Violent Interaction Detection in Video Based on Deep Learning

Peipei Zhou\(^1,2,3,4\), Qinghai Ding\(^1,5\), Haibo Luo\(^1,3,4\), and Xinglin Hou\(^1,2,3,4\)

\(^1\) Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016
\(^2\) University of Chinese Academy of Sciences, Beijing, 100049
\(^3\) Key Laboratory of Opto-Electronic Information Processing, CAS, Shenyang 110016
\(^4\) The Key Lab of Image Understanding and Computer Vision, Liaoning Province, Shenyang, 110016
\(^5\) Space star technology co., LTD, Beijing, 100086

Abstract. Violent interaction detection is of vital importance in some video surveillance scenarios like railway stations, prisons or psychiatric centres. Existing vision-based methods are mainly based on hand-crafted features such as statistic features between motion regions, leading to a poor adaptability to another dataset. Enlightened by the development of convolutional networks on common activity recognition, we construct a FightNet to represent the complicated visual violence interaction. In this paper, a new input modality, image acceleration field is proposed to better extract the motion attributes. Firstly, each video is framed as RGB images. Secondly, optical flow field is computed using the consecutive frames and acceleration field is obtained according to the optical flow field. Thirdly, the FightNet is trained with three kinds of input modalities, i.e., RGB images for spatial networks, optical flow images and acceleration images for temporal networks. By fusing results from different inputs, we conclude whether a video tells a violent event or not. To provide researchers a common ground for comparison, we have collected a violent interaction dataset (VID), containing 2314 videos with 1077 fight ones and 1237 no-fight ones. By comparison with other algorithms, experimental results demonstrate that the proposed model for violent interaction detection shows higher accuracy and better robustness.

1. Introduction

Violent interaction (fight) behaviour makes serious effect on social and personal security and the early warning of violent activity could greatly reduce these risks. Currently, there are millions of video surveillance equipment applied in public places, such as streets, squares and railway stations. It is of vital importance to investigate the harmful abnormal contents from vast amounts of surveillance video data. For the consideration of practical application, this work focuses on the challenging task of detecting violent interaction in videos and aims to propose a new method that could automatically detect violent behaviours using computer vision techniques.

Despite the potential significance, enough attention has not been paid on violent interaction detection compared with common action recognition which have made a great progress in recent years. In the field of action recognition and also the violence detection, there are two crucial and complementary aspects: appearances and dynamics. However, extracting such information suffers from a number of complexities, including illumination variations, view point changes, camera motions and so on. Different kinds of descriptors such as GMOF and OHOF\(^1\) or the statistics of motion regions \(^2,3\) are
extracted as the feature of a video. By feeding the features into classifiers (e.g. SVM), a model of violence detection is constructed. However, these features mainly represent the motion information estimated by optical flow images, leading to lack of appearance information. We aim to find a method that could catch both appearances and dynamics information to improve the violence detection performance.

Recently, Convolutional Networks (ConvNets) [4] have witnessed great success in classifying images of objects and video-based action recognition [5,6]. Great modelling capacity of deep ConvNets makes it possible to learn discriminative representation from raw visual data under the condition of large-scale supervised datasets. Violent interaction detection has a lot in common with general activity recognition but it differs from the common behaviours because it is random combination of multi-kinds of actions including kick, beat, push, etc. The irregular internal occlusion is the most remarkable feature and also one of the difficulties to describe the violence behaviours.

Enlightened by the design principle taken by the ConvNets on activity recognition, we make an attempt to apply the deep ConvNets to the detection of video-based violent interaction, which is impeded by two major obstacles. First, long-range temporal structure plays an important role in understanding the action dynamics in videos. We take the state-of-the-art activity network, Temporal Segment Network (TSN) [3] as the basis network for violent interaction detection, which is capable of modeling long-range temporal structure over the whole video. Based on TSN, we construct a network named FightNet fit for violent interaction detection. Second, in practice, for the sake of achieving optimal performance, a large volume of training samples are required to train the deep ConvNets. To solve this problem, we gather videos of violent interaction from datasets in the previous work [7,8,9], named Violent Interaction Dataset (VID). In addition, to prevent over-fitting of the deep ConvNets, we pre-train the model on the UCF101[8] dataset.

2. Related work
A number of researchers like Lam et al. [10] have made great effort to detect fight, violence or aggression. Some previous work [11,12] took the blood or explosion as the cues of violence, but these cues always miss alarm. Recently, Derbas et al. [13] proposed a feature which provides strong multi-modal audio and visual cues by first joining the audio and the visual features and then revealing statistically the joint multi-modal patterns. In [14], multiple kernel learning is applied to maximize the multimodality of videos based on the combined features of visual and audio. However, audio-based methods are always restricted in the real life since the absence of the audio channel.

Violent interaction detection is essentially a task of action recognition. Using computer vision technologies, the key point is to extract features that could represent the sequences during the fight. The features could be divided into two categories: hand-crafted features and learning features.

Hand-crafted features. We define the features designed by human as hand-crafted features. Among them, Space Time Interest Points (STIPs) [15] and Improved Dense Trajectories (iDTs) [16] are widely used in action recognition. Especially for violence detection, Deniz et al. [17] presented a novel method to detect violent sequences which uses extreme acceleration patterns as the main feature by applying the Radon transform to the power spectrum of consecutive frames. Recently, another features extracted from motion blobs between consecutive frames are proposed by [2] to distinguish fight and non-fight sequences. Similar work including [3] and [18] proposed a robust and intuitive approach based on motion statistical features from optical flow images. Zhang et al. [1] extract candidate violence regions using a Gaussian Model of Optical Flow (GMOF) and find out the fight regions through a linear SVM whose input vectors are constructed by Orientation Histogram of Optical Flow (OHOF). This kind of methods based on hand-crafted features are intuitive and efficient for a specific small-scale dataset, but in a big dataset their deficiencies are triggered, leading to low training speed, massive memory usage and inefficient execution.

Learning features. The features learned by deep learning networks are considered as learning features. With the enhancement of computer power brought by GPU and the collection of large-scale training set, action recognition based on deep learning makes a great progress. Designed by Simonyan et al. [19], two-stream ConvNets consist of spatial and temporal nets and exploit ImageNet dataset [20] for pre-training and exploit optical flow to explicitly capture motion information. Tran et al. [21] learned
both appearance and motion features by exploring 3D ConvNets [22] trained on a large-scale supervised dataset. Recently, Varol et al. [23] employed the neural networks with long-range temporal structure (LTC) to learn the video and concluded that LTC-CNN modelled with increased temporal extents improved the accuracy of action recognition. However, these methods are limited by the computational cost so that the video could be no more than 120 frames. The temporal segment network [6] combined a sparse temporal sampling strategy and video-level supervision to learn valid information from the whole action video, achieving the state-the-of-art performance on the two complex datasets of HMDB51 (69.4%) and UCF101 (94.2%).

Although the technology on action recognition is relatively mature, little effort has been spent on the violent interaction detection using deep ConvNets. Referring to the excellent TSN that could efficiently pick up representative action features from videos, we construct a violent interaction detection ConvNet, named FightNet.

3. FightNet for Violent Interaction Detection

As the velocity magnitude of actions in violent interaction is not constant but variable randomly, i.e., there is always a nonzero acceleration field when the violent interaction occurs. Besides the basic attributes of action, we employ the acceleration feature to design FightNet fit for fight detection. As depicted in figure 1, the general flow chart of violent interaction detection is designed to model the whole video.

![Figure 1. General flow chart of violent interaction detection, with three kinds of input modalities, i.e., RGB images for Net_rgb network, optical flow images for Net_flow network and acceleration images for Net_acceleration network.](image)

3.1. FightNet

For a video \( V \), it is divided into \( K \) segments \( \{S_1, S_2, \ldots, S_K\} \) of equal duration and a sequence of short snippets \( \{s_1, s_2, \ldots, s_K\} \) are randomly extracted in the segments. Then the \( K \) snippets are modelled as follows:

\[
M(s_1, s_2, \ldots, s_K) = h\left(g(f(s_1;W), f(s_2;W), \ldots, f(s_K;W))\right)
\]

(1)

Here, each snippet \( s_k \) is randomly sampled from the corresponding segment \( S_k \). \( f(s_i;W) \) is the class score produced by FightNet with the parameter \( W \) operated on the snippet \( s_i \). The fusion function \( g \) combines the outputs from different snippets. Thereon, the prediction function \( h \) provides the whole video the probability of each class, i.e., fight or non-fight. Here the widely used Softmax function is adopted to be the prediction function \( h \). According to the standard cross-entropy loss, the loss function with respect to \( G = g(f(s_1;W), f(s_2;W), \ldots, f(s_K;W)) \) is formed as equation (2):

\[
L(y, G) = -\sum_{i=1}^{C} y_i \left( G_i - \log \sum_{j=1}^{C} \exp G_j \right)
\]

(2)
where $C$ is the number of classes, $y_i$ is the ground truth label of class $i$. Here, we simplify the form of $G$ as $G_i = g(f_i(s_1), f_i(s_2), \ldots, f_i(s_K))$ and evenly averaging is used as the fusion function $g$. Depending on the fusion function $g$, multiple snippets are jointly used to optimize the model parameters $W$. In order to learn the model parameters, stochastic gradient descent (SGD) is chosen as the back-propagation algorithms and the gradients of model parameters $W$ with respect to the loss is derived as equation (3):

$$\frac{\partial L(y, G)}{\partial W} = \frac{\partial L}{\partial G} \sum_{i=1}^{K} \frac{\partial g}{\partial f_i(s_k)} \frac{\partial f_i(s_k)}{\partial W}$$

Here $K$ is the number of segments for all the videos.

### 3.2. Learning FightNet

As the TSN took a principle of temporal segment, the redundant frames in videos were largely reduced. Based on TSN, the FightNet makes appropriate modifications to better adapt to the detection of violent interaction.

#### 3.2.1. FightNet Architectures

The architecture of TSN adopted the Inception of Batch Normalization (BN-Inception) [24] to make the network wider and deeper, resulting to a huge structure as deep as 34 layers. Mathematically, violent interaction detection is modeled as binary classification problem which is not so complex as the action recognition. Therefore, in terms of architectures of FightNet, the most important enhancement compared with TSN is to cut off the last inception structure which includes 6 convolution layers. This measure could largely reduce the risk of over-fitting and shorten the training time.

#### 3.2.2. Input Data

Originally, the RGB images are applied for spatial stream and stacked optical flow fields for the temporal stream in two-stream ConvNets [19]. TSN has explored four input modalities to enhance the discriminative power, ranging from RGB images and optical flow fields to RGB difference and warped optical flow fields. Experimental results demonstrate that the former two input modalities make great contribution to the accuracy and the latter two input modalities plays little role in the improvement of accuracy. Here, we choose three modalities as the input of FightNet, i.e., RGB images, optical flow fields and acceleration fields. Considering acceleration fields as the input modality derives from the specific attributes of violent interaction that the actions are of short duration and forceful. Theoretically, acceleration fields are appropriate to learn motion features for FightNet. For two consecutive optical flow images $F(t-1)$ and $F(t)$, the difference $\text{diff}(t)$ is first computed using equation (4) and the acceleration fields $\text{Accel}(t)$ is defined as equation (5):

$$\text{diff}(x,y,t) = F(x,y,t) - F(x,y,t-1)$$

$$\text{Accel}(x,y,t) = \begin{cases} 0, & \text{diff}(x,y,t) + 128 < 0 \\ \text{diff}(x,y,t) + 128, & 0 \leq \text{diff}(x,y,t) + 128 \leq 255 \\ 255, & \text{diff}(x,y,t) + 128 > 255 \end{cases}$$

Here, $x = \{1, 2, \ldots, m\}$, $y = \{1, 2, \ldots, n\}$, $t = \{1, 2, \ldots, T\}$ and $(x,y,t)$ implies the pixel position $(x,y)$ in $t$-th frame. In this work, acceleration fields consist of two directions, x and y. In the acceleration image of each direction, we view the acceleration field as a general image. As shown in figure 2, RGB images are used to provide appearance information and optical flow fields and acceleration fields are the source of dynamic information.
3.2.3. FightNet Training. While there exist a number of well-studied datasets for action recognition [8,9], a few significant datasets with violent actions are available. For example, the most widely used fight dataset from [7] contains only 1000 video clips which is nowhere near enough to train FightNet. Therefore, we collect a violent interaction dataset (VID) from UCF101[8], HMDB51[9] and the dataset in [7], and the total number of 2314 videos is contained, including 1077 violent interaction clips and 1237 non-violent ones. Besides, to mitigate the risk of over-fitting, we take several strategies for training FightNet as follows.

(1) Pre-training. It has been proven that pre-training is an effective way to initialize deep ConvNets when the target dataset does not have enough training samples [19]. As a deep ConvNet, FightNet takes the parameters trained on ImageNet [20] as initial values. Then RGB images extracted from UCF101 is applied to pre-train the spatial network of FightNet. As mentioned on TSN [6], we utilize RGB models to initialize the temporal networks with respect to the other two modalities, optical flow field and acceleration fields. At last, the pre-trained networks are employed on the proposed VID.

(2) Data Augmentation. In TSN, four data augmentation techniques, i.e. corner cropping, scale-jittering, random cropping and horizontal flipping are employed to prevent over-fitting. In this work, another technique is proposed, video clipping. In video clipping, we clip the long video more than 150 frames into diverse clips. For instance, a long video sample contains 200 frames. We segment it into 2 samples with the frame index 1-100, and 101-200. Using this technique, the scale of training data is augmented by about 30%.

3.3. Testing FightNet
Similar to the testing scheme of original two-stream ConvNets[19] and TSN [6], we extract 25 RGB frames and optical flow stacks or acceleration stacks from every video sample. Besides, 4 corners and 1 centre, and their horizontal flipping from sampled frames are employed to evaluate FightNet. We take a weighted average strategy to fuse the spatial and temporal networks. Larger weight is placed on the temporal stream as the motion information plays more important role on the fight than the appearance information. We set the weight as 0.25 for spatial network and 0.75 for temporal. When both optical flow fields and acceleration fields are applied, the fusion weights are 0.6 and 0.4 respectively.

4. Evaluation
4.1. Dataset
Nievas et al [7] introduced two datasets explicitly designed for assessing fight detection. As shown in figure 3-top, the first dataset (“Hockey”) consists of 1000 clips, 500 fight ones and 500 non-fight ones.
extracted from hockey games of National Hockey League. The second dataset (“Movies”) contains 200 video samples in which fight clips were extracted from action movies and non-fight ones from public action recognition dataset, shown in figure 3-bottom. As the dataset about fight detection is too small to train deep ConvNets, we have collected a violent interaction dataset (VID), aggregating from four datasets, HMDB51, UCF101 and all the video samples of “Hockey” and “Movies”. The proposed dataset contains 2314 videos, including 1077 fight ones and 1237 no-fight ones. In HMDB51 dataset, 206 violent interaction videos are picked out from “hit”, “kick” and “punch”, which show the fight between two persons. In UCF101 dataset, 271 fight videos from “Punch” and “SumoWrestling” are selected. The similar videos like “BoxingSpeedBag” and “BoxingPunchingBag” from UCF101 are considered as the non-fight samples. Examples of frames extracted from the proposed dataset [25] are shown in figure 4. It is obvious that this dataset is challenging involving illumination variations, view point changes and camera motions.

![Example images](image1)

**Figure 3.** Examples dataset of [7], where (a) and (b) are fight and non-fight images from “Hockey” dataset and (c) and (d) from “Movies”.

![Example images](image2)

**Figure 4.** Frame examples in violent interaction dataset (VID), with images in red boxes selected from violent interaction (positive) videos and those in green boxes from negative samples.

### 4.2. Implication Details

For the sake of the extraction of optical flow fields and acceleration fields, we employed the TVL1 optical flow algorithm [26] to finish this task implemented on OpenCV with CUDA. The same as TSN, FightNet is speeded up through a data-parallel strategy with a modified version of Caffe [27] and OpenMPI.
As mentioned in Section 3.2, we firstly construct the architecture of FightNet by cutting off the redundant layers of TSN network. Secondly, FightNet is pre-trained on UCF101 dataset and the pre-trained parameters are considered as the initialization values of FightNet. Thirdly, because we predict 2 classes in VID instead of 101 classes in UCF101, we change the “InnerProduct” layer in the model and change the name from fc_action to fc_action_fight in the prototxt file. Since there is no layer named that, this layer will begin training with random weights. In the solver prototxt, we decrease the base_lr to 0.0001 for spatial networks and 0.0005 for temporal networks, but boost the lr_mult on the fc_action_fight layer. This strategy is to have the rest layers change slowly with the new dataset, but let the fc_action_fight layer learn fast. Additionally, test_iter is adjusted in terms of the size of test data. In the dataset VID, around 20% samples are used as testing data and the others as training data. The training procedure stops at 2000 iterations for spatial networks with RGB images as input and 6000 iterations for temporal networks with optical flow and acceleration fields as input. With 4 TITANX GPUs, the whole training time on VID is around 15 minutes for spatial TSNs and about 1.5 hours for temporal TSNs.

4.3. Experimental Results and Evaluation
In this subsection, we focus on the study of the FightNet framework and the comparison with the state of the art. As the effect of fusion functions and ConvNet architectures have been studied in TSN [6], we quote the conclusion of TSN directly. The average pooling is chosen as the default fusion function and we consider BN-Inception [24] as the ConvNet architecture in accordance with its nice performance in the image classification task. In this work, we adopt the traditional definition of accuracy, as formed in equation (6).

\[ \text{Accuracy} = \frac{TP + TN}{Total} \times 100\% \]  

(6)

where \( TP \) and \( TN \) denote True Positives and True Negatives, \( Total \) being the total number of testing samples.

With all the design choice prepared for FightNet, different combinations of three input modalities are compared on the VID. As reported in table 1, the networks with three different input modalities achieve accuracy greater than 94% while the spatial network taking RGB images as input yields the highest accuracy to 96.81%. We guess that the RGB images provide not only appearance information but also motion features to better detect violent interaction. Then we observe that the combination of any two modalities boosts the performance compared to that of one modality. The fusion of RGB images and acceleration fields could improve the performance to 97.05% which is approximately equal to the combination of all three modalities (97.06%). We conjecture that the reason of high accuracy from the combination of RGB images and acceleration field is the comprehensive effect of the texture information from the former and motion information from the latter. It is also concluded that the proposed input modality, acceleration field is effective and suitable for violent interaction detection in ConvNet.

| Modality             | RGB image | Optical flow | Acceleration | RGB & optical flow | RGB & acceleration | Optical flow & acceleration | All modalities |
|----------------------|-----------|--------------|--------------|-------------------|-------------------|---------------------------|----------------|
| Accuracy             | 96.81%    | 94.23%       | 94.69%       | 96.85%            | 97.05%            | 95.33%                    | 97.06%         |

After exploring the performance of different input modalities for FightNet, we are ready to compare FightNet with other fight detection algorithms. Specifically, we assemble three input modalities and test it on two widely used datasets, “Hockey” and “Movies”. Since the number of videos in “Hockey” and “Movies” datasets is not enough to train FightNet, the model is trained on VID and fine-tuned on the corresponding dataset. In this work, 80% samples on the dataset are randomly selected as training
data and the others as testing data. In addition, we want to classify that the testing data in “Hockey” and “Movies” is involved in the testing data of VID but there is no intersection with the training data in VID. Because there is still no deep learning approach on fight detection, we compare our method with some traditional representations, as shown in Table 2. In [7], by exploring STIP features and MoSIFT features on “Movies” and “Hockey” datasets, MoSIFT was considered as the best descriptor and the superiority of MoSIFT has also been proven in other applications of action recognition. In addition, violence detection using extreme acceleration features [17] and fight detection based on motion analysis [18] are employed to make comparison with the performance of FightNet.

Table 2. Comparison results on the “Hockey” and “Movies” datasets, the values in this table denoting the accuracy on the corresponding dataset.

| Features                | Classifier     | Dataset    | Movies | Hockey |
|-------------------------|----------------|------------|--------|--------|
| BoW (STIP) [7]          | SVM            | 82.5%      | 88.6%  |
|                         | AdaBoost       | 74.3%      | 86.5%  |
| BoW (MoSIFT) [7]        | SVM            | 84.2%      | 91.2%  |
|                         | AdaBoost       | 86.5%      | 89.5%  |
| Extreme Acceleration [17]| AdaBoost   | 85.4%      | 90.1%  |
| Motion Analysis [18]    | Random Forest  | 98.5%      | 84.5%  |
| FightNet                | Softmax        | 100%       | 97.0%  |

As shown in table 2, the previous performance on “Movies” dataset is about 98% in [17] and [18], but on “Hockey”, motion analysis [18] yielded a lower accuracy by around 6% than extreme acceleration features [17]. The feature of MoSIFT achieved nice performance to more than 91% on the “Hockey” dataset, but it is not well on the “Movies” dataset. It is to say that these features are not robust enough to violence detection with different disturbance. However, the robustness is of vital importance for violent interaction detection. The performance of FightNet on “Movies” dataset reaches to 100% partly because the non-fight videos on this dataset are quite different from the fight ones. On the contrary, positive and negative data on “Hockey” are roughly similar which are all collected from the hockey sport, leading to the lower accuracy (97%) than that on “Movies”. In terms of the performance of FightNet, it is concluded that deep ConvNets could capture more essential features and detect the violent interaction correctly with illumination variations, view point changes or camera motions.

5. Conclusion

In this work, we constructed a ConvNet named FightNet to model long-term temporal structure for violent interaction detection. By pre-training on UCF101 dataset of action recognition, the FightNet works well on the proposed violent interaction dataset (VID), which is collected from four public datasets, “Hockey”, “Movies”, HMDB51 and UCF101. A new modality, acceleration field is explored to capture the motion features and the contribution on accuracy is about 0.3%. Compared with other methods on fight detection, the FightNet achieves a higher accuracy with a reasonable computational cost. With the increasing scale of dataset, it is of vital importance to further explore efficient ConvNets to detect violent interaction in complex real-life scenarios in the future.

References

[1] Zhang T, Yang Z, Jia W, Yang B, Yang J and He X 2016 A new method for violence detection in surveillance scenes Multimedia Tools and Applications 75(12) pp 7327-49
[2] Gracia I S, Suarez O D, García G B and Kim T K 2015 Fast fight detection PloS one 10(4) p e0120448
[3] Fu E Y, Va Leong H, Ngai G and Chan, S 2016 Automatic Fight Detection in Surveillance Videos Proc. of the 14th Int. Conf. on Advances in Mobile Computing and Multi Media ACM pp 225-34
[4] LeCun Y, Bottou L, Bengio Y and Haffner, P 1998 Gradient-based learning applied to document recognition *Proc. of the IEEE* 86(11) pp 2278-324

[5] Zhang B, Wang L, Wang Z, Qiao Y and Wang H 2016 Real-time action recognition with enhanced motion vector CNNs *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* IEEE pp 2718-26

[6] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X and Van Gool L 2016 Temporal segment networks: towards good practices for deep action recognition *In European Conf. on Computer Vision* Springer International Publishing pp 20-36

[7] Nievas E B, Suarez O D, García G B and Sukthankar R 2011 Violence detection in video using computer vision techniques *Int. Conf. on Computer Analysis of Images and Patterns* Springer Berlin Heidelberg pp 332-9

[8] Soomro K, Zamir A R and Shah M 2012 UCF101: A dataset of 101 human actions classes from videos in the wild arXiv preprint arXiv 1212.0402

[9] Kuehne H, Jhuang H, Garrote E, Poggio T and Serre T 2011 HMDB: a large video database for human motion recognition 2011 IEEE Int. Conf. on Computer Vision IEEE pp 2556-63

[10] Lam V, Phan S, Le D D, Duong D A and Satoh S I 2016 Evaluation of multiple features for violent scenes detection *Multimedia Tools and Applications* pp 1-25

[11] Clarin C, Dionsio J, Echavez M and Naval PC 2005 Dove: Detection of movie violence using motion intensity analysis on skin and blood *ECCV* pp 150–6

[12] Liang-Hua C, Hsu HW, Wang LY and Su CW 2011 Violence detection in movies *Computer Graphics Imaging and Visualization (CGIV)* pp 119–24

[13] Tran D, Bourdev L D, Fergus R, Torresani L and Paluri M 2015 Learning spatiotemporal features with 3d convolutional networks *ICCV* pp 4489-97

[14] Ji S, Xu W, Yang M and Fei-Fei L 2013 3D convolutional neural networks for human action recognition *IEEE Trans. on Pattern Anal. And Mach. Intell.* 35(1) pp 221-31

[15] Ioffe S and Szegedy C 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift arXiv preprint arXiv p 1502.03167

[16] Zach C, Pock T and Bischof H 2007 A duality based approach for realtime TV-L 1 optical flow *Joint Pattern Recognition Symp.* Springer Berlin Heidelberg pp 214-23.

[17] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T 2014 Caffe: Convolutional architecture for fast feature embedding *ACM Multimedia* pp 675-8