积模块后,空间尺寸保持不变。



注:各模块之间的箭头表示信息流向;各模块右侧数字表示其输出数据的维数。

图 1 网络结构示意图

Fig. 1 Architecture of the designed convolutional neural network

针对本节开头提出的问题三,为使得输出图像保持足够的图像细节,本文借鉴文献[10]的思路进行网络结构设计。如图 1 左侧箭头所示,将卷积模块 2 的输出同时作为卷积模块 3 和反卷积模块 2 的输入。并以同样的方式使用卷积模块 3 和卷积模块 4 的输出。由于图像数据输入卷积神经网络后,随着处理层数的增加,得到的特征描述更抽象,同时图像细节损失越严重。将较浅层的输出作为较深层的额外输入,有利于最终输出结果保持丰富的图像细节。同时,类似于 Highway Network[11]和 Deep Residual Network[12]的设计原理,在网络中引入如图 1 左侧箭头所示的信息流,有助于帮助克服网络训练中的梯度弥散现象,使得随机梯度下降算法取得更好的训练效果。

对于本节开头提出的问题二,即网络应能对不同长宽比的输入图像进行处理的问题,本文设计的网络为一个全卷积神经网络。这种网络的优点一方面在于充分利用了图像数据的空间关联性,网络层数虽然很多,但其中包含大量共享参数,有利于避免过拟合(over fitting)问题并有效降

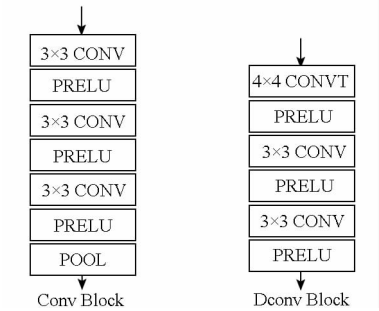


图 2　卷积模块和反卷积模块

Fig. 2　Conv block and dconv block

低运算的时间和空间复杂度。另一方面,全卷积网络的特点使得本文网络能够处理不同长宽比的输入图像。只需输入图像的高和宽方向尺寸分别为 16 的整数倍(即能被 2 整除 5 次),即可采用本文网络进行处理。

2　深度卷积神经网络的训练

2.1　训练数据

如前文所述,可以利用现有的海量自然视频数据对上节设计的卷积神经网络进行训练,而不需要对数据进行费时费力的人工标注。这是因为自然视频图像序列通常具有时序一致性,即短时间内可以认为摄像机和拍摄物体均仅进行速度均匀的运动,十分适合于对本文网络的训练。

本文采用文献[13]提供的 KITTI 原始数据和 Durian 开源电影项目 Sintel 视频构造训练样本。KITTI 视频由安装在汽车上的摄像机采集于德国 Karlsruhe,其作为公开数据集主要面向自动驾驶应用,如光流场计算、Stereo、视觉里程计以及图像分割等。Sintel 视频原本由文献[14]进行改编以构造用于评估光流场估计算法的公开测试集。对于 KITTI 和 Sintel 数据,本文均仅使用其原始视频数据提取训练样本。

KITTI 数据集包含 56 个序列共 16 951 帧图像。在每个序列中,取每连续三帧图像(正序或反序)构成一个训练样本,第一帧和第三帧作为输入图像,第二帧作为运动插值图像的真值。同时采用对各训练样本中包含的图像进行左右翻转、上下翻转以及镜像的方式构造增广样本,共生成 133 921 个训练样本。对于 Sintel 视频,根据时序一致性标准共采集 63 个图像序列,包含 5670 帧图像。采用与 KITTI 数据类似的方式共构造 44 352个训练样本。对于 KITTI 数据,输入图像被降采样为 384 ×128,对于 Sintel 数据,采用的图像大小为 256 ×128。

2.2　目标函数与训练过程

设训练样本集为 $\{I_1^i, I_2^i, I_3^i\}_{i=1}^n$,其中 I_1^i 和 I_3^i 为两张待插值图像,I_2^i 为插值结果真值,n 为样本总数。设上文设计的神经网络输出 $\tilde{I}_2 = I_{\mathrm{CNN}}(w, I_1, I_3)$,其中 w 为神经网络中如卷积滤波器权值等参数,I_1 和 I_3 为输入图像。则训练本文网络等价于求解以下最优化问题:

$$\arg\min_w \sum_{i=1}^{n} d[I_2^i - I_{\mathrm{CNN}}(w, I_1^i, I_3^i)] \tag{1}$$

由于直接采用 *L2* 范数衡量真值图像和合成图像之间的差异(即令 $d(I_2 - \tilde{I}_2) = \|I_2 - \tilde{I}_2\|^2$)容易导致训练得到的神经网络输出丢失过多图像纹理等细节信息[15-16]。本文利用鲁棒光流场计算[17]中常用的 Charbonnier 函数 $\rho = \sqrt{\Delta x^2 + \varepsilon^2}$ 作为距离测度训练本文网络,则式(1)应写作:

$$\arg\min_w \sum_{i=1}^{n} \sum_{x,y,c} \sqrt{[I_2^i(x,y,c) - \tilde{I}_2^i(x,y,c)]^2 + \varepsilon^2} \tag{2}$$

式中 ε 设为 0.1。

本文采用修改后的 Caffe[18] 在安装有两片 NVIDIA Tesla K40c 显卡的高性能工作站上进行实验。在神经网络的训练过程中,首先采用文献[19]的方法对待优化参数 w 进行初始化,然后采用 Adam 方法[20]迭代优化求解式(2)描述的最优化问题。Momentum 设为 0.9;初始学习率(learning rate)设为 0.001,并在优化过程中当观察到目标函数不再下降后手动对学习率进行调整。训练时使用的批(batch)大小为 16。整个训练过程耗时约 5 d。

3　实验与分析

本节对上文设计和训练的用于视频图像帧间运动补偿的卷积神经网络进行实验。采用文献[21]提供的方法作为对插值结果进行评估的基准算法,其中光流场计算部分采用目前在公开数据集[22]上排名靠前且提供源代码的 DeepFlow 方法[23]进行。以下将采用本文提出的基于深度卷积神经网络的方法(Deep Convolutional Neural Network, DCCN);将采用传统逐像素插值策略(并基于 DeepFlow 计算密集图像匹配)的方法简称为 DeepFlow 方法。

参与本节实验评估的数据分为两部分:第一部分为从 2.1 节描述的 KITTI 数据和 Sintel 数据中随机抽取的各一个图像序列(下文简称为

KITTI 序列和 Sintel 序列),注意这两个图像序列在训练神经网络时仅作为用于监控网络训练的验证数据使用;第二部分为 MiddleBurry 数据集[21]中 RubbleWhale 序列以及一组用于医学目的的 DICOM 图像[24]。这两组图像序列(下文简称为 Rubble 序列和 DICOM 序列)主要用于评估本文训练的神经网络的泛化能力(generalization ability)。采用与 2.1 节类似的方法构造用于评估算法性能的图像样本:取序列中每连续三帧构造一个评估样本,其中第一帧和第三帧图像作为输入图像,第二帧图像作为真值图像。

3.1 视觉效果评估

首先从视觉效果方面评估分别由 DCNN 方法和 DeepFlow 方法进行视频图像帧间运动补偿的结果。图 3 展示了两种方法对 KITTI,Sintel 和 Rubble 序列的代表性图像进行运动补偿即生成插值帧的效果,可见两种方法均较好地对输入图像进行了运动插值。注意到 DeepFlow 方法虽然较 DCNN 方法保持了更多的图像细节,但是存在部分错误插值的情况(如图 3 中用矩形框标注的区域)。图 4 展示了两种方法在 DICOM 图像序列上的效果,与图 3 展示的结果类似,DCNN 方法虽然较 DeepFlow 方法在图像细节方面稍有损失,但是不存在如图 4 矩形框中标出的 DeepFlow 方法明显发生错误的情况。

从视觉效果评估结果看,DCNN 方法不但能对与训练数据类似的 KITTI 和 Sintel 图像序列进行正确的运动插值,而且对与训练图像差别较大的图像序列,如 Rubble 和 DICOM 序列,仍能进行正确的运动插值图像生成,这表明训练得到的神经网络具有较好的泛化能力。

3.2 数值评估

采用文献[21]提出的运动插值图像评价准则对 DCNN 和 DeepFlow 方法生成的运动插值图像进行数值评估。插值图像 I 与真值图像 I_{GT} 之间的插值误差(Interpolation Error, IE)由式(3)定义:

$$IE = \sqrt{\frac{1}{3HW}\sum_{(x,y,c)}\left[I(x,y,c) - I_{GT}(x,y,c)\right]^2} \tag{3}$$

图 3 KITTI,Sintel 和 Rubble 序列上的运动补偿帧插值效果

Fig. 3 Example interpolated images in KITTI, Sintel and Rubble sequence

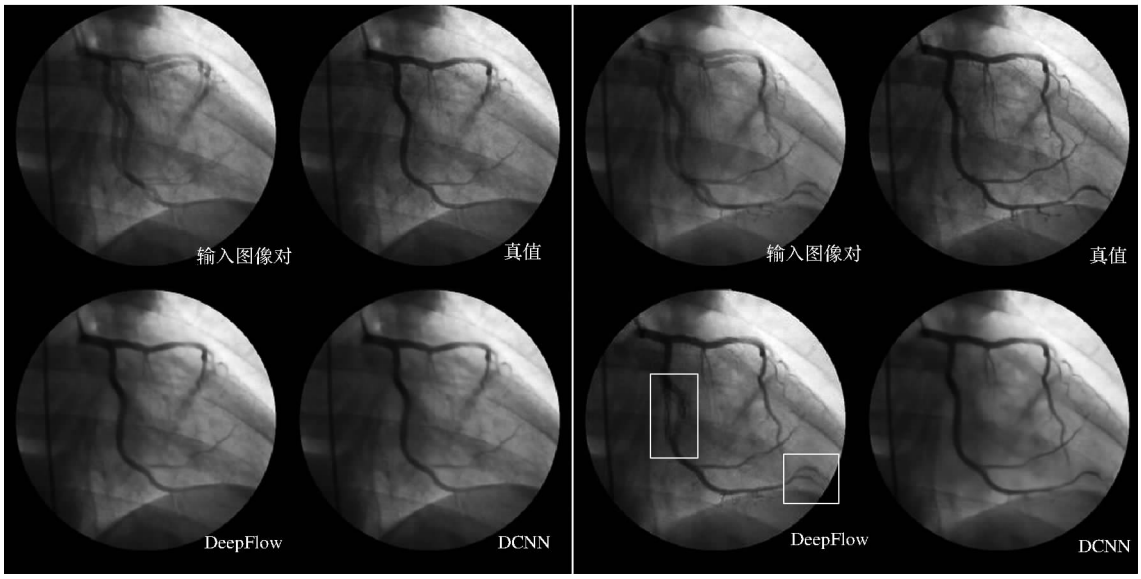


图 4　DICOM 序列上的运动补偿帧插值效果

Fig. 4　Example interpolated images in DICOM sequence

归一化插值误差(Normalized interpolation Error, NE)由式(4)定义:

$$NE = \sqrt{\frac{1}{3HW}\sum_{(x,y,c)}\frac{[I(x,y,c) - I_{GT}(x,y,c)]^2}{\|\nabla I_{GT}(x,y,c)\|^2 + \tau}} \tag{4}$$

其中,∇I_{GT} 表示梯度图像,τ 取 1.0。

对参与实验的四组图像序列分别采用 DCNN 和 DeepFlow 方法进行运动补偿帧插值并计算插值误差 *IE* 和归一化插值误差 *NE*,得到的结果如表 1 所示。

表 1　插值图像的插值误差和归一化插值误差(均值)

Tab. 1　Mean interpolation error and mean normalized interpolation error of the interpolated images

	KITTI	Sintel	DICOM	Rubble
IE(DCNN)	30.99	27.43	6.00	9.31
IE(DeepFlow)	33.30	26.40	6.23	9.00
NE(DCNN)	7.29	9.36	1.65	1.91
NE(DeepFlow)	7.82	9.06	1.70	1.84

从表 1 中可见,在 Sintel 和 Rubble 序列上,以 *IE* 和 *NE* 评价,DCNN 插值效果差于 DeepFlow;对于 Sintel 序列,经过分析,发现导致 DCNN 插值误差增大的原因在于该序列中某些帧之间运动过大;而对于 Rubble 序列,其帧与帧之间运动很小,插值误差较大主要反映了 DCNN 保持图像细节的能力弱于 DeepFlow。

在 KITTI 和 DICOM 序列上,DCNN 插值效果优于 DeepFlow。观察到这两组序列图像中包含大量弱纹理区域,采用 DeepFlow 方法估计帧间图像运动即计算光流场难度较大,而 DCNN 方法直接基于卷积神经网络进行插值图像生成,不需显式计算精确的光流场,取得了较好的效果。同时,在 DICOM 序列上取得的良好数值评估效果进一步验证了 DCNN 方法具备良好的泛化性能。

4　结论

设计并训练了一个深度卷积神经网络,对深度学习方法在视频图像帧间运动补偿问题中的应用进行了探索。实验结果表明,本文深度卷积神经网络具备良好的泛化能力,能有效生成运动插值图像,尤其适用于存在较多弱纹理区域的图像序列。针对实验中发现的问题,后续工作将围绕以下三个方面展开:①深入分析现有卷积神经网络适用于处理存在多大运动量的图像序列;②研究具有更好保持图像细节能力的网络结构;③探索对全尺寸图像进行运动插值的方法。

参考文献(References)

[1] Choi B T, Lee S H, Ko S J. New frame rate up-conversion using bi-directional motion estimation [J]. IEEE Transactions on Consumer Electronics, 2000, 46(3): 603 – 609.

[2] Park D, Jeong J. Motion compensated frame rate up conversion using modified adaptive extended bilateral motion estimation [J]. Journal of Automation and Control Engineering, 2014, 2(4): 371 – 375.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Proceedings of Advances in Neural Information Processing Systems, 2012.

[4] Szegedy C, Liu W, Jia Y, et al. Going deeper with

convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[5] Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[6] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets[C]//Proceedings of BMVC, arXiv preprint arXiv:1405.3531, 2014.

[7] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification [C]//Proceedings of the IEEE International Conference on Computer Vision, 2015.

[8] Vedaldi A, Lenc K. MatConvNet-convolutional neural networks for MATLAB [R]. arXiv preprint arXiv: 1412.4564, 2014.

[9] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]//Proceedings of Computer Vision-ECCV 2014. Springer International Publishing, 2014: 818 – 833.

[10] Dosovitskiy A, Fischery P, Ilg E, et al. Flownet: learning optical flow with convolutional networks [C]//Proceedings of IEEE International Conference on Computer Vision, 2015: 2758 – 2766.

[11] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks[C]//Proceedings of Advances in Neural Information Processing Systems, 2015.

[12] He K M, Zhang X Y, Ren S Y, et al. Deep residual learning for image recognition [R]. arXiv preprint arXiv: 1512.03385, 2015.

[13] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: the KITTI dataset [J]. International Journal of Robotics Research, 2013, 32(11): 1231 – 1237.

[14] Butler D J, Wulff J, Stanley G B, et al. A naturalistic open source movie for optical flow evaluation [C]//Proceedings of European Conference on Computer Vision, Springer-Verlag, 2012: 611 – 625.

[15] Wang X L, Gupta A. Unsupervised learning of visual representations using videos [C]//Proceedings of IEEE International Conference on Computer Vision, IEEE, 2015: 2794 – 2802.

[16] Goroshin R, Mathieu M, LeCun Y. Learning to linearize under uncertainty [R]. arXiv preprint arXiv: 1506.03011, 2015.

[17] Sun D, Roth S, Black M J. A quantitative analysis of current practices in optical flow estimation and the principles behind them [J]. International Journal of Computer Vision, 2014, 106(2): 115 – 137.

[18] Jia Y Q, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding [C]// Proceedings of the ACM International Conference on Multimedia, ACM, 2014.

[19] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [J]. Journal of Machine Learning Research, 2010, 9: 249 – 256.

[20] Kingma D, Ba J. Adam: a method for stochastic optimization[R]. arXiv preprint arXiv:1412.6980, 2014.

[21] Baker S, Scharstein D, Lewis J P, et al. A database and evaluation methodology for optical flow [J]. International Journal of Computer Vision, 2007, 92(1): 1 – 31.

[22] Andreas G, Lenz P, Urtasun R. Are we ready for autonomous driving? the KITTI vision benchmark suite [C]//Proceedings of the IEEE International Conference on Computer Vision, 2012: 3354 – 3361.

[23] Weinzaepfel P, Revaud J, Harchaoui Z, et al. Deepflow: large displacement optical flow with deep matching [C]// Proceedings of the IEEE International Conference on Computer Vision, 2013.

[24] DICOM sample image sets [EB/OL]. [2016 – 04 – 20] http://www.osirix – viewer.com/datasets/.