Hyperspectral Data To Relative Lidar Depth: An Inverse Problem For Remote Sensing

Savas Ozkan, Gozde Bozdagi Akar
Middle East Technical University
Department of Electrical and Electronics Engineering
Ankara, Turkey
ozkan.savas@metu.edu.tr

Abstract

Hyperspectral data provides rich information about a scene in terms of spectral details since it encapsulates measurements/observations from a wide large range of spectrum. To this end, it has been used in different problems mostly related to identification and detection processes. However, the main limitation arises for the accessibility of data. More precisely, there is no sufficient amount of hyperspectral data available compared to visible range data for trainable models. In this paper, we tackle an inverse problem to estimate the relative lidar depth from hyperspectral data. To solve its limitation, we integrate semantic information existed in data with supervised labels to decrease the possibility of parameter overfitting. Moreover, details of the output responses are enhanced with Laplacian pyramids and attention layers in which the model makes predictions from each subsequent scale instead of a single shot prediction from the top of the model. In our experiments, we use the 2018 IEEE GRSS Data Fusion Challenge dataset. From the experimental results, we prove that use of hyperspectral data instead of visible range data improves the performance. Moreover, we show that results are significantly improved if a sparse set of depth measurements is used along with hyperspectral data. Lastly, the integration of semantic information to the solution yields more stable and better results compared to the baselines.

1. Introduction

The solution of an inverse problem aims to obtain unknown information from a limited set of observations. For instance, to recover high-resolution (HR) image from its low-resolution (LR) versions or to transform data to a different domain, content and neighborhood embedding of data can be prominent clues that should be exploited in the solutions with advanced models [13, 31, 32]. For this purpose, performance highly depends on the distinctive power of observations and models.

The underlying idea behind these problems relies on the assumption of manifold learning that ultimately retrieves target output data by learning input counterparts to represent the overall manifold space with example-pairs (i.e. training data) [13, 1, 7, 23]. For this reason, manifold space should be modeled precisely for high performance. In the recent years, deep learning models (i.e. fully convolution neural networks) obtain state-of-the-art performance due to their large body of learnable parameters and higher-order non-linearity captured by network models [22, 11]. Hence, the solutions in the literature are frequently based on the variants of this phenomena.

For remote sensing domain, similar applications that we mentioned can be found [10, 4]. In particular, sensors used for this domain can provide richer information about a scene (i.e. hyperspectral/multispectral data). However, the main drawback is that the accessibility to these sensor data is either limited or quite expensive. Moreover, their dimensionality and sensitivity to noise aggravates to obtain a stable solution from data [3, 14, 19]. Therefore, models should be robust to these cases and take these limitations into consideration in the learning step especially for inverse problems.

In the scope of the paper, depth estimation refers to determine the relative distance from sensors. In the literature, this can be achieved by using either binocular [2] or monocular [8, 24, 9] clues exhibited from data. Texture (i.e. content), lighting and shading are some of features that can be leveraged in course of obtaining relative depth information. Indeed, complex sensor outputs can be more robust to extract these features from data.

In this paper, we tackle an inverse problem (i.e. domain transformation) that aims to estimate lidar depth data from a single shot hyperspectral image. As recalled, content information and neighborhood embedding are essential and need to be learned automatically from data for high

performance. For this purpose, we exploit fully convolution neural networks (FCN) by presenting several contributions/improvements to the architecture as well as parameter optimization. More precisely, use of semantic information of hyperspectral data (i.e. grass, trees, buildings etc.) by training an auxiliary NN model (i.e. has shared parameters with the model of the main task) ultimately improves the results. Note that this step intuitively corresponds to extraction of high-level semantic information from data and this will be called as content of data for the rest of the paper. By this way, the proposed solution has a similar notion with manifold learning since manifold space is transformed by exploiting both building atoms of hyperspectral data (i.e. endmembers and categories [19]) and objective of domain transformation. To this end, this reduces the dependency to a large number of data (i.e. parameter overfitting) and improves the robustness to weakness of data (i.e. dimensionality and weakness to noise) that will be summarized in the following sections.

To ease the understandability, contributions/improvements in the paper are presented as follows:

- We prove that use of hyperspectral data can be more reliable for inverse problems, since it allows us to capture richer observations (i.e. texture, lighting and shading etc.) about a scene compared to RGB sensors.

- Besides its rich information, dimensionality and accessibility are main factors which adversely affect the stability of learning. For this purpose, semantic information of hyperspectral data is accounted with an auxiliary NN model. To this end, high-level latent representations are estimated from hyperspectral data and encapsulated in the FCN model which is dedicated to solve the inverse problem for an optimum transformation from hyperspectral data to lidar depth data.

- We present several improvements to the structure of NN model as well as its optimization step. As explained [16], estimated object boundaries with FCN are generally smooth (i.e. even if skip-connections re used [20]) and lower evaluation performance can be observed due to these characteristics of FCN models. As a remedy, we output lidar depth predictions at different scales and compute one final prediction at the top of the network by fusing the responses from each subsequent scale. Ultimately, the error is propagated from different scales instead of only top-layer of the network. Moreover, attention layers are employed to increase the selectivity of parameters and the global context information [29]. Lastly, parameters are regularized with a L1 penalty which encourages responses to be zero practically that correspond to the background parts of data (i.e. depth will be zero for furthest objects). To this end, generalization capacity for responses (i.e. distribution of responses) is eventually improved.

- In the experiments, we show that by providing a sparse set of lidar samples from the scene, depth predictions can be significantly improved. This shows that the proposed model can be also used as a refinement/auxiliary technique in case of lacking depth samples.

The related literature and solutions are summarized in Section 2. Section 3 presents the details of the proposed method with the optimization procedure. Lastly, experimental results and final remarks are explained.

2. Related Work

There is a large body of works that aim to achieve an optimum solution for various inverse problems in the literature. Indeed, the literature for remote sensing domain is relatively narrow compared to visible range domain. Therefore, instead of focusing only one domain, we will survey the studies mainly related to our problem/solution from both domains.

Manifold Learning: As indicated, the definition of inverse problem covers a wide range of studies from super resolution to domain transformation [13, 1, 7, 31, 10, 4]. Even if their formulations are different, a similar notion is exploited to handle the problems. More precisely, [13, 31] summarizes that learning-based methods are built on the assumption of manifold learning where optimum solutions ultimately correspond to retrieve common pattern transformations from input-output example pair space in the training data. Note that more input-output example pairs eventually tend to learn more reliable manifold space [13].

Similarly, the number of parameters in the model and parameter convergence are two critical factors since filtering operations (i.e., matrix multiplications [6, 22, 19]) are intuitively equal to affine transformations that form the manifold space based on these example pairs. For this purpose, FCNs draw lot of interests in the recent years due to their large parameter sizes and non-linearity features.

Advantage of Supervised Learning: For an unsupervised setup, holes/irregularity can be observed in the learning step which can generate unstable results as explained in [18, 28]. The main reason is that Euclidean distance (i.e. L2 norm) does not reflect the semantic similarities between example pairs. Moreover, [25] observes that to fill these holes/irregularities, dataset size need to be high since the distributions of example pairs become denser and this closes the semantic similarities in feature space.

To solve the issue, use of supervised information (i.e. labels related to common patterns of data) is the most straightforward solution. By this way, the dependency to large scale

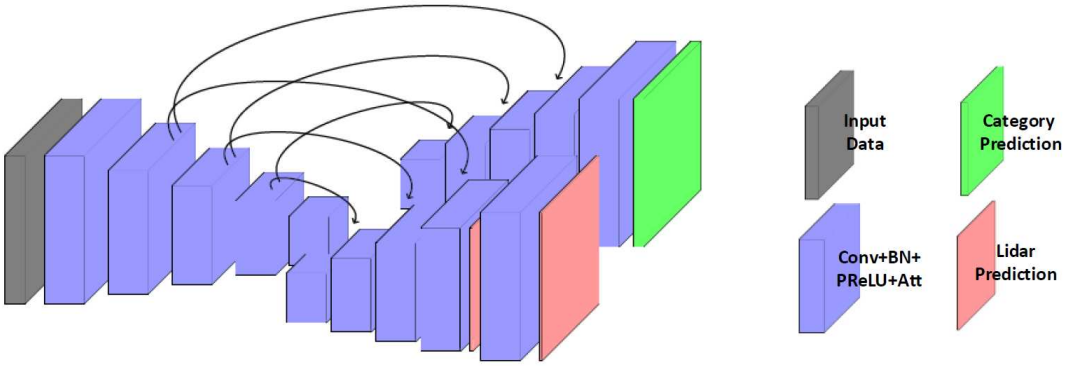

Figure 1. The flow of the proposed method. Each distinct color denotes different components of the model as summarized. Each filter size is set to $3 \times 3 \times 64$ and stride is equal to 2 for each convolution in the encoder module.

data is mitigated [18, 22] and more reliable manifold space can be estimated.

Depth Estimation From A Single Shot Image: For depth estimation, the phenomena of over-segmentation (i.e. super pixels) is utilized to maintain the cardinality of data especially for their spatial information. By this way, 3D location and orientation of local planes are estimated from data containing RGB-laser scans pairs [21]. Moreover, Markov Random Field (MRF) is combined to refine the predictions. Similarly, [15] uses FCN model to learn the parameters in this transformation. However, [9] explains that predictions of these models are generally weak by which global context of input does not preserved in the learnable models.

The most similar works to our method in the literature are based on the assumption that a single shot depth estimation and semantic segmentations are tied to the property of perspective geometry [12, 27, 30]. Hence, these two information can be used together to obtain more straightforward solutions and it reduces to the number of possible solutions. Note that this assumption is only studied for visible range data.

Lastly, [17, 5] show that depth estimation can be further improved by using a sparse set of depth information with RGB data from a scene.

3. Proposed Method

Formally, an inverse problem transforms input data $\mathbf{X}$ to target output data $\mathbf{Y}$ by optimizing a function $f(.)$ where the relation corresponded to $\mathbf{Y} \approx f(\mathbf{X})$ is achieved with a minimum error. Note that this function makes predictions by exploiting pre-given observations from input-output data examples.

For the estimation of lidar depth data (i.e. $\mathbf{Y} \in \mathbb{R}^{\mathrm{N} \times \mathrm{M}}$) from hyperspectral data (i.e. $\mathbf{X} \in \mathbb{R}^{\mathrm{N} \times \mathrm{M} \times \mathrm{D}}$) (Here, N, M and D denote the spatial dimensions and spectral channels for hyperspectral data respectively.), semantic content and neighborhood reasoning are two distinct clues that can be used to optimize the parameters of the function $f(.)$.

For this purpose, the objective of the proposed method is to extract content of data and neighborhood relations as much as possible from data with an end-to-end learning manner. To this end, we compute semantic information and neighborhood relations with two FCN models that intuitively have a shared parameter set. Therefore, both semantic information and neighborhood reasoning are maintained in the solution. The flow of the model is illustrated in Figure 1.

We divide this section into two parts. First, we will mention the FCN architecture and our contributions to improve the performance. Second, the idea behind use of content of hyperspectral data will be explained. Moreover, we will formulate the loss function that is used to optimize the parameters of the models.

3.1. Hyper2Lidar Network

The proposed network architecture (i.e. Hyper2Lidar) aims to transform hyperspectral data to lidar depth data. For this purpose, we use FCN architecture which takes hyperspectral data and predicts lidar data. It formulates the solution with two modules as encoder and decoder (details will be provided). In course of this transformation, trainable parameters are optimized by minimizing the error between ground truth and prediction:

$$\mathcal{L} = ||\mathbf{Y} - \hat{\mathbf{Y}}||_2 \tag{1}$$

$\mathcal{L}$ is the reconstruction loss function by using L2 norm (i.e. Euclidean distance).

Ideally, a FCN architecture used for domain transformation consists of two modules as encoder $e(.)$ and decoder $d(.)$ (i.e. $f(.) \approx d(e(.))$). Encoder part computes latent features related to neighbourhood embedding and content of data based on hidden responses. On the other hand, decoder network fuses these features in order to reconstruct output data with high accuracy. To this end, parameters learn the transformation from given example-pairs.

However, current state-of-the-art architectures have apparent weakness in which object boundaries of the responses can be smooth. Either by revisiting the architecture [20] or by adding an additional loss function [16], this weakness is studied in the literature.

In the proposed solution, a full sized hyperspectral image is similarly given as an input and high-level features related to texture, lighting, shadows etc. are extracted in the encoder part. To boost the visual quality of decoder part, we reconstruct output predictions at each subsequent layer by computing a reconstruction error as follows:

$$\mathcal{L}_{reg} = \sum_{k=1}^{K} \alpha_k ||\mathbf{Y}_k - \hat{\mathbf{Y}}_k||_2 + \beta|\hat{\mathbf{Y}}_K| \tag{2}$$

Note that along with L2 norm that minimizes the reconstruction loss between ground truth and prediction at each subsequent layer ($K = 2$), L1 penalty is utilized to regularize the loss function to enforce the predictions to be zero. β is set to 0.01 throughout the paper. Moreover, α_1 and α_2 are equal to 0.25 and 1.0 respectively that are proportional to spatial resolutions of the prediction outputs.

Moreover, in conventional FCN architectures, convolution kernel, batch normalization and parametric ReLU are used at l^{th} layer [11]. Although we follow this conventional way, an attention gate is added to their end. By this way, the outputs of the activation function are rescaled based on the spatial attention of pixel channels. Note that this step has a similar notion to the global content of responses as highlighted in [9] and selectivity of filters are increased by regularizing the responses before the convolution operation similar to batch normalization:

$$\mathbf{o}_l = \mathbf{o}_l \odot \psi_l(\mathbf{o}_l) \tag{3}$$

Here $\psi_l(.)$ denotes the attention gate that is composed of convolution kernels (i.e. with activations) and it computes attention scores for individual input channels according to spatial mean of input data. These scores are also rescaled with tanh activation function. Moreover, $\odot$ denotes the element-wise multiplication operator. Note that we tested the variants of attention gate in the experiments [26] yet no significant change is observed.

3.2. Content of Hyperspectral Data

Hyperspectral imaging is a technique to collect measurements about a scene. These measurements ultimately correspond to responses of a sensor array operating in different electromagnetic spectrum. To this end, rich information about a scene can be obtained.

However, there are several theoretical and practical limitations for hyperspectral data. One of which is that its operational cost is expensive, thus accessibility of data is constrained. Moreover, sensitivity to noise, spectral variability and high-correlation between channels are some of other obstacles that should be handled [19]. Lastly, for an unsupervised setup (e.g. inverse problem), high-dimensionality and lack of data lead to holes/irregularity in the learning step of manifold space which can generate unstable results. For all these reasons, direct use of hyperspectral data for inverse problem ultimately underestimates the solution.

Hence, we opt to integrate content of hyperspectral data with high-level representations and the objective of the inverse problem. Benefits of the method can be explained in twofold. First, supervised data is used to reduce the dimensionality by highlighting structural predictions about data (i.e. the idea behind hyperspectral unmixing). Therefore, manifold space can encapsulate semantic information about data at the end. Second, inter-obstacles of data are also mitigated since the network learns to cluster manifold space by accounting different obstacle such as different noise-levels and illumination conditions (i.e. spectral variability) with supervised labels.

Similar to Hyper2Lidar network, a different FCN model $g(.)$ is utilized to extract content patterns related to data. For this purpose, we use pixel-level semantic annotations $\mathbf{Z} \in \mathbb{R}^{\mathrm{N}\times\mathrm{M}\times 20}$ that are already available in the dataset along with hyperspectral and lidar data (Note that 20 semantic categories are available in the dataset). To this end, this model solves a pixel-level classification problem that is formulated as $\mathbf{Z} = g(\mathbf{X})$. To optimize the parameters, softmax cross-entropy is utilized:

$$\mathcal{L}_{cls} = -\mathbf{Z} log(\hat{\mathbf{Z}}) \tag{4}$$

Indeed, to capture high-level semantic information also in Hyper2Lidar network, encoder parts $e(.)$ of both models have a shared parameter set. By this way, content of data that is highlighted in the classification problem can be taken into account in course of transformation of hyperspectral data to lidar data.

3.3. Parameter Learning

For the decoder part of each model (i.e. either classification or reconstruction model), corresponding loss functions can be directly used to optimize the parameter sets. However, since a shared parameter set is used for the encoder part for both models, two loss functions are accounted as follows:

$$\mathcal{L} = \mathcal{L}_{reg} + 0.01\mathcal{L}_{cls} \tag{5}$$

Note that classification loss is weighted with a lower coefficient due to the fact that loss of the inverse problem should be higher related to the objective of this work. Adam solver is utilized by exploiting a stochastic mini-batch scheme. Mini-batch size is set to 16 and spatial resolution is fixed to 64×64 in the training step. Lastly, we set the learning

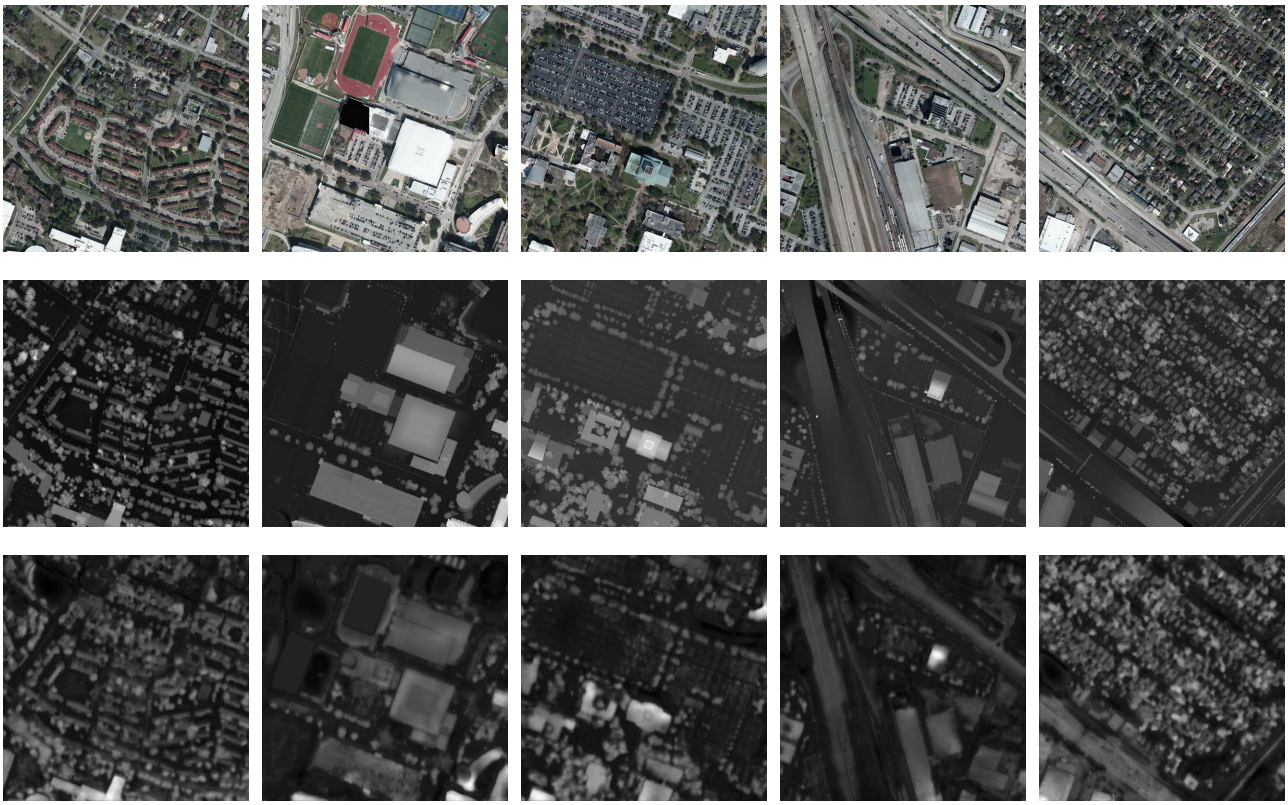

Figure 2. Qualitative results on 2018 IEEE GRSS Dataset for hyperspectral data and semantic information. RGB data, ground truth depth data and estimated depth data are illustrated in each columns.

rate and the number of training iterations to 0.001 and $10K$ respectively.

4. Experiments

In this section, we will explain the experimental results conducted to show the performance of our contributions. First, we will mention the details of the dataset that we used in the experiments. Later, impacts of expansions will be discussed in detail.

4.1. 2018 IEEE GRSS Data Fusion Challenge Dataset

Dataset used in our experiments is initially proposed for 2018 IEEE GRSS Data Fusion Challenge[1]. It is composed of registered hyperspectral, lidar and RGB data with 20 land cover classes.

In particular, hyperspectral data covers a 380-1050 nm spectral range with 48 bands and 1-m ground resolution. Ground resolutions of lidar data and RGB data are 50-cm and 2-cm respectively. Their spatial resolutions are rescaled as in hyperspectral data throughout the experiments. Also, lidar data is normalized to [0, 1] and all predictions are computed based on these assumptions.

The overall dataset is splitted into 14 pieces where each piece contains a 512×512 image to ease the operation of the dataset. To exploit symmetries and rotations in convolution kernels, data is also augmented in different rotations (i.e. in 4 main directions).

Moreover, performance is evaluated with Root Mean Square Error (RMSE) metric to compare prediction with ground truth annotation:

$$RMSE(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{L}\sqrt{\|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2}. \tag{6}$$

Lastly, tests are repeated 20 times. Therefore, mean and standard deviation are reported in the paper.

4.2. Impact of Hyperspectral Data

To understand the impact of RGB and Hyperspectral data for lidar depth estimation, we first evaluate the model trained by each data type. After that, the impact of semantic information is also reported.

In Table 1, lidar depth predictions are reported for three different data models (i.e. Hyper-Smnt indicates hyperspectral with semantic data). Moreover, the impact of Laplacian

[1] http://www.grss-ieee.org/community/technical-committees/data-fusion/2018-ieee-grss-data-fusion-contest/

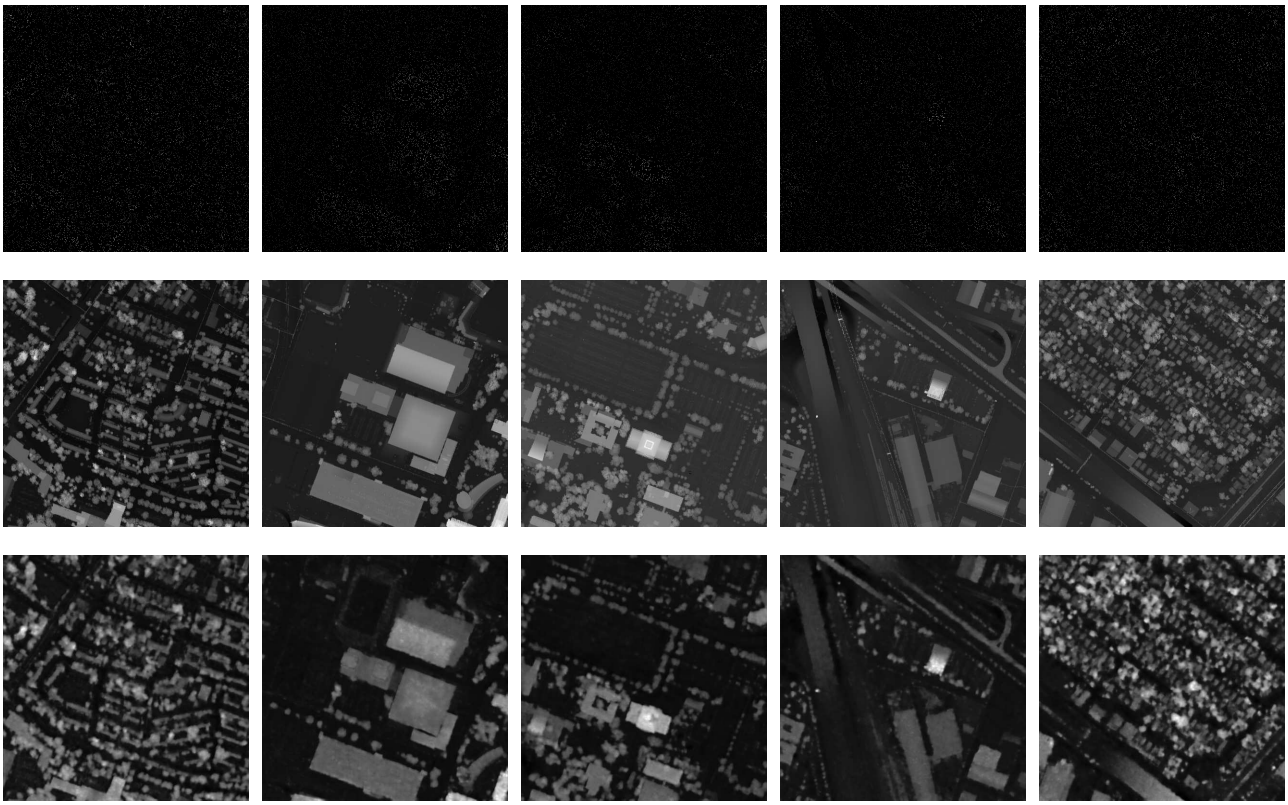

Figure 3. Qualitative results on 2018 IEEE GRSS Dataset for hyperspectral data and sparse depth data. Sparse depth data, ground truth depth data and estimated depth data are illustrated in each columns. The proposed method refines and improves the quality of predictions drastically.

Table 1. RMSE performance to evaluate the impact of hyperspectral data with semantic content. Note that RMSE performance is normalized with $\times 10^{-2}$

	RGB	Hyper	Hyper-Smnt
Att-Lap w/o	1.46 ± 0.27	1.38 ± 0.15	1.34 ± 0.13
Att-Lap	1.41 ± 0.26	1.35 ± 0.16	1.29 ± 0.11

Table 2. RMSE performance to evaluate the impact of sparse depth data for RGB and Hyperspectral data. Note that RMSE performance is normalized with $\times 10^{-2}$

$Data(\%)$	RGB	Hyper
0.1	0.31 ± 0.18	0.30 ± 0.11
0.07	0.38 ± 0.21	0.33 ± 0.15
0.05	0.57 ± 0.38	0.46 ± 0.19
0.01	1.21 ± 0.34	1.11 ± 0.21

and attention function is also computed. From the experimental results, hyperspectral data yields better performance compared to RGB data for lidar depth estimation. The reason is that hyperspectral data encapsulates more meaningful clues about the scenes as explained throughout the paper. Moreover, performance variations significantly reduce for hyperspectral data.

For the impact of semantic data, RMSE performance is improved compared to both RGB and hyperspectral data. As explained, accessibility of hyperspectral data, dimensionality and noise-prone characteristics are some of the major factors that ultimately affect performance. By incorporating semantic information, these limitations are mitigated. Some of the visual results are illustrated in Figure 2.

4.3. Impact of Sparse Depth Data

For this experiment, the objective is to evaluate the impact of sparse depth data provided along with input data. This assumption is meaningful since there can be more than one possible solution for inverse problems and sparse samples can inherently reduce the ambiguity in the solution. To this end, the proposed methods takes sparse depth data and RGB/Hyperspectral data as inputs and estimates depth predictions.

From the experimental results reported in Table 2, sparse depth data drastically improves depth estimation performance as expected. More precisely, use of 0.05% sparse

depth samples reduces RMSE from 1.29 to 0.46. The main reason is that sparse depth data gives clues about the scene and the proposed method refines its predictions by exploiting these insights. To this end, more confident solutions can be attained. In Figure 3, the ground truth and prediction depth data are illustrated respectively. On the contrary to the baseline models, depth prediction for stadium (second column) is significantly enhanced under the guidance of sparse data.

Moreover, we repeat this experiment for both RGB and hyperspectral data. Similarly, due to the rich insights collected from scenes, the combination of sparse depth and hyperspectral data yields higher performance while boosting confidences for the predictions.

5. Conclusion

In this paper, we tackle an inverse problem to estimate lidar depth from hyperspectral data. For this purpose, we present several contributions and observations about the problem throughout the paper. More precisely, use of hyperspectral data instead of RGB data ultimately improves performance by which it provides richer information about scenes which is meaningful due to texture, lighting and shading clues. Moreover, semantic content of data is imposed to the inverse problem by exploiting label annotations (i.e. endmembers) with an auxiliary NN model. By this way, limitations of data are mitigated as explained in detail. Lastly, prediction results are enhanced by providing sparse depth data along with input data and optimum performance saturates quickly. This shows that the proposed model can be used as a refinement step as well.

6. Acknowledgments

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research. Moreover, the authors would like to express their gratitude to Mehmet Efendioglu and Feray Oztoprak who have helped in the preparation of data.

References

[1] E. Agustsson, R. Timofte, and L. Van Gool. Anchored regression networks applied to age estimation and super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1643–1652, 2017.

[2] R. S. Allison, B. J. Gillam, and E. Vecellio. Binocular depth discrimination and estimation beyond interaction space. *Journal of Vision*, 9(1):10–10, 2009.

[3] Y. Chang, L. Yan, H. Fang, and C. Luo. Anisotropic spectral-spatial total variation model for multispectral remote sensing image destriping. *IEEE Transactions on Image Processing*, 24(6):1852–1866, 2015.

[4] Y. Chang, L. Yan, T. Wu, and S. Zhong. Remote sensing image stripe noise removal: From image decomposition perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7018–7031, 2016.

[5] Z. Chen, V. Badrinarayanan, G. Drozdov, and A. Rabinovich. Estimating depth from rgb and sparse sensing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 167–182, 2018.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[7] A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[8] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[9] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[10] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, (99):1–19, 2018.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[12] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014.

[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[14] J. Li, Q. Yuan, H. Shen, and L. Zhang. Noise removal from hyperspectral image with joint spectral–spatial distributed sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9):5425–5439, 2016.

[15] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[16] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[17] F. Mal and S. Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[18] M. Norouzi and D. M. Blei. Minimal loss hashing for compact binary codes. In *Proceedings of the 28th international*

conference on machine learning (ICML-11), pages 353–360. Citeseer, 2011.

[19] S. Ozkan and G. B. Akar. Improved deep spectral convolution network for hyperspectral unmixing with multinomial mixture kernel and endmember uncertainty. *arXiv preprint arXiv:1808.01104*, 2018.

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[24] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 673–680, 2013.

[25] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[26] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[27] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

[28] L. Zhang, L. Wang, and W. Lin. Semi-supervised biased maximum margin analysis for interactive image retrieval. *IEEE Transactions on Image Processing*, 21(4):2294–2308, 2012.

[29] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[30] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.

[31] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.

[32] A. Zia, J. Zhou, and Y. Gao. Relative depth estimation from hyperspectral data. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2015.