Channel Attention Networks

Alexei A. Bastidas
alexei.a.bastidas@intel.com
Intel AI Lab
2200 Mission College Blvd
Santa Clara, CA 95054

Hanlin Tang
hanlin.tang@intel.com
Intel AI Lab
2200 Mission College Blvd
Santa Clara, CA 95054

Abstract

Multi-band images beyond RGB are becoming popular in both commercial applications and research datasets, yet existing deep learning models were designed for academic RGB datasets. In this talk, we propose Channel Attention Networks (CAN), a deep learning model that uses soft attention on individual channels. We jointly train this model end-to-end on Spacenet, a challenging multi-spectral semantic segmentation dataset. In a comparative study, CAN outperforms previous models. We also demonstrate that CAN is significantly more robust to noise in individual bands than the other models, because the attention network allocates attention away from the noisy channels. Our proposed method marks the first step in designing deep learning algorithms specifically for multi-spectral imagery. Semantic Segmentation; Convolutional Neural Networks; Attention

1. Introduction

Multi-band data beyond the visible spectrum are becoming prevalent in computer vision, from satellite imagery [1] to infrared and depth measurements in autonomous driving [2, 3]. These data range from a few extra channels to potentially dozens for hyperspectral imaging [4]. For example, remote sensing platforms collect imagery beyond the visible spectrum, which are valuable in applications ranging from agriculture [5] to surveillance [6, 7] to land type classification [8]. Recent work has used convolutional neural networks to leverage hyperspectral imagery for image classification [8, 9, 10, 6] and semantic segmentation [11, 12, 13, 14].

However, these and other approaches typically merge the spectral bands into a single multi-channel image and employ existing deep learning models [15, 16]. On the other hand, multi-stream models are common in deep learning when combining information from different modalities, such as in video recognition [17, 18, 19] or image captioning [20]. With the exception of recent work [21, 3], these approaches have not been applied to multi-spectral data. In addition, feature maps from the different streams are fused by concatenation, leaving the network vulnerable to noise in the individual streams. It is unclear the optimal approach to exploiting multi-band data. Furthermore, existing approaches leverage models that were designed for RGB images, which may not readily transfer to this domain.

In this paper, we examine which architectures best leverage multispectral imagery for semantic segmentation. We propose Channel Attention Networks (CAN), which employ an attention network to merge multiple network streams. Our main inspiration is from previous work [22], who applied a similar mechanism to different scales of the same image. We use a similar concept, but with an orthogonal application. We split the multi-band channels into subsets, each presented to different streams. The predictions are then combined with a soft attention network, which is jointly trained with the streams.

We evaluate single stream and different multi-stream models on the multi-band Spacenet dataset for semantic segmentation. Controlled experiments demonstrate that CAN ($F_1 = 66.2$) outperforms both multi-stream concatenation ($F_1 = 64.4$) and single-stream models ($F_1 = 65.0$). More important that the modest performance increase, however, are several additional benefits. First, from perturbation experiments, the attention network is more robust to injected noise in the channels than other models, which is an important property in applications where sensors are susceptible to noise. Second, attention models are also attractive for the potential interpretability of the learned attention masks [22, 23]. However, developing quantitative measurements has proven difficult, with previous approaches relying on connecting visualizations with human intuition on individual examples. In this work, we quantify the attention weights at a population level and show that the robustness to noise is due to the network shifting attention away from the affected streams.

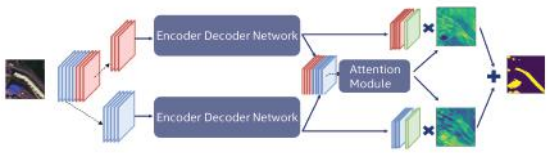

Figure 1. **Channel Attention Network.** The input bands are split into a visible stream (blue - 5 bands) and an infrared stream (red - 3 bands). In these experiments, we use the U-Net architecture [15] as the encoder-decoder network. A small attention network generates attention masks that are used to combine the stream predictions (green) to produce the final segmentation.

2. Related Work

Attentional mechanisms have been shown to enhance human perception, and play an integral role in visual intelligence [24]. Soft attention are also used in recurrent neural networks [25, 26]. In vision, recent work has also applied attention to augment residual networks to produce competitive benchmarks with fewer parameters [27].

Our proposed network is most similar to Chen *et al.* [22], who introduced the concept of merging with an attention network and applied this method to merge streams from multiple scales of the same image. They demonstrate competitive performance and increased interpretability by visualizing the attention maps. Our proposed network utilizes this same mechanism, but with a different goal. We attend to channels instead of scale, thus differentiating the visual input to each stream. We also introduce interpretability methods for interrogating the model behavior.

Recent work [3] explores multi-stream approaches for object detection, but by merging the streams with concatenation. The authors experimented with different merge points. This approach servers as one of the models we compare. In the image captioning domain, a combination of spatial and channel-wise attention was used with competitive results [28]. Their channel attention mechanism, however, is embedded in individual layers of a single stream model, and orthogonal to our proposal.

3. Methods

We chose to benchmark models on the multi-band Spacenet dataset[1], which contains satellite imagery in 8-bands from four cities (Vegas, Paris, Shanghai, and Khartoum) with labeled building segmentation. Other datasets with multiple bands exist (e.g. RGB-D [29], or RGB-IR [30], but we chose this dataset for the number of bands, and the diversity of environments. The source images consist of 10,557 images in $650 \times 640 \times 8$ GeoTIFF format. In total, the images cover 5,555 square kilometers of area, with 302,601 labeled building polygons. Prior to training, we filtered images that had $> 70\%$ blank pixels. For training, we applied a random zoom factor uniformly sampled from 100% to 200%, and rotations drawn from $[0°, 90°, 180°, 270°]$. After these data augmentations, the images were resized to 256×256 pixels.

Spacenet uses F_1 to evaluate models, with a reference implementation provided via an open source tool[2]. First a threshold T is applied to the confidence scores to produce the predicted segmentation mask. The segmentation masks are then converted to discrete polygons. The matching procedure is similar that to used in PASCALVOC [31] and ILSVRC [32]. Each ground truth polygon is compared the list of proposed polygons. True positives occur when the polygons have an intersection over union (IoU) of

$$IoU(A, B) = \frac{\text{Area}(A \cap B)}{\text{Area}(A \cup B)} > 0.5 \tag{1}$$

with the constraint that only one match – the highest IoU – can be assigned to each polygon. Otherwise, the proposed polygon is a false positive. We then compute the F_1 as:

$$F_1 = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

To match the Spacenet competition, we used $T = 0.5$ to threshold the confidence scores. We also ran experiments measuring F_1 while modifying T.

3.1. Single Stream Baseline

For our controlled comparison, we use a single stream encoder-decoder architecture based on U-Net [15]. We improve on the version used by the winner of the Spacenet competition[3] by adding batch normalization and dataset-specific augmentation (zoom, rotation). This provides a robust baseline for comparison to the below multi-stream models.

3.2. Multi-stream models

Multi-stream models allow each stream to specialize. Although widely used in multi-modal tasks, the advantage of stream segregation has not been fully explored for different channels. In our experiments, each stream uses the same encoder-decoder network as the single stream baseline. Each stream s extracts a feature map $\mathbf{X}_s$. We explore two main methods of merging these feature maps to generate the final prediction.

We limited our experiments to a visible stream $\mathbf{X}_{vis}$ and a non-visible infrared stream $\mathbf{X}_{ir}$ but this approach can be generalized to arbitrary subsets of the image channels. We chose this split to help with interpretability of the learned attention weights.

[1] See: https://spacenetchallenge.github.io/

[2] https://github.com/SpaceNetChallenge/utilities

[3] https://github.com/SpaceNetChallenge/BuildingDetectorsRound2/

3.2.1 Concatenation

In the first approach, we are motivated by prior work to concatenate the feature maps along the channel dimension [3, 11] . This merged feature map is then used to generate the final semantic segmentation prediction. Formally, the final prediction is generated as:

$$\mathbf{Y} = F\left(X_{vis} \oplus X_{ir}\right) \tag{3}$$

where F is a 1×1 Convolution layer with Pixelwise Softmax, and $\oplus$ denotes channel-wise concatenation.

3.2.2 Channel Attention Network

Our proposed channel attention network applies soft attention to learn how to weight the individual streams (Figure 1). Given feature maps $\mathbf{X}_s$ for each stream $s \in \{1, \ldots, S\}$, the final prediction is generated as:

$$\mathbf{Y} = \sum_{s=1}^{S} \mathbf{w}_s \cdot \text{Softmax}(\mathbf{X}_s) \tag{4}$$

The weights $\mathbf{w}_s$ are computed by an attention network Φ that takes as input the concatenated feature maps from the streams, and employs two convolutional layers, the first with a 5×5 kernel with K filters followed by a 1×1 Convolution with S filters and a pixel-wise softmax.

$$\mathbf{w} = \text{Softmax}\left(\Phi\left(\mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \ldots \mathbf{X}_S\right)\right) \tag{5}$$

Through joint training of the entire network, the attention network learns to weight contributions of each stream in an image-dependent manner. Note that both the attention and concatenation have approximately the same number of total parameters, since both have multiple streams with non-shared parameters.

3.2.3 Loss variants

We explored several variants of the loss for the attention model. We start with pixelwise cross-entropy loss (CE) on the final prediction $\mathbf{Y}$ used in semantic segmentation tasks. In CAN-ES, we added extra supervision to each stream by attaching a cross entropy loss to the pre-attention predictions, Softmax($\mathbf{X}_s$). The total loss is then,

$$\mathcal{L} = CE(\mathbf{Y}, \mathbf{Y}_{gt}) + \alpha \sum_{s=1}^{S} CE\left[\text{Softmax}(\mathbf{X}_s), \mathbf{Y}_{gt}\right] \tag{6}$$

The hyper-parameter α controls the contribution of the extra supervision term. The extra supervision forces each stream to learn reasonable predictions before the attention merge.

For segmentation tasks with a strong imbalance in foreground and background pixels, the dice coefficient loss can help stabilize training [33]. The dice loss is defined as

$$DICE = 1 - \frac{2 * |\mathbf{Y} \cup \mathbf{Y}_{gt}|}{|\mathbf{Y}|^2 + |\mathbf{Y}_{gt}|^2} \tag{7}$$

In our experiments denoted as CAN-DICE, our total loss is then

$$\mathcal{L} = CE(\mathbf{Y}, \mathbf{Y}_{gt}) + DICE(\mathbf{Y}, \mathbf{Y}_{gt}) \tag{8}$$

4. Experimental Results

In all experiments, we used Adam optimizer with a learning rate of 0.0005 and a batch size of 4. We trained an individual model for each city. Each city in the Spacenet dataset has different number of images and characteristics, so we optimized several hyper-parameters for each city: the number of filters in the attention network (K), the extra supervision weight (α), and the learning rate of the attention network. Results from the best performing model are shown in Table 1. We removed dropout, and used batch normalization throughout.

4.1. Performance comparisons

The single stream baseline model is drawn from the recent winner of the Spacenet challenge (xDxD), which is similar to the U-Net model [15]. We added batch normalization and data augmentations to build a better baseline. Rotation augmentations help exploit the symmetry in overhead imagery, and due to the large distribution of small objects, we also add random zoom to motivate scale-invariance (c.f. [34]). With these changes, and the addition of batch normalization, we arrive at the baseline UNET model that will be used to benchmark our attention module results. This baseline UNET model has significant improvements in the mean F_1 score across all cities compared to the Spacenet model (see Table 1, $F_1 = 65.0$ compared to $F_1 = 61.92$).

Bands beyond the visible spectrum, such as infrared, contain additional information used in many remote sensing applications. However, when we trained the baseline UNET model on individual bands, the contribution of the infrared bands was minimal. Using just the visible stream (UNET-VISIBLE) obtained an $F_1 = 64.67$, and using both streams had a minimal improvement of $\Delta F_1 = 0.33$ to $F_1 = 65.00$. (Table 1). An illustrative failure mode is shown in Figure 2, where the 3-band and 8-band models each miss different components of the building.

We explored two ways to merge the multiple streams: the commonly used concatenation approach (UNET-CONCAT), and our proposed Channel Attention Network (CAN). Concatenation reduced performance compared to

	Mean F_1	Vegas	Paris	Shanghai	Khartoum	# params
Single Stream Models						
Spacenet (xDxD)	61.92	83.66	67.33	54.33	42.37	7.9M
UNET	**65.00**	**85.17**	**69.69**	**54.83**	**50.32**	7.9M
UNET-VISIBLE	64.67	85.04	69.43	54.29	49.90	7.9M
UNET-IR	63.63	84.88	68.68	52.16	48.79	7.9M
Multi-Stream Models						
UNET-CONCAT	64.40	84.76	69.24	53.86	49.75	15.7M
CAN	52.69	75.06	57.59	48.84	29.27	15.9M
CAN-ES	**66.17**	**86.00**	**71.16**	**56.30**	**51.20**	15.9M
CAN-DICE	60.00	81.46	69.12	48.68	40.75	15.9M
CAN-ES-DICE	60.59	81.13	66.91	49.39	44.93	15.9M

Table 1. F_1 scores for each model, reported for each city in the Spacenet dataset individually, and also as the mean F_1.

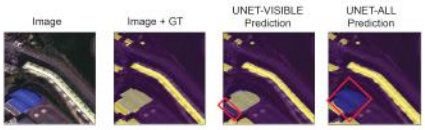

Figure 2. **Example failure image for single stream UNET model**. Segmentation shown in yellow for the ground truth (GT) and also the individual model predictions. The visible-only model (UNET-VISIBLE) correctly segments the blue roof, but misses the adjacent building (see red boxes). The all band model (UNET-ALL) has the opposite prediction.

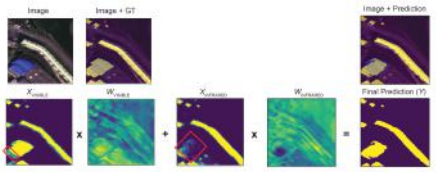

Figure 3. **Example prediction from our Channel Attention Network.** Same image as shown in Figure 2. The individual stream $\mathbf{X}_{vis}$ and $\mathbf{X}_{ir}$ predictions reproduce the failures of the single stream models in Figure 2 (see red boxes), but the attention masks $\mathbf{W}_{\text{VIS}}$ and $\mathbf{W}_{\text{IR}}$ combines the results for a more accurate prediction.

the single-stream approach ($F_1 = 64.4$). For the attention-based networks, with extra supervision (CAN-ES), performance outperformed the naive single-stream approach, reaching $F_1 = 66.12$. Interestingly, the dice coefficient loss did not improve performance.

As shown in Figure 3 using the same image as the previous figure, even though the individual streams ($\mathbf{X}_{vis}$ and $\mathbf{X}_{ir}$) miss different components of the building, the attention module appropriately combines the predictions to produce an accurate segmentation. The modest performance improvement from the attention network is not due to the number of parameters, since the concatenation (CONCAT) and attention (CAN) models have similar number of parameters, yet concatenation yielded worse results.

4.2. Model analysis

We performed several analyses to better understand the model's performance and behavior. We measured performance across object area, visualized the learned attention masks, and conducted experiments where we added noise to the visible or infrared bands with a gamma correction, and observed the model response.

4.2.1 Model struggles with small objects.

To better understand model performance, we binned the ground truth buildings by area and measured the F_1 score between the predictions and the individual bins. Similar to findings in object detection models [34, 35], our model performs poorly on small objects (Figure 4), but for different reasons. In object detection models, the network's feature downsampling reduces the representation of small objects. In addition to this issue, the loss function (pixel-wise cross-entropy loss) used in our segmentation models is not instance-aware, and penalizes misses of small objects equally to errors in the fine detail of large objects. However, the F_1 metric has the opposite incentive; it weighs missing small objects more than the fine detail of large objects. Furthermore, the metric considers IoU > 0.5 as a match, and does not credit the model for predicting closer overlaps with the ground truth.

4.2.2 Attention network is more robust.

We observed anecdotally that the attention network adapts which stream to attend to based on lighting conditions. In Figure 5, compare two images from Khartoum in **A** and **B**. For the left image, the model uses the visible spectrum to

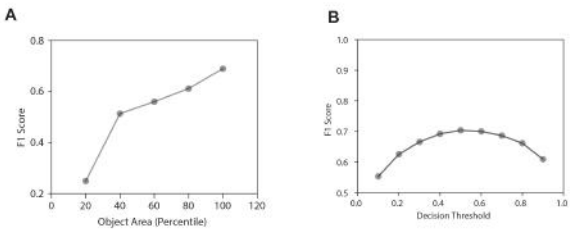

Figure 4. **Sensitivity of model to object area and decision threshold.** For the CAN-ES model on the Paris dataset, (**A**) we measured the F_1 based on the building footprint size. Despite our zoom augmentations, the model struggles with small buildings. In (**B**), the F_1 changes significantly based on the choice of threshold T applied to the confidence maps to generate the predicted maps, with $T = 0.5$ being the optimal threshold.

define the building edges, and the infrared bands to fill the interior (see attention masks). Note the heavier weighting on the visible stream. For the right image, with cloud cover reducing the usability of the visible stream, the attention masks are instead re-weighted to utilize the infrared stream.

To quantify this across the population, we applied a gamma correction with a factor γ to the test set, then measured the average attention weight to the visible versus the infrared stream. The gamma correction transforms the image via the power law expression,

$$I' = I^{\gamma} \tag{9}$$

with $\gamma = 1$ representing the original untransformed images. This transformation has the effect of modifying the overall luminance in the image. We applied the gamma correction to the test set in three conditions, transforming (1) all channels, (2) infrared channels only, and (3) visible channels only. We measured the F_1 performance using the trained models from Table 1 while varying the correction strength γ. Importantly, the tested models were not trained on the transformed data.

As shown in Figure 6A, when this luminance shift is applied across all channels, all models perform similarly. The multi-stream models have a performance advantage ($F_1 \approx 0.52$) compared to the single stream model ($F_1 \approx 0.3$) for small shifts ($\gamma = 1.0 - 1.5$). However, when individual streams are affected (Figure 6B and C), our model is significantly more robust than other models. For example, after shifting the infrared bands with $\gamma = 1.25$, our model drops to $F_1 = 0.63$, compared to UNET-CONCAT ($F_1 = 0.42$) and the single-stream UNET model ($F_1 = 0.0$).

We hypothesized that this robustness to noise was due to the attention model shifting attention away from the affected streams. To test this, we computed the average attention mask value, $|\mathbf{w}|$, for the visible and infrared streams. As shown in Figure 7, the model shifts attention to the visible

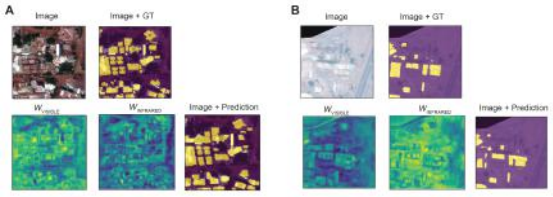

Figure 5. **Model adapts to lighting conditions.** Compare two images drawn from Khartoum. In the image from (**A**), the attention masks favor information from the visible stream. However, in (**B**), with cloud cover changing the image lighting, the attention masks are more heavily weighted towards the infrared stream.

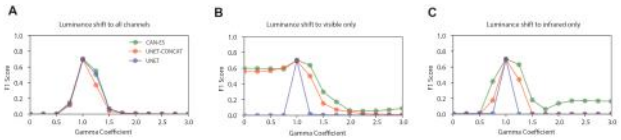

Figure 6. **Model sensitivity to noise.** We applied the gamma correction in Equation 9 to the test data for (**A**) all channels, (**B**) visible channels only, and (**C**) infrared channels only, and measured the F_1 score for our model (CAN), the concatenation model (UNET-CONCAT), and the single stream model (UNET). When noise is added to all channels, the model performs similarly, with a slight performance benefit to the multi-stream models in $\gamma = 1.0 - 1.5$ range. However, when noise is added to one of the streams, our model is significantly more robust than other models.

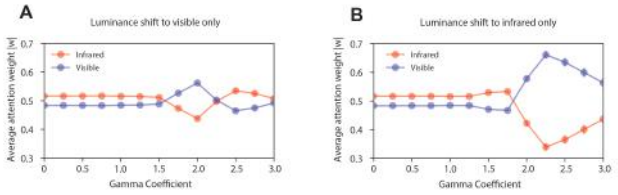

Figure 7. **Model shifts attention away from affected streams.** The average attention weight across the test set for Paris when applying the luminance shift to individual bands. For the visible stream (**A**), the model had a subtle response, but when noise is added to the infrared band (**B**), the model shifted more attention to the visible stream (blue).

stream when the infrared stream (B) is affected. Changing the luminance of the visible stream (A), yielded a smaller response.

4.2.3 Trade-off between performance and interpretability.

We also visualized the contribution of the different streams to the final prediction for the tested attention models (Figure 8). For the CAN-ES model ($F_1 = 66.2$), the extra supervi-

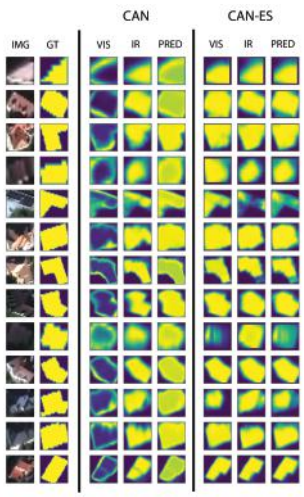

Figure 8. **Stream contributions for the attention models.** For the base model (CAN), and the model with extra supervision (CAN-ES), the contributions from the visual (VIS) and infrared (IR) streams, as well as the final prediction (PRED).

sion term forced the individual streams to also produce reasonable predictions, leading to less differentiation between the contributions of the two streams. The lower performing CAN model ($F_1 = 52.7$), however, did not have the extra supervision term. As shown in Figure 8, the visible stream clearly contributing the building edges, whereas the infrared stream was leveraged to fill the interior. We hope future work can resolve this tension between performance and interpretability.

5. Conclusion

With the coming prevalence of multi-band datasets, we took the first step of designing architectures to better exploit the additional information in ways that are more robust to noise in individual bands. We assessed several models that achieve state-of-the-art results on the Spacenet building segmentation dataset. We proposed Channel Attention Networks, which incorporate soft attention to weight channel contributions. Attention not only improves performance for segmentation, but also allows better interpretability of the network function. We demonstrate the CAN is significantly more robust to channel noise than other models. By quantitatively measuring the attention mechanism, we show that this robustness is due to the attention network allocates attention away from the affected streams.

Interpretability of the attention masks has always been appealing but difficult to quantify across the population. In our luminance perturbation experiments, we measured the overall effect on the attention weights, averaged over the dataset. In visualizing the attention masks, we also observed a trade-off between interpretability, as in connection to human intuition, with overall model accuracy. As models become deployed, avoiding this trade-off gains significance. Future work could close this gap by exploring alternatives to extra supervision that induce equally performing models.

By segregating the streams, we also allow individual streams, such as an RGB stream, to leverage pre-trained weights on RGB datasets. We speculate that our attention-based mechanism to fusing multiple streams can also be applied more generally to other multi-stream models such as video recognition and image captioning, as well as other multi-band datasets that are contributed to the community.

References

[1] Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K.: A large contextual dataset for classification, detection and counting of cars with deep learning. In: ECCV. (2016) 1

[2] Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: RGB-Infrared Cross-Modality Person Re-identification. ICCV (2017) 5390–5399 1

[3] Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral Deep Neural Networks for Pedestrian Detection. BMVC (2016) 1, 2, 3

[4] Debes, C., Merentitis, A., Heremans, R., Hahn, J., Frangiadakis, N., van Kasteren, T., Liao, W., Bellens, R., Pizurica, A., Gautama, S., Philips, W., Prasad, S., Du, Q., Pacifici, F.: Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(6) (July 2014) 2405–2418 1

[5] Lacar, F.M., Lewis, M.M., Grierson, I.T.: Use of hyperspectral imagery for mapping grape varieties in the Barossa Valley, South Australia. In: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217). (2001) 2875–2877 vol.6 1

[6] Yuen, P.W., Richardson, M.: An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. The Imaging Science Journal **58**(5) (2010) 241–253 1

[7] Uzkent, B., Rangnekar, A., Hoffman, M.J.: Aerial Vehicle Tracking by Adaptive Fusion of Hyperspectral Likelihood Maps. CVPR Workshops (2017) 233–242 1

[8] Chen, Y., Zhao, X., Jia, X.: Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief

Network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **8**(6) (July 2015) 2381–2392 1

[9] Zhao, W., Du, S.: Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing **113**(C) (March 2016) 155–165 1

[10] Makantasis, K., Karantzalos, K., Doulamis, A.D., Doulamis, N.D.: Deep supervised learning for hyperspectral data classification through convolutional neural networks. IGARSS (2015) 4959–4962 1

[11] Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W., Munteanu, A.: Hourglass-ShapeNetwork Based Semantic Segmentation for High Resolution Aerial Imagery. Remote Sensing **9**(12) (December 2017) 522–24 1, 3

[12] Cavigelli, L., Bernath, D., Magno, M., Benini, L.: Computationally efficient target classification in multispectral image data with Deep Neural Networks. In Stein, K.U., Schleijpen, R.H.M.A., eds.: SPIE Security + Defence, SPIE (October 2016) 99970L–12 1

[13] Kemker, R., Kanan, C.: Self-Taught Feature Learning for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing **55**(5) (March 2017) 2693–2705 1

[14] Aytaylan, H., Yuksel, S.E.: Semantic segmentation of hyperspectral images with the fusion of LiDAR data. IGARSS (2016) 2522–2525 1

[15] Ronneberger, O., Fischer, P., Brox, T.: U-Net - Convolutional Networks for Biomedical Image Segmentation. MICCAI **9351**(Chapter 28) (2015) 234–241 1, 2, 3

[16] He, K., Zhang, X., Ren, S., 0001, J.S.: Deep Residual Learning for Image Recognition. CVPR (2016) 770–778 1

[17] Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. NIPS **cs.CV** (2014) 1

[18] Jiang, Z., Rozgic, V., Adali, S.: Learning Spatiotemporal Features for Infrared Action Recognition with 3D Convolutional Neural Networks. CVPR Workshops (2017) 1

[19] Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional Two-Stream Network Fusion for Video Action Recognition. CVPR (2016) 1

[20] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4) (September 2016) 652–663 1

[21] Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M.A., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. IROS (2015) 1

[22] Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to Scale - Scale-Aware Semantic Image Segmentation. CVPR (2016) 3640–3649 1, 2

[23] Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: ICLR. (2015) 1

[24] Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience **2** 194 EP – 2

[25] Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.: Recurrent Models of Visual Attention. NIPS (2014) 2

[26] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, Attend and Tell - Neural Image Caption Generation with Visual Attention. ICML (2015) 2

[27] Wang, F., Jiang, M., 0006, C.Q., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual Attention Network for Image Classification. CVPR (2017) 6450–6458 2

[28] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: SCA-CNN - Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. CVPR (2017) 6298–6306 2

[29] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. CVPR (2012) 2

[30] Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection - Benchmark dataset and baseline. CVPR (2015) 1037–1045 2

[31] Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision **111**(1) (June 2014) 98–136 2

[32] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.F.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3) (2015) 211–252 2

[33] Milletari, F., Navab, N., Ahmadi, S.A.: V-Net - Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 3DV (2016) 565–571 3

[34] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C.Y., Berg, A.C.: SSD - Single Shot MultiBox Detector. In: ECCV. (2016) 3, 4

[35] Ren, S., He, K., Girshick, R.B., 0001, J.S.: Faster R-CNN - Towards Real-Time Object Detection with Region Proposal Networks. NIPS (2015) 4