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 perform joint registration and fusion for RGB and Infrared (IR) video pairs. While RGB is related to human perception, IR is associated with heat. However, IR images often lack contour and texture information. The goal with the fusion of the visible and IR images is to obtain more information from them. This requires two completely matched images. However, classical methods assuming ideal imaging conditions fail to achieve satisfactory performance in actual 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 that 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 handle their tasks effectively in challenging visibility conditions such as 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. Challenging working conditions such as smokes, toxic fumes, and superheated glasses make it very difficult to save lives. 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 effectively in smoke-filled environments where low visibility and high temperature are present. Poor visibility might lead to influences of human behavior such as redirection of movement and their initial response speed [2, 3, 11]. As a solution, firefighters use IR cameras to aid in seeing through smokes and detect the variation of temperature in the surroundings. While IR cameras' response is related to the temperature within the captured frame, RGB cameras' response is related to human perception.

In order to overcome these limitations, we provide an effective framework to align and fuse the scenes captured by an IR/RGB camera pair. This helps to have a full view of the environment and facilitates firefighters' missions in extreme conditions. 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 attached to a head-mounted wearable device. However, it suffers from misalignment and the sensing differences between the two cameras.

Contribution: This paper presents a deep-learning-based framework for unsupervised joint registration and fusion for RGB-IR video pairs to produce well-aligned and fused videos. Thus fusion mutually complements the drawbacks of each sensing device and maximizes the vision capability within the environment. Our proposed method is evaluated on a very challenging dataset consisting of videos mimicking the extreme working conditions of firefighters.

Related Work

Image Registration: Demand for fast and accurate registration methods motivated the development of deep learning methods based on transformation estimation techniques. Challenges associated with generating ground truth data have recently led many researchers to develop unsupervised frameworks. Two recent works [4, 17], presents an unsupervised learning-based image registration methods. Both propose a neural network consisting of a CNN and a spatial transformation function [10] that warps images to one another. However, these two initial methods are only demonstrated on limited subsets, such as 3D sub-regions [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 the method relies on detecting corners between input pairs and evaluating using a similarity metric. In their seminal work, Firmenichy et al. [6] applies registration between Near-Infrared and RGB pairs based on detection of the feature points.

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. The method proposed in [15] does not require training, which alleviates the necessity of collecting data by using a pre-trained network as a feature extractor.

Background and Preliminaries

Image Registration

Image registration is the process of transforming different images into one coordinate system with matched scene contents. Applications include medical imaging, remote sensing, and 3-D computer vision. Registration may be necessary when analyzing a pair of images acquired from different viewpoints, at different times, or using various sensors/modalities like IR and visible images. Recently, the need for fast and accurate registration

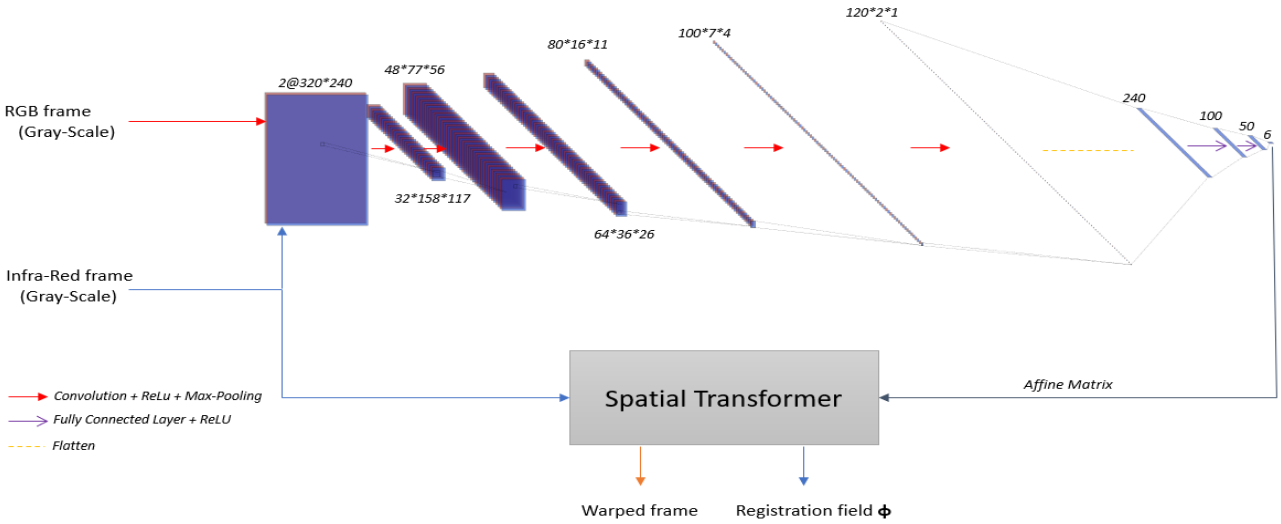

Figure 1. *Affine network architecture: The affine network is a spatial transformer network. It takes a 2-channel image obtained by concatenating the IR frame and gray-scale of the RGB frame as input. After a number of convolution and non-linear layers, it generates a warped frame and a registration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

methods encouraged researchers to propose deep-learning-based transformation estimation techniques.

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

The VoxelMorph model is similar the UNet architecture [25] consisting of an encoder-decoder with skip connections. The network takes fixed f, moving m volumes and apply convolutions. This is 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. The authors propose an unsupervised loss consisting of two terms defined as:

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

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

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

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:

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

Minimizing L_{sim} will help $m\cdot\Phi$ approximate f but may generate a non-smooth Φ that could be physically impractical. Here $Z(\tilde{p})$ are the pixel neighbors of p, and d iterates over dimensions of Φ. Because it is possible to compute gradients, the model can backpropagate errors during optimization.

The fact that this method is unsupervised, we need a way to know if the deformation field Φ is doing good by making sure

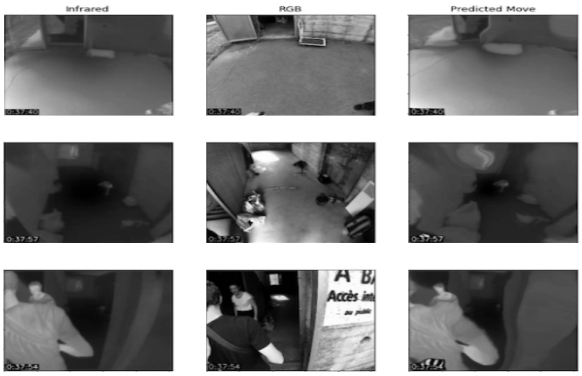

Figure 2. *Applying the state-of-the-art VoxelMorph model directly on our IR-RGB pairs performs poorly on the challenging scenes in our dataset. We employed a segmentation procedure to tackle this problem leveraging the IR camera's response based on temperature.*

that $m\cdot\Phi$ (m warped by Φ) is close to f. Regularizing θ makes the deformation field smooth. A spatial transformation function, $T(\Phi)$, is used to interpolate neighboring voxels' heights to overcome the shortcoming of the model being spatially invariant to the input data. The use of spatial transformer helps the model remain invariant to translation, scale, rotation. This also makes the overall system capable of modeling 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. The goal of the STN 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 that contains a better description of the scene than the one provided by any. The applications include night-time surveillance, military reconnaissance missions, and firefighting by fusing visible and IR images. In this paper, we are targeting fusion of IR and RGB images using state-of-the-art method [15] which uses a pretrained VGG19 model [26] considered as feature extractor. Therefore, this method alleviates

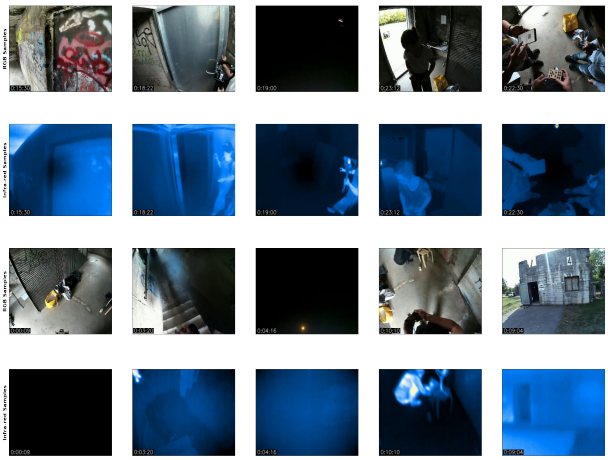

Figure 3. *Image samples from our dataset. The data is captured with the two cameras attached to a head-mounted wearable device in various indoor and outdoor conditions to mimic the working situations for firefighters ranging from extremely dark scenes to very bright ones.*

the need to do training and collection of data, and can generalize well to any type of fusion. The 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 reflecting the image's essential features. The detail layers are combined using deep feature maps extracted from detail image sources using a 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 where high pixel values correspond to high activity. Finally, after obtaining both the base and detail layers, the final fused image combines both layers pixel-by-pixel and removes out-of-range pixel values of the fused image.

Proposed Method

Our framework consists of two stages: First, we created a spatial transformer network named *affine network*. The affine network takes IR-RGB pairs as input and produces a flow field. Flow field output of the affine network will apply the transformation on the IR frames. Second, a fusion network takes a registered image and produces an image that will complement all the essential 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 a 2-channel image obtained by concatenating the IR frame and gray-scale of the RGB frame as input. The input size is $2 \times 320 \times 240$ 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. While the first output has a size of 320×240, the second output is $2 \times 320 \times 240$. The two channels are because we have a channel

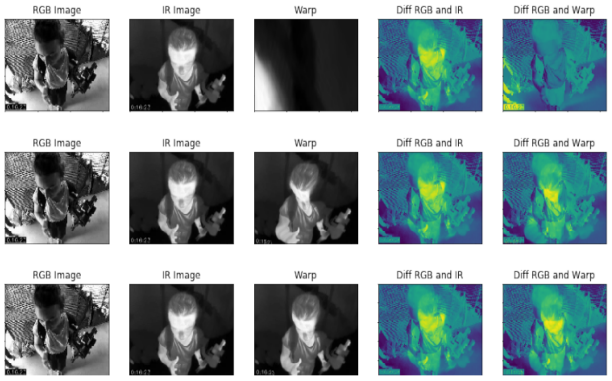

Figure 4. *IR-RGB registration with the VoxelMorph architecture and the Affine network. For small λ values, such as $\lambda = 0.2$, VoxelMorph (1st, 2nd row) poorly warps the output image and causes misaligned results. For larger λ values, such as $\lambda = 0.6$, the output is less warped but still poorly aligned. With the affine network (3rd row), the output looks only scaled and misaligned with the RGB frame.*

for each direction (x, y). Figure 1 shows the visualization of the architecture.

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

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

where F is the RGB image, M is the IR image and $M(\Phi)$ is the IR image after processed by 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 deformed because we apply an affine matrix on the image. VoxelMorph applies a flow field in the spatial transform component. So the pixels can move anywhere. This is the reason why $L_{smooth}(\Phi)$ loss term exists in the loss function. This is unnecessary for the affine network because the linearity of the transformation is guaranteed with the affine matrix.

Segmentation Applying VoxelMorph or Affine model directly on our IR-RGB pairs performs poorly as in Figure 2. This is mostly because the input pairs are not well aligned, and the objects in the images varied in size. This makes the problem more challenging than medical imaging benchmarks, as RGB/IR pairs are not segmented. The difference between IR and RGB images is mostly at the borders of the captured scene, which is different from the well centered and segmented medical imaging benchmark datasets where VoxelMorph is mostly evaluated. In order to resolve this, we apply semantic segmentation on the RGB pairs using the pretrained Mask-RCNN model [8]. Similarly, the IR images are segmented using the adaptive thresholding method depending on the frame sequences. Then the image registration is applied to the segmentation result. This way, the model learns to align the result of the segmentation.

Mask-RCNN [8] is a deep neural network proposed to tackle the semantic segmentation problem in computer vision. It takes an image as input and produces the bounding boxes for objects, classes, and masks. Mask-RCNN is a two-stage framework: The first 10 stage scans the image and generates proposals (i.e., areas likely to contain an object). The second stage classifies the proposals and generates bounding boxes and masks. Both stages are connected to the backbone structure. Mask-RCNN is an extension of Faster R-CNN [24] with an extra mask head. The

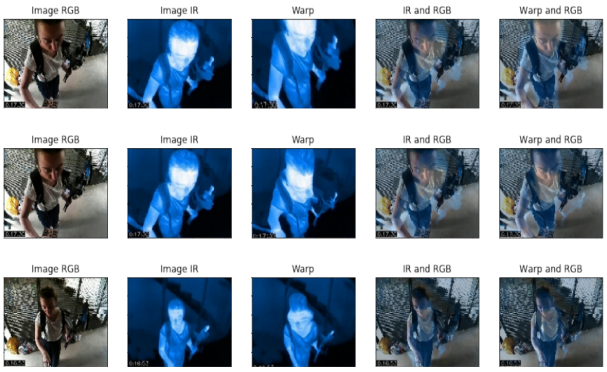

Figure 5. *IR-RGB registration with the VoxelMorph architecture and the Affine network. When VoxelMorph (2nd and 3rd row) is trained with $\lambda = 0.6$ and $\lambda = 1$, it poorly deforms the shape of the objects in the image. When compared visually, the affine network (1st row) achieves better results.*

extra mask head allows us to pixel-wise segment each object and extracts each object separately from the background.

Experimental Setup and Analysis

Dataset We demonstrate our work on IR and visible video pairs with varying duration ranging from 6 minutes to 12 minutes. The data is captured with the two cameras attached to a head-mounted wearable device to mimic the working situations for firefighters ranging from extremely dark scenes to very bright ones. Therefore, before tackling the registration, we had to apply pre-processing 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. The masks were obtained only for persons. So in this part, we use only frames that contain humans. For the RGB video frames, we use a pre-trained Mask-RCNN to segment the images. For the IR video frames, we applied thresholding 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 image 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 pre-trained 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). Then we fuse the resulting image using [15]. This is 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, then using a semantic segmentation procedure for humans.

Without segmentation After training both networks for 20 epochs, the results without segmentation are not satisfactory (see Figure 4). When we increase the number of epochs, the results become worse. In another approach, we vary λ between 0.1 to 0.6 for the first network. For small λ values, the output image is overly warped and not aligned. Larger values for λ further warp the output but make it poorly aligned.

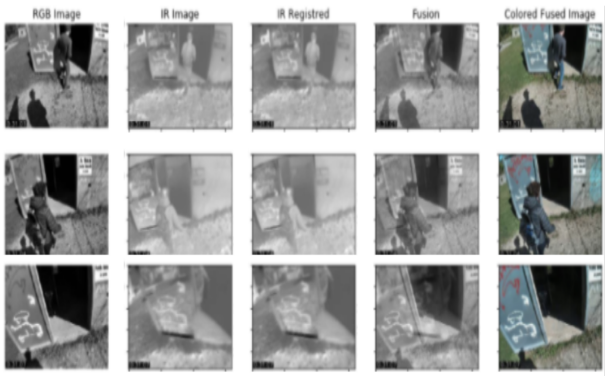

Figure 6. *Fusion of IR/RGB wrapped after segmentation, using ZeroFusion [15]. The right-most column corresponds to applying alpha-blending to fused IR-RGB images.*

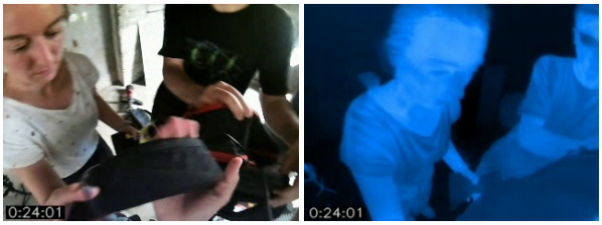

Figure 7. *The white t-shirt and the black t-shirt on the RGB frame appear the same on the IR frame. Although the IR-RGB pair are quite different, this results in a very low loss score.*

Using semantic segmentation for humans: When we employ semantic segmentation for humans, we train the network 100 epochs and vary λ between 0.6 and 1. This way, VoxelMorph achieves better alignment performance. However, it deforms the image's shape. So visually, the affine network achieves better results. Results are shown in Figure 5.

IR-RGB Fusion The fusion process is applied after IR-RGB registration. Figure 6 illustrates the results from the fusion process using the registered IR frame with original gray-scale image to complement two imaging modalities. It brings essential 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 leads to poor results because we cannot compare the RGB frame with the IR frame. After all, pixel values are not correlated. We explain this with an example in Figure 7. As in Figure 7, the IR camera's response is the same for the two persons regardless of the two different colors and textures of the t-shirts. So if we compute the mean square error between the IR and RGB images, this will measure the similarity between the two images which is not desired. The only correlation between these two is the borders of the objects.

Concluding Remarks

We presented a system for tackling the joint IR-RGB image registration and fusion problem to assist firefighters in performing their missions in extreme visibility conditions. For this challenging task, we made significant progress in our attempts at targetting cases very similar to the working conditions of firefighters. Our results provide a backbone for future improvements to overcome our limitations on the segmentation of the video pairs.

Acknowledgments This study was funded by the Hasler Foundation (grant no. 16076, S.A.V.E.). We thankfully acknowledge their support.

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 Here we summarize our dataset cleaning procedure: Initially, we selected only IR frames containing humans. Using the pixels belonging to human objects, we threshold the IR frames by sequences of frames. After that, we segment the associated RGB frame with Mask-RCNN. This helps us to identify matching frames between the RGB and IR pairs. 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 into 8000 frames for the training set, 1000 frames for the validation set, and 1400 frames for the test set.

Visualization Figure 8 shows how we visualize to see if the frames are aligned to reconstruct the video after merging the warped frame (i.e., IR frame after the transformation) and RGB frame.

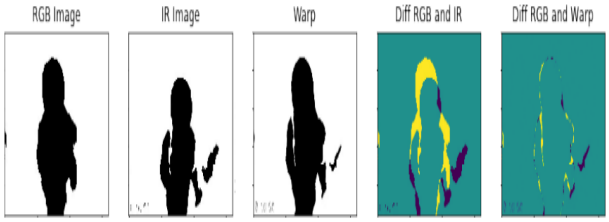

Figure 8. *Our visualization method shows the IR image, RGB image, warped image, the difference between the RGB and IR image, and the difference between the RGB and the warped image. This is useful when we work with the masked version of the images. Here we show an example of the visualization method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

Ablation Study Before applying human semantic segmentation, we investigated the registration procedure based on contours. To extract the contours of the shapes, first, we applied a Sobel filter in each direction (x, y) and computed its magnitude. The Canny filter, adaptive thresholding, and Laplacian filter followed this. When we apply the registration to this final image, the resulting image did not move significantly. This could be due to the contrast on RGB frames (see Figure 9). We observed similar results when we tried this approach on very clean images (i.e., images of contours of circles). So we understood that the problem was coming from the loss function. When the loss L_{sim} function is very flat, the loss function will have almost the same value for the cases of a slight motion. 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. We present a demonstration video on the project page.

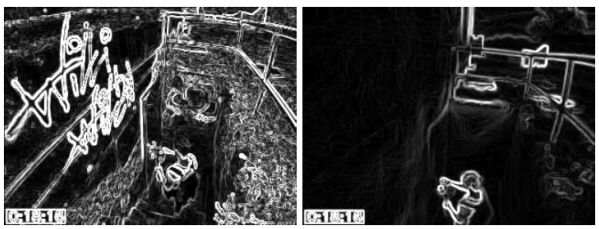

Figure 9. *Extracting contours from RGB-IR pairs: It is clear that there is much more contrast on the RGB image (on the left). Such samples can be found in our dataset resulting in poor registration based on contours.*