Joint Unsupervised Infrared-RGB Video Registration and Fusion

Imad Eddine Marouf [1,2] ***Luca Barras[2]*** ***Hakki Can Karaimer[2,3]*** ***Sabine Süsstrunk[2]***
[1]Institut Polytechnique de Paris
[2]School of Computer and Communication Sciences (IC), Ecole Polytechnique Fédérale de Lausanne (EPFL)
[3]Advanced Micro Devices, Inc. (AMD)
imad.marouf@ip-paris.fr luca.barras@epfl.ch hakki.karaimer@epfl.ch sabine.susstrunk@epfl.ch

Abstract

We present a system to achieve joint registration and fusion of RGB and Infrared (IR) videos. While RGB is related to human perception, IR is related to the heat. However, IR images often lack good contour and texture information. An increasing number of researchers work on fusing visible and IR images to obtain more information from them, which requires two completely matched images. However, classical methods assuming ideal imaging conditions fail to achieve satisfactory performance in real cases. From the data-dependent modeling point of view, labeling the dataset is costly and impractical.

In this context, we present a framework that tackles two challenging tasks. First, a video registration procedure aims to align IR and RGB videos. Second, a fusion method brings all the essential information from the two video modalities to a single video. We evaluate our approach on a challenging dataset of RGB and IR video pairs collected for firefighters to do their missions more efficiently due to the difficulties in the vision, such as environments with heavy smoke after a fire, see our project page.

Introduction and Motivation

The mission of firefighters is to rescue people and animals from hazardous fire, but due to difficult working conditions within places of fire makes it very difficult to save lives, such as smokes, toxic fumes and super heated glasses. In 2011, 70,090 firefighters in the U.S. alone were injured in the line of duty with 61 deaths [5, 12, 16, 22]. Over 60% of the firefighter deaths and over 20% of the firefighting injuries are caused by exposure to fire conditions such as smoke inhalation, burns, overexertion/stress, or being trapped [7, 13, 16]. Thus, they cannot perform well in fire smoke-filled environments where low visibility and high temperature are present. Lack of visibility might lead to influences of human behavior such as redirection of movement and their initial response speed [2, 3, 11]. As solution to this visibility issue, firefighters use IR cameras to aid in seeing through smokes and detect the variation of temperature in the surroundings. IR camera is merely based on temperature variation within the frame captured, it cannot detect stable, non-variation temperatures comparing to RGB camera where it rely on human perception.

In order to overcome these limitations, we provide an efficient framework to fuse the information captured by IR/RGB cameras to have a full view of the environment and facilitate firefighters' mission in smoke and nighttime operations. We present our results on a dataset containing IR/RGB video pairs of scenes that mimic working environments for firefighters. The dataset comprises IR-RGB video pairs captured by two cameras, but they are not well aligned.

Contribution: In this paper, we present a deep-learning-based framework for unsupervised joint registration and fusion for RGB-IR videos in order to produce well aligned pairs of RGB and IR. Thus fusion mutually complements the drawbacks of each camera type and maximizes the vision capability within the environment. Our proposed method is evaluated on a very challenging dataset consisting of videos mimicking the working conditions of firefighters.

Related Work

Image Registration: Demand for faster registration methods motivated the development of deep learning methods based on transformation estimation techniques and challenges associated with generating ground truth data have recently motivated many groups to develop unsupervised frameworks. Two recent papers [4, 17], were the first to present an unsupervised learning based image registration methods. Both propose a neural network consisting of a CNN and spatial transformation function [10] that warps images to one another. However, these two initial methods are only demonstrated on limited subsets, such as 3D subregions [17] or 2D slices [4], and support only small transformations [4]. All above methods were demonstrated on medical images, rare works has been done on non-medical images such as [9], where it relies on detecting corners between input pairs and evaluating using similarity metric. Another approach [6] applies registration between Near-Infrared and RGB pairs based on feature points using a variant of SIFT detector.

Image Fusion: The earliest fusion work involving neural networks poses multi-focus fusion as a classification task [19]. Three focus measures define the input features to a shallow network which outputs the weight maps corresponding to the source images. Due to architectural constraints, the method can only run on image patches, and generates boundary artifacts. More recently, convolutional neural networks have been trained to generate decision maps for multi-focus [21], multiexposure [23], medical [20], and thermal fusion [18]. Although these approaches often achieve better performance than their classical counterparts, they still have major drawbacks. First, they require large datasets for training. Second, deep networks often overfit the datasets they are trained on, e.g., a network trained for multi-focus image fusion will only be suitable for that task. Method proposed in [15] does not require training, which alleviates the necessity of collecting data, by using pre-trained network as feature extractor.

Background and Preliminaries

Image Registration

Image registration is the process of transforming different images into one coordinate system with matched imaging contents, which has significant applications in medical imaging, remote sensing, and 3-D computer vision. Registration may be necessary when analyzing a pair of images that were acquired from different viewpoints, at different times, or using different sensors/modalities like IR and visible images. Further, demand for faster registration methods later motivated the development

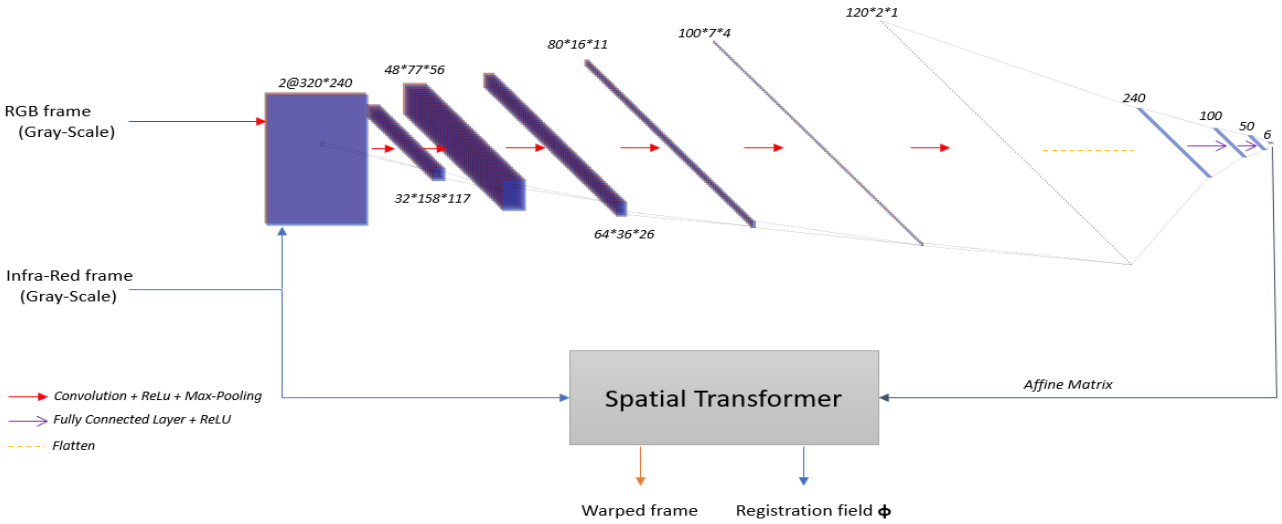

Figure 1. *Affine network architecture: The affine network is a spatial transformer network. It takes as input an 2-channel image that is the concatenation of the IR frame and the RGB frame. After a number of convolution and non-linear layers it generates a warped frame and a registration field.*

of deep learning based one-step transformation estimation techniques and challenges associated with generating ground truth data have recently motivated many groups to develop unsupervised frameworks for one-step transformation estimation.

VoxelMorph Network Firstly, we evaluated an unsupervised registration model [1] since it is state-of-the-art in medical image registration with a significant speed in the registration process. Dalca et al. [1] propose a CNN function with parameters shared across population, enabling registration to be achieved through a function evaluation, which can be optimized for a various cost functions. The model is fed by pair of fixed and moving images of equal size that are concatenated, which are RGB and IR volumes in our case.

The model is similar to well-know architecture called UNet [25] consisting of an encoder-decoder with skips connections. The network takes fixed f, moving m volumes and apply convolutions followed by Leaky ReLU activations in both the encoder and decoder stage. The convolutional layers capture hierarchical features of the input image pair necessary to estimate the correspondence registration field Φ defined as $g_\theta = (m; f) = \Phi$, where the goal is to optimize the learnt parameters θ to estimate a deformation field Φ. Authors propose unsupervised loss consisting of two terms defined as:

$$L_{us}(f,m,\Phi) = L_{sim}(f,m\circ\Phi) + \lambda L_{smooth}(\Phi), \tag{1}$$

$$L_{sim}(f,m\circ\Phi) = \frac{1}{\Sigma}\sum_{p\in\Sigma}[f(p) - [m\circ\Phi](p)]^2, \tag{2}$$

$$L_{smooth}(\Phi) = \sum_{p\in\Sigma} ||\bigtriangledown u(p)||^2. \tag{3}$$

For each pixel p, we compute a (subpixel) pixel location $\tilde{p} = p + u(p)$ in m. Because image values are only defined at integer locations, we linearly interpolate the values at the eight neighboring voxels:

$$m\circ\Phi(p) = \sum_{q\in Z(\tilde{p})} m(q) \prod_{d\in x,y,z} (1 - |\tilde{p}_d - q_d|). \tag{4}$$

Minimizing L_{sim} will encourage $m\circ\Phi$ (m to approximate f) but may generate a non-smooth Φ that is not physically realistic. Where $Z(\tilde{p})$ are the pixel neighbors of p, and d iterates over dimensions of Φ. Because we can compute gradients or subgradients, we can backpropagate errors during optimization.

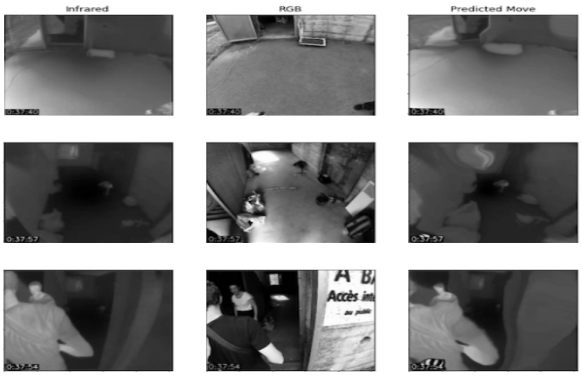

Figure 2. *Applying the state-of-the-art VoxelMorph model directly on our IR-RGB pairs performs poorly on the challenging scenes. We employed a segmentation procedure to tackle this problem leveraging the IR camera's response change based on temperature.*

Due to the fact that this method is unsupervised, we need a way to know if the deformation field Φ is doing good by making sure that $m.\Phi$ (m warped by Φ) is close to f, regularizing θ make it smooth. A spatial transformation function $T(\Phi)$ to interpolate the height neighboring voxels, in order to overcome their lack of ability to be spatially invariant to the input data. The use of spatial transformer results in models which learn invariance to translation, scale, rotation and more generic warping.

Spatial Transformation The spatial transformer network (STN) proposed by Jaderberg et al. [10] was one of the first methods that exploited deep learning for image alignment. The STN is designed as part of a neural network. Its task is to spatially transform input images such that the image registration is simplified. Transformations might be performed using a global transformation model or a thin plate spline model. In the application of an STN, image registration is an implicit result.

Image Fusion

Image fusion is the process of combining multiple input images into a single output image which contain better description of the scene than the one provided by any. Image fusion is an important image processing and computer vision application. For example, it is used in night-time surveillance, military reconnaissance mission and firefighting by fusion of visible and IR images. In this paper, we are targeting fusion of IR and RGB im-

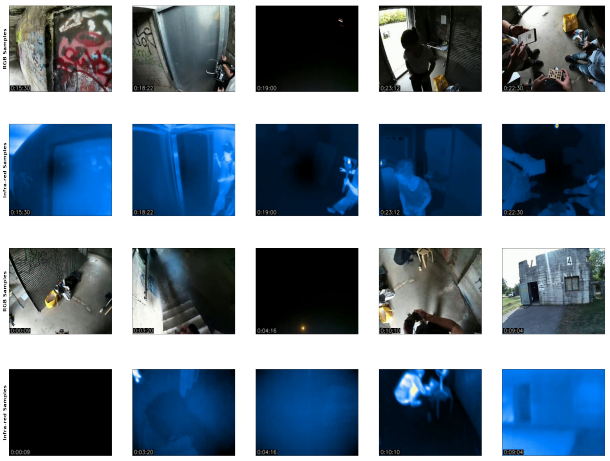

Figure 3. *Image samples from our dataset. The dataset contains indoor, outdoor with different contrast and brightness conditions.*

ages using state-of-the-art method [15] which uses a pretrained VGG19 model [26] considered as feature extractor. Therefore, this method alleviates the need to do training and collection of data, and can generalize well to any type of fusion. The proposed method works as follows: First, it decomposes both input images (IR, RGB) into base layer representing large scale intensity variation, and a detail layer containing small scale changes to avoid mixing low and high frequency information and reduces halo artifacts. Base layer is obtained by applying smoothing filter on the image, and the detail layer is the difference between the original image and the base layer. Then, the base layers are fused based on visual saliency maps calculated which reflects the important features in the image. The detail layers are fused using deep features maps extracted from detail image sources using pre-trained network according to their activation levels at each layer in the pre-trained network. Each feature map of the detail layer indicates the contribution of the image at a specific pixel, high pixel values correspond to high activity. Finally, after obtaining both base, and detail layers, the final fused image will be combination of both layers pixel-by-pixel, and cropping pixel values of the fused image to remove out of range values.

Proposed Method

Our framework consists of two stages: First, we created a spatial transformer network, *Affine network*, which takes as input IR-RGB pairs, in order to create flow field. Flow field output of the affine network will apply the transformation on the original frames (IR frames). When we have this flow field, we just need to put each channel (r, g, b) of the original frame in the spatial transformer and we obtain the IR moved frame that must be better aligned to RGB frame. Secondly, fusion network will take registred IR image, with original RGB image to produce an image which will complement all the missing information of the well matched IR-RGB pairs. This is followed by alpha-blending to color the resulting image.

Affine network The affine network is a spatial transformer network. It takes as input an 2-channel image that is the concatenation of the IR frame and gray-scale of the RGB frame. The input size is of $2 \times 320 \times 240$ in our case. The network begin with a succession of 2D convolutions followed by ReLU activation with kernel size of 7×7, 5×5, 5×5, 5×5, 3×3, 3×3 respectively. All the convolutions are done with a stride of size 1. Successively, we flatten the output of convolution features, followed by fully connected layers of size $240, 100, 50, 6$. Output of the network can be seen as affine matrix of size 2×3. In the spatial transformer we apply the affine matrix on the IR image and get the first output (warped image) of the network. The second output is the flow field generated by the affine matrix. The first output has a size of 320×240 and the second a size of $2 \times 320 \times 240$. 2 channels because we have a channel for each direction (x, y). Figure 1 shows the visualization of the architecture.

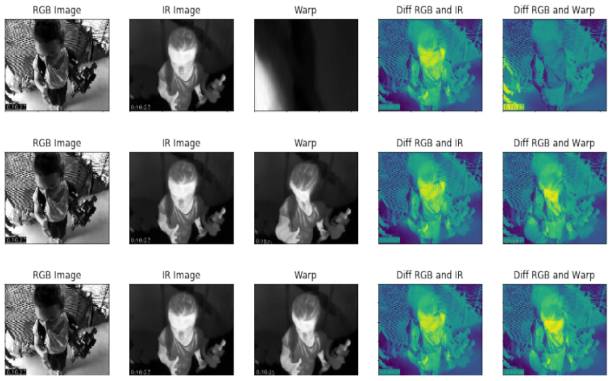

Figure 4. *Image Registration with VoxelMorph architecture (1st, 2nd row), and Affine network (3rd row). For VoxelMorph, the output image is too warped when we use a small value for $\lambda : 0.2$ and not aligned. For bigger value $\lambda : 0.6$, the result is not much warped but not aligned too. With Affine network, the output is too scale and not at all aligned with the RGB frame*

The loss function is the similitude between the RGB image (fixed input) and the warped image. It can be expressed as:

$$L(F,M,\theta) = MSE(f, m \circ \Phi) = \frac{1}{\Sigma} \sum_{p \in \Sigma} [f(p) - [m \circ \Phi](p)]^2 \quad (5)$$

where F is the RGB image, M is the IR image and $M(\Phi)$ is the IR image after to be went out of the spatial transformer. L_{sim} is the mean square error in our experiment. We use an architecture very similar to VoxelMorph but we start with a CNN to regress to a 2×3 affine matrix, and apply this matrix on the image. So the shape in the image cannot be deform because we apply an affine matrix on the image. VoxelMorph apply a flow field in the spatial transform component. So the pixels can move anywhere, it is for that they add the $L_{smooth}(\Phi)$ loss term in loss function. Affine network do not need it because the linearity of the transformation is guaranteed by the affine matrix.

Segmentation Applying VoxelMorph or Affine model directly on our IR-RGB pairs did not perform good enough as illustrated in Fig 2. Due to the fact that the input pairs frame provided are not well aligned, bigger size, and challenging compared to medical images since RGB/IR pairs are not segmented. Mostly, the difference between IR and RGB images is at the borders of the captured scene, which is different from the well centered and segmented medical imaging benchmark datasets where VoxelMorph is evaluated. We have to apply semantic segmentation on the RGB pairs using pretrained Mask-RCNN model [8], on the other hand IR images are segmented using adaptive threshold method depending on the frames sequences, and apply image registration on the resulting segmentation where it learns to align the segmentations.

Mask-RCNN [8] is a deep neural network aimed to solve semantic segmentation problem in computer vision. It takes an image as input and produces the object bounding boxes, classes and masks. Mask-RCNN is a two-stage framework: The first 10 stage scans the image and generates proposals (areas likely to contain an object). Second stage classifies the proposals and generates bounding boxes and masks, both stages are connected to

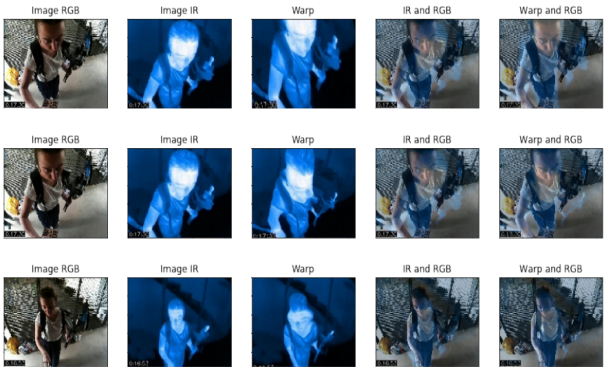

Figure 5. *Image Registration with VoxelMorph architecture, and Affine network respectively. We achieve much better results and remark that the VoxelMorph (2nd row, 3rd row) trained with $\lambda : 0.6, \lambda : 1$ respectively align better the frames but deform the shape in the image. When compared visually, Affine network (1st row) achieves better result.*

the backbone structure. Mask-RCNN is an extension of Faster R-CNN [24] with an extra mask head. The extra mask head allows us to pixel-wise segment each object and extracts each object separately without any background.

Experimental Setup and Analysis

Dataset We demonstrate our work on challenging video-pairs of IR and visible with varying duration ranging from 6 minutes to 12 minutes. Recorded in many places with different brightness and contrast conditions to emulate the working conditions for firefighters ranging from extremely dark scenes to very bright ones. Therefore, before tackling the registration part we had to carry out pre-processing steps, by converting our videos into frames and resizing them to 320 $\times$ 240 size. We generated the masks using Mask-RCNN [8] corresponding to the frames. We generated only masks for persons. So in this part we use only frames that contain humans. For the RGB images, we use a pre-trained Mask-RCNN to segment the images. For the IR images, we threshold them based on the pixels belong to humans.

Training Details For the Voxelmorph network, we adapted the official Pytorch implementation with adjustments to fit our 2D images dataset, VoxelMorph is initialised with the number of features for encoder and decoder as [256, 256, 256, 256], [256, 256, 256, 256, 256, 128]. For the two networks, we use Adam optimizer [14] with a learning rate of 10^{-4}. To speed up the learning, we use a batch of 20 pairs of images for the two networks. To segment the RGB frames, we use the official pretrained Mask-RCNN. To better see if an IR frame is aligned with its corresponding RGB frame, we implemented a visualization method that shows IR image, RGB image, warped image, the difference between the RGB and IR image and the difference between the RGB and the warped image (see Figure 5). Consequently, we fuse the resulting image using [15], followed by alpha blending as an interpolation between the two images. The formula is given by: $\alpha Y + (1 - \alpha)F$, where F is the original RGB frame and Y is the fused IR-RGB frame.

Results

In this section we present our results. First, without employing any segmentation procedure, then using a semantic segmentation for humans.

Without segmentation The results without segmentation are not very good (see Figure 4). We trained both networks for 20 epochs. We tried to do more epochs but the results getting worse. We varied λ between 0.1 to 0.6 for the first network and we have

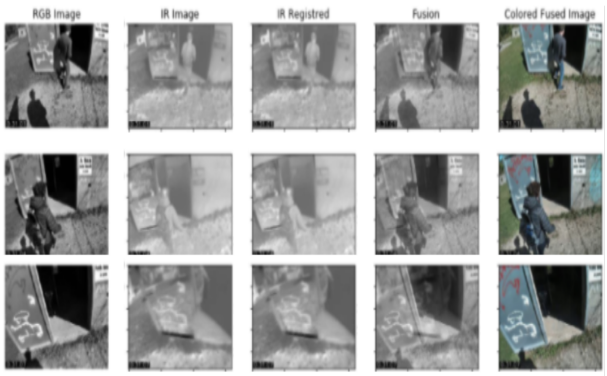

Figure 6. *Fusion of IR/RGB wrapped after segmentation, using Zero-Fusion [15]. Right-most column corresponds to applying alpha-blending to fused IR-RGB image.*

Figure 7. *It is clear that the t-shirt white (very light) and the t-shirt black (very dark) on the RGB frame appear the same color on the IR frame. Which results to very low loss score, although, the IR-RGB pair are quite different from each other.*

the output image too warped and not aligned when we use a small value for λ. Bigger values for λ lead to much warped output but not well aligned.

Using semantic segmentation for humans For this part, we trained 100 epochs and varied λ between 0.6 and 1. This achieves much better results and the VoxelMorph align better the frames but deform the shape in the image. So visually, the Affine network give better results. Results are shown in Figure 5.

IR-RGB Fusion Fusion process is applied after IR-RGB registration. Figure 6 illustrates the results from the fusion process using the registred IR frame with original gray-scale image to complement each image modality, and have full-information about the scene. The fusion process is followed by alpha blending to highlight the original colors of the image.

Discussion The results turn much better with the segmentation (see Figure 5). Applying VoxelMorph registration directly lead to poor results because we cannot compare the RGB frame with IR frame because pixel's values are not correlated. For example, if on the RGB image a person has a black t-shirt or a white t-shirt, on the IR image it will appear like it was the same ones (see Figure 7). So if we compute the mean square error between the IR and RGB images, it is going to measure the similitude between the two images which is not what we want. The only correlation between these two images is the shape borders in the image.

Concluding Remarks

We presented a system for solving the joint IR-RGB image registration and fusion problem to aid firefighters in implementing their missions in challenging visibility conditions. For this difficult task, we made a significant advancement to tackle the issue. Our results provide a backbone for future improvements to overcome the limitations we had regarding the segmentation of IR and RGB image pairs.

References

[1] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *arXiv*, 2018.

[2] J. Bryan. Behavioral response to fire and smoke. *Society of Fire Protection Engineers*, 2, 2002.

[3] G. Cook and M. Wright. The effects of smoke on people's walking speeds using overhead lighting and wayguidance provision. *International Symposium on Human Behaviour in Fire*, 2001.

[4] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Isgum. End-to-end unsupervised deformable image registration with a convolutional neural network. *Lecture Notes in Computer Science*, page 204–212, 2017.

[5] Fire Incident Data Organization. Deadliest Fires in the U.S. with 5 or more Firefighter Deaths at the Fire Grounds, 1977–2012. *Technical report*, 2013.

[6] D. Firmenichy, M. Brown, and S. Süsstrunk. Multispectral interest points for RGB-NIR image registration. In *IEEE International Conference on Image Processing*, 2011.

[7] S. Gwynne and E. Ronchi. Fire loss in the United States during 2010. *Technical Report*, 2010.

[8] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *IEEE International Conference on Computer Vision*, 2017.

[9] T. Hrkac, Z. Kalafatic, and J. Krapac. Infrared-visual image registration based on corners and hausdorff distance. In *Image Analysis, 15th Scandinavian Conference*, 2007.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 2016.

[11] T. Jin. Visibility through fire smoke. *Report of Fire Research Institute of Japan*, 42, 2001.

[12] M. Karter. Fire loss in the United States during 2010. *Technical Report*, 2011.

[13] M. Karter. Selected special analyses of firefighter fatalities. *Technical Report*, 2011.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[15] F. Lahoud and S. Süsstrunk. Fast and efficient zero-learning image fusion. *arXiv*, 2019.

[16] P. LeBlanc and R. Fahy. Firefighter fatalities in the united states-2011. *Technical Report*, 2012.

[17] H. Li and Y. Fan. Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv*, 2017.

[18] H. Li, X.-J. Wu, and J. Kittler. Infrared and visible image fusion using a deep learning framework. *IEEE International Conference on Pattern Recognition*, 2018.

[19] S. Li, J. T. Kwok, and Y. Wang. Multifocus image fusion using artificial neural networks. *Pattern Recognition Letters*, 23, 2002.

[20] Y. Liu, X. Chen, J. Cheng, and H. Peng. A medical image fusion method based on convolutional neural networks. *IEEE Fusion*, 2017.

[21] Y. Liu, X. Chen, H. Peng, and Z. Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36, 2017.

[22] J. Molis and M. Karter. US Firefighter Injuries-2011. *Technical Report*, 2012.

[23] K. R. Prabhakar, V. S. Srikar, and R. V. Babu. A deep unsupervised approach for exposure fusion with extreme exposure image pairs. *IEEE International Conference on Computer Vision*, 2017.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv*, 2016.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, 2015.

[26] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015.

Supplemental Material
Data Cleaning

We summarize our dataset cleaning procedure as follows: Initially, we selected only good IR frames we can easily threshold based on humans with their associated RGB frames on each video. Then we threshold the IR frames by sequences of frames. After that, we segment the associated RGB frame with Mask-RCNN. Furthermore, we checked again and removed all the poorly segmented RGB frames. Finally, we have 10431 4-tuples *(RGB, IR, mask RGB, mask IR frames)*. We split it in 8000 frames for the training set, 1000 frames for the validation set and 1400 frames for the test set.

Figure 8 visualizes if the frames are aligned to reconstruct the video after merging the warped frame (IR frame after the transformation) and RGB frame.

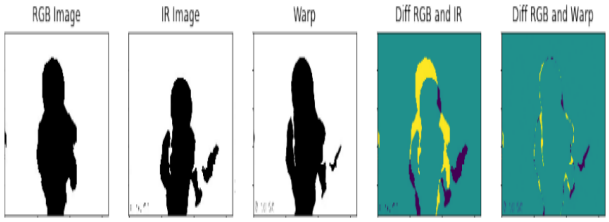

Figure 8. *Our visualization method shows the IR image, RGB image, warped image, the difference between the RGB and IR image and the difference between the RGB and the warped image. This is useful when we work with the masked version of the images. Here we show an example of the visualization method on the masks.*

Ablation Study

Before applying human semantic segmentation, we investigated registration based on contours. We extract contours of the shapes within the frames. First, we applied a Sobel filter in each direction (x, y) and computed its magnitude. The Canny filter followed this, then adaptive thresholding and Laplacian filter. Then, we apply the registration phase. The resulting image does not move globally or moves very little.

Figure 9. *Extracting contours from RGB-IR pairs: It is clear that there is much more contrast on the RGB image (on the left). Such samples can be found in our dataset resulting into poor registration based on contours.*

We claim it is due to on contrast on RGB frames (see Figure 9). In fact, we tried on very clean images, just images of contour of circles and we get the same results, the moving image does not move globally. So we understood that the problem was coming from the loss function. The problem is that the loss L_{sim} function is very flat. So if we move a little bit the image, the loss function will have almost the same value, thus the optimizer will not update the parameters of the network because the gradient will be very close to 0.

We solved this by the segmentation approach. The segmentation procedure significantly boosted the registration performance—a demonstration video is provided in the project page.