The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Multispectral Transfer Network: Unsupervised Depth Estimation for All-Day Vision

Namil Kim,[*1,2] **Yukyung Choi,**[*1,3] **Soonmin Hwang,**[1] **In So Kweon**[1]
[1]Korea Advanced Institute of Science and Technology (KAIST), Korea
[2]NAVER LABS Corp., Korea [3]Clova, NAVER Corp., Korea
http://multispectral.kaist.ac.kr

Abstract

To understand the real-world, it is essential to perceive in all-day conditions including cases which are not suitable for RGB sensors, especially at night. Beyond these limitations, the innovation introduced here is a multispectral solution in the form of depth estimation from a thermal sensor without an additional depth sensor. Based on an analysis of multispectral properties and the relevance to depth predictions, we propose an efficient and novel multi-task framework called the *Multispectral Transfer Network* (MTN) to estimate a depth image from a single thermal image. By exploiting geometric priors and chromaticity clues, our model can generate a pixel-wise depth image in an unsupervised manner. Moreover, we propose a new type of multitask module called *Interleaver* as a means of incorporating the chromaticity and fine details of skip-connections into the depth estimation framework without sharing feature layers. Lastly, we explain a novel technical means of stably training and covering large disparities and extending thermal images to data-driven methods for all-day conditions. In experiments, we demonstrate the better performance and generalization of depth estimation through the proposed multispectral stereo dataset, including various driving conditions.

Introduction

Depth estimation from a single RGB image is a fundamental problem in computer vision. Many industrial (Google, Tesla etc.) and academic approaches (Geiger et al. 2013; Cordts et al. 2016) have utilized the depth entailing complementary sources of RGB images. In recent years, deep learning-based approaches have advanced significantly for single-image depth estimations. Because these data-driven approaches require large amounts of RGB-D data, supplementary depth sensors are typically used to capture the ground truth accurately. In outdoor scenarios, a 3D laser scanner is usually used to capture depth data. However, such devices are limited in terms of the range and resolution and usually fail when used with specular or transparent objects such that depth measurements with them do not capture detailed variations in the images. Moreover, due to the physical limitation of RGB sensors, these measurements have yet to be broadly applied to various or less well-lit conditions, such as nights or sunsets and sunrises. Hence, this led to the question of *how it would be possible to estimate dense and accurate depth images all day*.

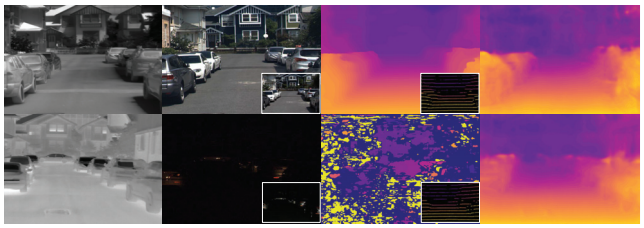


Figure 1: The result of an accurate depth estimation using a *Multispectral Transfer Network* (MTN) in both day (top) and night (bottom) conditions. From left to right: input thermal images, RGB stereo pairs, depths from RGB stereo/Velodyne HDL-32E, and our results. The proposed method can predict high-quality pixel-wise depths at night compared to the depths from the RGB stereo/Velodyne.

We believe that the answer will rely on the use of *alternatives to RGB sensors*. Among the promising options, the thermal sensor has a strong advantage if used to capture images in the world, as this type is less affected by light changes under highly lit and dark conditions. Therefore, various thermal sensors have been increasingly used in modern robotics and computer vision research on all-day recognition. Recently, multispectral[1] approaches (Hwang et al. 2015; Choi et al. 2015; Jingjing et al. 2016; Treible et al. 2017) have demonstrated some degree of correspondence between RGB and thermal images as well as complementary information. Spectral images from both types of sensors share global contexts such as silhouettes, boundaries, and structures regardless of the loss of fine visual details in the thermal images, even if their spectra are wholly different. From these observations, we argue that thermal images can be incorporated into RGB-based depth estimation methods.

There are two main challenges when estimating depths from thermal images. The first is the scarcity of large-scale multispectral datasets for depth estimations. Therefore, we created a new multispectral stereo dataset which includes co-

*Authors contributed equally, listed in alphabetical order

[1]Denote **thermal** as Long Wavelength InfraRed (LWIR) and **multispectral** as RGB and thermal spectrum.

aligned multispectral pairs and 3D measurements. Another issue is the different physical principles between the RGB and thermal domains. Depth from RGB includes certain visual details which are infeasible to reconstruct from thermal images. To alleviate this problem, our approach is based on the unsupervised deep learning approaches (Garg et al. 2016; Godard, Aodha, and Brostow 2017; Zhou, Brown, and Lowe 2017) using the L2 loss, though this is apt to produce blurry images. Due to the assumption that pixels are drawn using a single Gaussian distribution (Mathieu, Couprie, and LeCun 2016), our model is trained to induce degraded uncorrelated details.

We name the proposed method the *Multispectral Transfer Network* (MTN). It can generate the RGB-based depth from a single thermal image based on the spectral/geometric relationships between multispectral domains. Technically, we introduce three novel contributions for accurate depth estimations. First, we introduce *efficient multi-task learning* for depth estimation using the concept of chromaticity, which is well known as the main feature in various tasks related to depth. The proposed multi-task learning approach can improve the performance without additional annotated data or measurements (Eigen and Fergus 2015). Secondly, compared to general skip-connections (Long, Shelhamer, and Darrell 2015; Bell et al. 2016; Hariharan et al. 2015; Farabet et al. 2016), we propose the *Interleaver* module which simultaneously encodes the finer details of the lower layers and chromaticity features onto skip-connected activation. The proposed module supplements the standard CNN architecture without directly sharing the feature layers for multi-task learning. Lastly, we provide several technical considerations as *an adaptive scaled sigmoid* to cover large disparities during training and *photometric correction* to handle thermal contrast variations at different times for all-day depth estimation.

Related Work

Multispectral Vision

Multispectral vision has been proposed for robust recognition in all-day environments. With the promising thermal sensors, Hwang *et al*. (Hwang et al. 2015) proposed a multispectral benchmark using a beam splitter to capture optically aligned multispectral image pairs. From this milestone work, the multispectral approach was extended to various applications, such as place recognition (Choi et al. 2015), image enhancement (Choi et al. 2016), visual odometry (Poujol et al. 2016) and object detection (Jingjing et al. 2016) for all-day recognition. However, most of these works only focused on fusing multispectral pairs through image concatenation or stacking multiple DNN models. Compared to the RGB-D domain (Gupta, Hoffman, and Malik 2016; Hoffman, Gupta, and Darrell 2016), it is still an open question as to how multispectral images can be properly combined to obtain optimal synergy. Recently, there have been several works related to colorization with RGB and near-infrared (NIR) (Limmer and Lensch 2016; Patricia L. Suarez and Vintimilla 2017) for multispectral transfer learning.

In the paper, we attempt to bridge the gap between RGB and thermal images, which have completely different spectrums. To do this, we propose an unsupervised multispectral framework that transfers depth information from RGB pairs to thermal input images for all-day recognition, as well as a the large-scale multispectral stereo dataset for real-world scenarios.

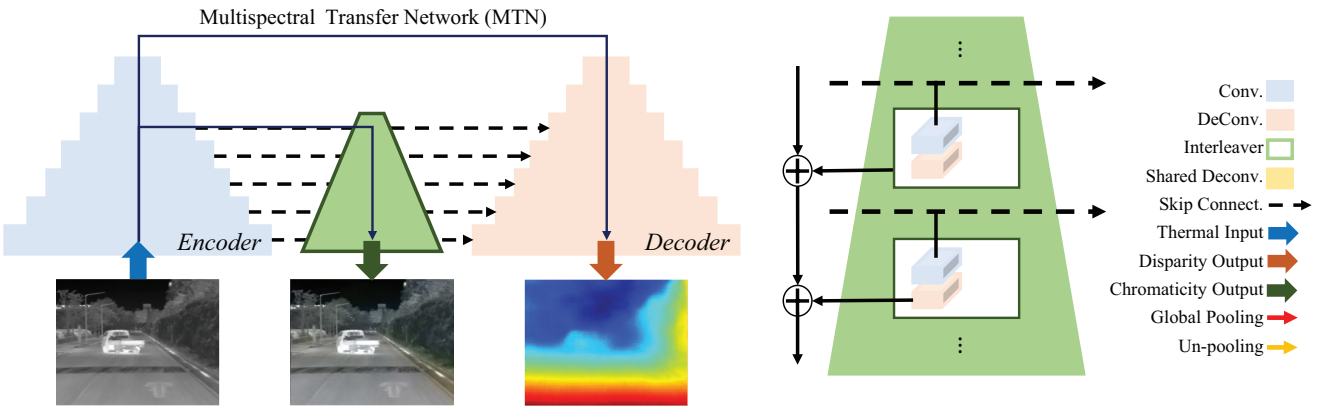
Learning-based depth estimation

Supervised data-driven methods which adapt a CNN to general depth predictions (Liu et al. 2015; Ladicky, Shi, and Pollefeys 2014), multi-scale predictions (Eigen, Puhrsch, and Fergus 2014), multi-task learning (Eigen and Fergus 2015), CRFs (Li et al. 2015), and robust objective functions (Laina et al.) outperform conventional approaches. While supervised methods can generate better results, the preparations necessary to handle a large amount of ground truth data are not trivial, especially in outdoor scenarios. To overcome these limitations, unsupervised methods have recently been presented. These typically use the geometric properties of a single image or rectified stereo pairs (Xie, Girshick, and Farhadi 2016; Garg et al. 2016; Steinbrucker and Pock 2009; Godard, Aodha, and Brostow 2017; Zhou, Brown, and Lowe 2017). In another approach, Chen *et al*. (Chen et al. 2016) proposed a model that learns to estimate metric depths using annotated relative depths, Kuznietsov *et al*. (Kuznietsov, Stückler, and Leibe 2017) proposed a semi-supervised approach using 3D measurements as supervised and stereo pairs for unsupervised learning, and depth estimation through synthetically rendered images (Gaidon et al. 2016).

Our approach is based on unsupervised learning models (Godard, Aodha, and Brostow 2017; Zhou, Brown, and Lowe 2017) which transfer RGB-based depth data to the thermal image domain. For this purpose, we propose an efficient multi-task approach with a new module termed *Interleaver* to obtain more accurate depth results. Lastly, we generalized a sigmoid function which stably scales up to large disparities and introduce an augmentation method which learns the necessary time-invariant features for robust all-day depth prediction.

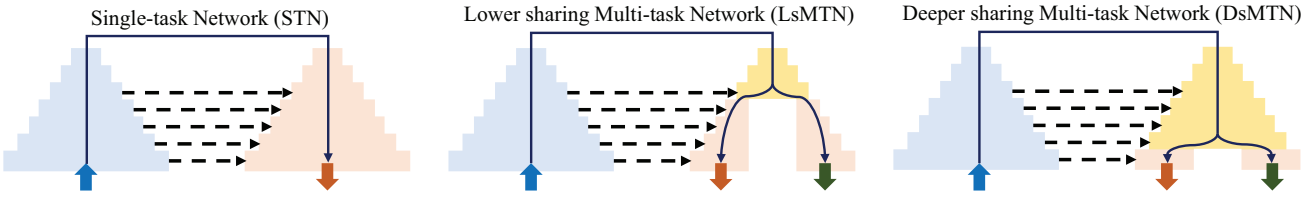
Multispectral Stereo Dataset

Multispectral datasets (Hwang et al. 2015; Choi et al. 2015) have recently been utilized in the computer vision and robotics communities. However, most existing datasets focus on recognition tasks and are thus not suitable for data-driven methods without refined depth ground truths. Therefore, we introduced for the first time *a large-scale multispectral dataset* for use in day and night conditions. Our multispectral stereo dataset provides a calibrated RGB stereo pair, a co-aligned thermal image with left-view RGB stereo images and 3D measurements, making the dataset compatible with various supervised and unsupervised methods. As shown in Fig. 3, compared to other multispectral stereo datasets (Barrera, Lumbreras, and Sappa 2013; Treible et al. 2017), we focus on real-world driving conditions, such as those on campuses, residential areas, urban



(a) Proposed Multispectral Transfer Network (MTN)

(b) Individual components



(c) Baseline architectures: (from left to right) STN, LsMTN, and DsMTN

Figure 2: Proposed multispectral transfer model (a) and baseline models (c) for the comparison. Each module is denoted in (b)

areas, and suburbs in day and night-times. We also provide fully aligned RGB and thermal pairs using a beam-splitter (Hwang et al. 2015) without clipped rectification regions. More specifically, our dataset covers day [7*am* to 2*pm*], night [10*pm* to 2*am*], and high and under saturated conditions. In total, we provide (#7383) stereo/thermal images for training (#4534) and testing (#2853) during the daytime, while also testing (#1583) pairs at night. We split the training/testing samples using GPS and time-logging data without unnecessary duplication or consistency issues. To achieve more accurate ground truth data, we used 3D measurements and RGB stereo results (Žbontar and LeCun 2016) under daytime conditions. Due to the poor RGB visibility at night, we provide new evaluation metrics for depth images in less-lit conditions using only 3D measurements.

Approach

This section describes the architecture of MTN as shown in Fig. 2-(a), including the details of the unsupervised framework and the efficient multi-task learning strategy for MTN with the new module *Interleaver*. It also discusses how the time-invariant features are learned in a broader disparity range.

Unsupervised Depth Estimation

Given a single thermal image I_T, our goal is to learn a function that can predict the pixel-wise depth estimation $\hat{d}$. Most existing methods require a pair of images and the depth, working in a supervised manner. However, it is not easy to acquire the depth ground truth in an outdoor environment due to sensor limitations. Because our goal is to generate thermal-specific depth images, we designed the model based on unsupervised depth estimation methods (Garg et al. 2016; Godard, Aodha, and Brostow 2017). The basic conception is that, given a calibrated pair of binocular RGB cameras I_R^L and I_R^R, the model is trained to predict the disparity D_w that would enable the warping of the right-view image I_R^R to reconstruct the left-view image I_R^L, as shown below.

$$O_{R_{dist}}^{RGB} = \| I_R^L - I_R^R \otimes D_w(\mathbf{I_R^L}) \|^2 \quad (1)$$

This is the objective function for unsupervised methods, where $\otimes$ denotes the warping I_R^R using estimated the disparity $D_w(\mathbf{I_R^L})$. With prior knowledge of the camera intrinsic/extrinsic parameters, we can predict the pixel-wise depth using the camera focal length f and the baseline distance B as follows: $\hat{d}_R = \frac{f \times B}{D_w(\mathbf{I_R^L})}$.

We extend this method to our multispectral transfer framework for depth estimation from a single thermal image. To train the multispectral transfer network, we instead feed the thermal image I_T^L to the model to estimate the disparity D_w^m to warp the right-view RGB image I_R^R to the left-view RGB image I_R^L.

$$O_{R_{dist}}^{RGBT} = \| I_R^L - I_R^R \otimes D_w^m(\mathbf{I_T^L}) \|^2 \quad (2)$$

The key insight is that multispectral images share global context information such as the boundary of the scene and the silhouettes of objects despite the loss of detail in thermal (Hwang et al. 2015; Choi et al. 2015; Jingjing et al. 2016). To generate thermal-specific depth images, we used the L2 loss to optimize the objective function. According

Figure 3: Examples of the proposed multispectral stereo dataset. From top to bottom: an image taken on *campus*, in a *residential* area, in an *urban* area, and in a *suburb*. From left to right: the rectified RGB stereo pair, and the co-aligned thermal image. On the right-bottom side of the thermal image, we denote the number of training/testing frames for each scenario.

to generative methods (Mathieu, Couprie, and LeCun 2016; Yoo et al. 2016; Kingma and Welling 2014), pixels are actually drawn from a complex multi-modal distribution. However, the L2 loss induces the pixel intensity to the average variable of multiple modes such that the trained model produces blurry predictions without complex details of the original images. Therefore, we optimize the objective function with the L2 loss in an unsupervised learning framework.

Multispectral Transfer Network

Efficient Multi-task Learning For more discriminative feature learning, multi-task learning is generally used with a deep neural network to model related tasks jointly in various computer vision and machine learning tasks (Ren et al. 2015; Iizuka, Simo-Serra, and Ishikawa 2015; Eigen and Fergus 2015; Kokkinos 2017). Although multi-task learning typically induces positive feedback between each task, additional efforts are required to prepare subsequent tasks. Moreover, the relevance and sharing between tasks can affect the performance. In depth estimations, multi-task learning was applied simultaneously to estimate surface-normal and segmentation labels (Eigen and Fergus 2015) in an indoor dataset. The surface normal is difficult to obtain in outdoor conditions, and manually labeling such that involving scenarios is not feasible. For multi-task learning to succeed, we define the problem such that it simultaneously predicts the depth and *chromaticity* of the aligned left-view RGB image. The chromaticity has been used in various works related to depth (Heo, Lee, and Lee 2016; Park et al. 2011). As a unique property of the visible spectrum, it has been demonstrated that it is relevant to contextual information (Iizuka, Simo-Serra, and Ishikawa 2015; Zhang, Isola, and Efros 2016) to improve the depth quality. Moreover, we do not require additional works to obtain the source. We propose an efficient multi-task method for depth estimation which simultaneously estimates the depth and chromaticity. In the ensuing experiments, we show that the proposed multi-task based method can generate a more accurate depth map than a model based on the learning of a single task.

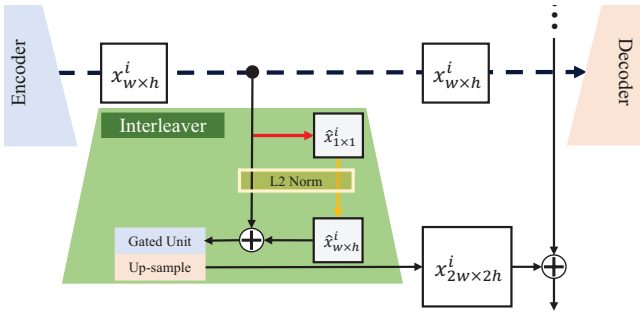


Figure 4: The configuration of Interleaver. Individual components are indicated in Fig. 2. The proposed module is composed of the global/up-pooling (red/yellow arrow), gating mechanism (blue box), and up-sampling unit (red box).

Proposed Architecture (Interleaver) An overview of the proposed framework is illustrated in Fig. 2-(a). Our framework is based on a general skip-connected network, and the standard convolution and deconvolution are represented by blocks of layers. A general extension to multi-task learning is the sharing of feature layers in both tasks, with the splitting of the intermediate layer via a task-specific approach, as shown in the two rightmost models in Fig. 2-(c). This inbuilt sharing mechanism (sharing or split-architecture) is decided after experimenting with splits at multiple layers and picking the best one. Therefore, this approach relies on enumerating multiple network architectures specific to the tasks, because it is challenging to define the inter-relationship or dependency between tasks. Therefore, we propose a novel multi-task module called *Interleaver* to explore the best model without having to train all of them. The goal of the interleaver is to combine multi-task into a single network in a way such that the tasks supervised how much sharing is needed. Motivated by gating mechanism in recurrent neural networks (RNNs), we model sharing of representations by learning gated weights using convolution modules at each skip-connected layers. As shown in Fig. 4, each Interleaver takes a skip-connected feature $x^i_{w\times h}$ and pools the feature via global average pooling and then up-pools it to add to the input feature map. This pooling mechanism has been qualified to effectively enlarge the receptive field and improve the generative results. This pooled feature passes the gated convolution to learn the control flow of the chromaticity information for the skip-connected features, after which

(a) AM07, AM09, AM11, PM04, and AM02

(b) γ_1, γ_2, original, γ_3, and γ_4

Figure 5: Example images captured at different times (top) and our photometric correction results (bottom). The original image can be augmented by certain parameters (γ_n). Our result can reasonably represent images captured at different times, because these parameters are estimated by a data-driven approach from training samples.

it passes the upsampling layer to serve as the feature $x^i_{2w\times 2h}$ to the corresponding to following layers. Through the proposed module, we can obtain a better representation model for learning to capture finer details from lower convolutional layers with local receptive fields and contextual information from the chromaticity. Moreover, the proposed multi-task architecture can minimize the effort needed to design optimal networks and reduce the adverse effects of previous approaches. During the training process, we optimize the same strategy as the general multi-task learning.

Adaptive Activation For a fully differentiable model, we replaced the linearized warped model (Garg et al. 2016) with a bilinear interpolation sampler (Jaderberg et al. 2015). The main difference from earlier work (Godard, Aodha, and Brostow 2017) is the penultimate activation, which scales the estimated disparity. Godard *et al*. (Godard, Aodha, and Brostow 2017) used a fixed-scale sigmoid function to control the maximum disparity level. However, the maximum disparity level can vary depending on the sensor configuration and dataset used. Accordingly, finding the optimal scale is not a trivial problem without the ground truth. Moreover, the bilinear sampler module would be unstable if used to increase the scale of the sigmoid in the initial steps, and its derivative cannot generally handle cases in which the current focus is outside of the region of the pixel space. Therefore, we propose an adaptive scaled sigmoid function (S_{ass}) which iteratively increases the scale of the sigmoid for stable convergence while covering a large-scale maximum disparity range.

$$S_{ass}(x) = \frac{\beta}{1+e^x}, \quad \begin{cases} \beta = \beta_0, & \text{if } epoch = 1 \\ \beta = \beta + \alpha, & \text{otherwise} \end{cases} \tag{3}$$

As we increase the initial variable β_0 by α in certain epochs, our model can undertake learning with large-scale disparities without interruptions during training. The entire network configuration is illustrated in the website.

Photometric Correction Even if thermal sensors are robust to illumination changes, there are changes in the thermal contrast ratio relative to the amount of heating energy over time. More specifically, this variation can occur during both day and night, during different seasons, and at sunset and sunrise, implying that this is an important issue in thermal-based applications. Moreover, we can only use multispectral pairs taken in daytime conditions due to the poor visibility of RGB sensors at night. To alleviate these issues, we propose a data-driven photometric correction $P_{cor}(x) = \lambda x^\gamma$ method. The basic concept is that we estimate the parameters from various contrast ratio images. To do this, we collect temporally ordered images during the time range of 7*am* to 2*am* (Fig. 5-(a)) and then convert the training images into corrected images using pre-defined correction parameters (γ, λ). We then compare the similarity of the intensity and gradient histogram-based features between the corrected images and the temporally ordered images to vote on each parameter. With our photometric correction method (Fig. 5-(b)), we can augment the realistic contrast variation of the thermal images at different times. This method is simple to implement, but it is crucial to learn the time-invariant features for all-day depth predictions. In our experiments, we show that our proposed photometric correction approach can greatly affect the quality of the depth image, particularly at night.

Training Loss

We formulate a multi-task objective function that incorporates the reconstruction of disparities, the chromaticity, and the smoothness prior to the disparity.

$$E_{MTN} = O^{RGBT}_{R_{disp}} + \lambda_s O_{S_{disp}} + \lambda_c O_{R_{chrom}} \tag{4}$$

$O^{RGBT}_{R_{disp}}$ encourages the reconstructing of warped images to corresponding pairs to learn disparities from multispectral correspondences. Because the disparity discontinuity generally corresponds to the edge of the image I, we use simple l_1 regularization on the gradient of the disparity D, similar to (Godard, Aodha, and Brostow 2017) for smoothness priors.

$$O_{S_{disp}} = |\bigtriangledown D_x|\ e^{-\|\bigtriangledown I_x\|} + |\bigtriangledown D_y|\, e^{-\|\bigtriangledown I_y\|}. \tag{5}$$

For the chromaticity estimation $O_{R_{chrom}}$, we tested several color codings to extract the chromaticity from RGB images and finally used the *YCbCr* color coordinates shown below,

$$O_{R_{chrom}} = \sum_{i=1}^{2} \| \ C_i^L - X_i^L \ \|^2, \tag{6}$$

where C_i^L and X_i^L denote the ground truth and the prediction of each *CbCr* channel respectively.

Experiments

Implementation Details

The network is implemented in MatConvnet (Vedaldi and Lenc 2015), and takes 20 hours to train using a single NVIDIA TITAN X GPU on 4.5 thousand pairs for 40 epochs. During the training process, we set the weights of the objective terms as $\lambda_s = 0.01$ and $\lambda_c = 0.01$ and use SGD for optimization with a momentum(0.9)/weight decay(0.0005) from scratch within a learning rate of 10^{-5}.

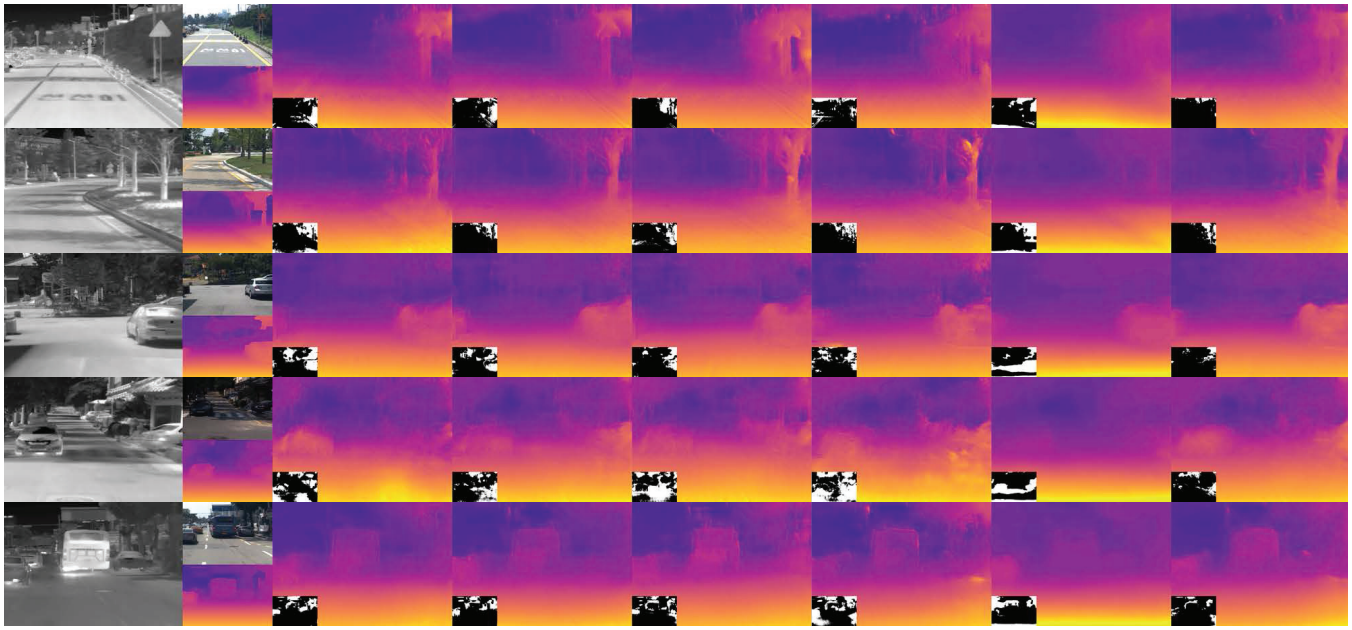


Figure 6: Qualitative results from the day scenario. (From left to right) thermal image, the co-aligned RGB image, depth images and several results (STN, LsMTN, DsMTN, MTN-P, DIW, and MTN). The left-bottom frame of the results is a binary error map which represents the error over three pixels of disparity. Our method outperforms when representing global scenes and objects such as cars, trucks, and those in nature.

According to the normalized coordinate of the bilinear sampler (Jaderberg et al. 2015), we set an adaptive scaled sigmoid function which initially sets $\beta_0 = 0.3$, with an increase by $\alpha = 0.01$ every two epochs. During the training process, we found that the proposed function can encourage stable convergence to cover a large range of disparity. Instead of ReLU, we used ELU, similar to (Godard, Aodha, and Brostow 2017) and batch normalization (Ioffe and Szegedy 2015). For every iteration, we conduct photometric correction and data augmentation. The correction parameters are pre-computed by the proposed method as γ=[0.5, 0.75, 1.25, 1.5] and $\lambda = 1$, and we randomly cropped the center region in the margin of [64, 96].

Evaluation

Baselines We evaluate our approach on the proposed multispectral stereo benchmarks. To do this, we designed the unsupervised baseline methods shown in Fig. 2-(c). Note that essentially all baselines share the same architecture with skip-connections. First, we set our full model as (**MTN**). The single transfer network (**STN**) is trained by only depth estimation tasks in the leftmost model shown in Fig. 2-(c) to verify the effect of the chromaticity-based multi-task learning. To prove the superiority of Interleaver, we provide two general multi-task models which directly share the feature layer. According to the amount of feature sharing, one is termed Low-shared MTN (**LsMTN**) and the other is referred to as Deep-shared MTN (**DsMTN**), as shown in the center and rightmost models in Fig. 2-(c). Lastly, we set the model without photometric correction, as (**MTN-P**).

Our aim is to demonstrate multispectral transfer learning via depth prediction so as to compare the outcome with those of other depth prediction methods based on the thermal image input. We select various approaches for the comparison, including a supervised method by Eigen *et al*. (Eigen, Puhrsch, and Fergus 2014) and an ordinal-based depth estimation method (Chen et al. 2016), denoted as **Eigen** and **DIW** respectively. Although our framework is not suitable for these types of methods, we compare MTN to the cornerstone models to verify that our generated depth offers better quality than general methods. Lastly, we conducted an experiment using RGB images as inputs with the compared methods (**STN-RGB, Eigen-RGB, DIW-RGB**). Through these additional baselines, we demonstrate that our depth result from a single thermal image has reasonable quality compared to the depth results from a single RGB image. It also offers the advantage of being able to estimate the depth at night.

Evaluation Metrics In day scenarios, we evaluate the accuracy of the proposed method for depth predictions using conventional metrics from earlier work (Eigen, Puhrsch, and Fergus 2014). These metrics measure the error in terms of both the physical distance from the ground truth and the accuracy levels within certain threshold depth ranges. In night scenarios, due to the poor visibility of RGB images, we cannot measure the performance using these metrics. Moreover, the simple comparison of 3D laser measurements does not represent reasonable performance because these points cannot cover the entire image, and depths from far distances, reflective objects, and boundaries are not accurate, causing

Table 1: Quantitative results in day scenarios.

	Distance Metric								Accuracy Metric					
	RMS		***log* RMS**		**Absolute relative**		**Square relative**		**Accuracies** $\delta < 1.25^1$		**Accuracies** $\delta < 1.25^2$		**Accuracies** $\delta < 1.25^3$	
	$1{\sim}80m$	$1{\sim}50m$	$1{\sim}80m$	$1{\sim}50m$	$1{\sim}80m$	$1{\sim}50m$	$1{\sim}80m$	$1{\sim}50m$	$1{\sim}80m$	$1{\sim}50m$	$1{\sim}80m$	$1{\sim}50m$	$1{\sim}80m$	$1{\sim}50m$
STN	10.418	7.7737	0.2137	0.2000	0.1616	0.1531	2.7168	2.2767	0.7759	0.8060	0.9349	0.9337	0.9784	0.9776
LsMTN	9.5155	6.6967	0.1981	0.1801	0.1422	0.1325	2.1029	1.6322	0.8053	0.8358	0.9472	0.9492	0.9835	0.9842
DsMTN	9.3808	6.3671	0.2016	0.1761	0.1390	0.1259	1.9780	1.4394	0.7966	0.8407	0.9469	0.9544	0.9828	0.9855
MTN-P	10.5755	7.0058	0.2236	0.1951	0.1573	0.1413	2.4506	1.7251	0.7540	0.8040	0.9314	0.9440	0.9787	0.9827
MTN(Ours)	**8.7387**	**6.0786**	**0.1933**	**0.1714**	**0.1307**	**0.1207**	**1.7394**	**1.3119**	**0.8124**	**0.8451**	**0.9508**	**0.9557**	**0.9842**	**0.9868**
STN-RGB	10.3758	7.5876	0.2326	0.2094	0.1657	0.1570	2.5682	2.0618	0.7395	0.7772	0.9276	0.9378	0.9769	0.9806
Eigen-RGB	12.9946	10.1792	0.2513	0.2386	0.2105	0.1992	4.6046	4.0629	0.7095	0.7551	0.8985	0.8965	0.9649	0.9612
Eigen-T	12.9632	10.266	0.2505	0.2384	0.2090	0.1976	4.6110	4.0835	0.7126	0.7561	0.8980	0.8947	0.9656	0.9618
DIW-RGB	9.3927	6.4993	0.2029	0.1934	0.1660	0.1644	2.3764	1.8030	0.7743	0.7956	0.9485	0.9482	0.9840	0.9842
DIW-T	10.4869	6.4427	0.2105	0.1967	0.1754	0.1697	3.0885	1.7543	0.7585	0.7825	0.9413	0.9454	0.9828	0.9851

*Note that *distance metrics* are that lower variable is better, and *accuracy metrics* are that higher is better.

erroneous evaluations. Therefore, we propose a new metric, f (Eq. (7)), which considers the ordinal information of the estimated depth $\mathbb{D}$ on projected LiDAR points instead of distance metrics. Because the simple ordinal comparison has some degree of ambiguity given the erroneous points, we introduce a penalty for ordinal pairs in $\mathcal{L}$, which has distance error exceeding a certain threshold ($\xi(m)$), similar to accuracy metrics (Eigen, Puhrsch, and Fergus 2014).

$$\begin{aligned}
&\mathbb{M} = \{(d_i, d_j) \mid d_i, d_j \in \mathbb{D}, i \neq j\} \\
&\mathbb{L} = \{(d_i, d_j) \mid (d_i - d_i^{gt}) < \xi, (d_i, d_j) \in \mathbb{M}\} \\
&f = \frac{1}{|\mathbb{M}|} \sum_{(d_i,d_j)\in\mathbb{L}} C(d_i, d_j) \\
&C(d_i, d_j) = \begin{cases} 1, & \text{if } \textit{sign}(d_i - d_j) = \textit{sign}(d_i^{gt} - d_j^{gt}) \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{7}$$

Results Table. 1 shows our results in relation to the baselines and the state-of-the-art methods of multispectral stereo benchmarks. We conducted the experiments in two aspects [1 to 50 meters, 1 to 80 meters] to cover all evaluations of the previous single-view depth estimation. For most metrics, our proposed method clearly performs the best. In both ranges, the chromaticity multi-task model (the series of MTN) outperforms the single-task method. Moreover, our proposed Interleaver module (MTN) predicts the more accurate depth than the feature-sharing models (LsMTN, DsMTN). The deep-shared model (DsMTN) can improve the quality of depth than that of the low-shared model (LsMTN). The most noticeable point is that MTN allows all connections between layers. The Interleaver encouraged models to be automatically tuned by learning the gated weights between tasks in every connections. Totally, our full model (MTN) can improve the performance of (STN) by more than around 22% and 16% in terms of RMSE in 50 and 80 meters respectively. Furthermore, we can see that the performance of MTN-P is worst in most measurements as similar to the single task baseline (STN). Therefore, we conclude that our photometric correction is very important for handling the thermal image invariant property. Compared to other supervised methods, all our baselines outperform Eigen and our efficient multi-task approach mostly shows the better performance than DIW. When evaluating 50 meters, DIW shows good results in most of the metrics. However, since the relative depth has an ambiguity in the longer range, the depth accuracy became worse. Our MTN is still robust to the longer range compared to other methods. In qualitative results (Fig. 6), other predictions may appear more plausible and seem to smoother. However, these things are not always consistent with ground truth depth maps. According to error maps, our predictions for global and local scene boundaries and depth consistency levels are more accurate. In Table. 2, we compare the proposed method to the same baselines at night with the proposed metric in at [10, 20, and 30] meters. We find that the average accuracy is lower than the results of day scenarios due to the different conditions of the scenes. Regardless of this fact, our results clearly demonstrate the benefit of using thermal images compared to RGB-based baselines during night conditions, as the accuracy is greater by threefold in the tightest threshold. For most metrics and setups, MTN performs best with similar tendencies, indicating that our efficient learning and Interleaver can provide meaningful training cues to depth estimations. DIW shows the better results at [20,30] meters because that method is robust to relative depth relationships. However, the accuracy of the metric depths is not higher than that of the proposed

Table 2: Quantitative results in night scenarios.

	Ordinal Accuracy Metric		
	$\xi = 10$	$\xi = 20$	$\xi = 30$
STN	0.3233	0.6237	0.7317
LsMTN	0.3405	0.6855	0.7753
DsMTN	0.3745	0.6820	0.7797
MTN-P	0.3096	0.6225	0.7397
MTN	**0.4666**	0.7026	0.7757
STN-RGB	0.2508	0.3284	0.3592
Eigen-RGB	0.1728	0.2442	0.3064
Eigen-T	0.2033	0.6178	0.7516
DIW-RGB	0.1404	0.3176	0.3805
DIW-T	0.3744	**0.7459**	**0.8401**

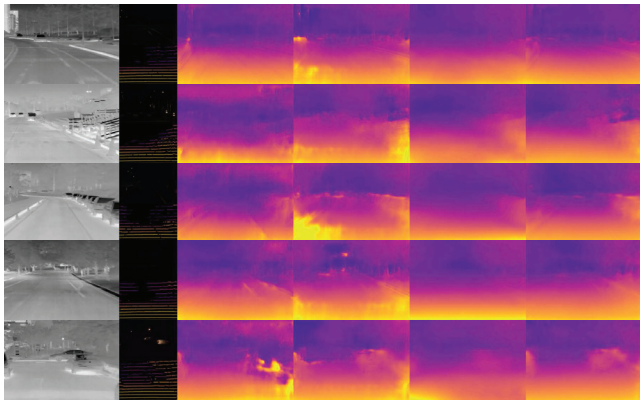


Figure 7: Qualitative results in the night scenario. From left to right: input thermal image, result of STN, result of MTN-P, result of DIW, result of the proposed MTN. Our method achieves better qualitative results at night despite being trained during day conditions. Compared to MTN-P, photometric correction is not a trivial function to resolve thermal time-variant properties.

method. As shown in Fig. 7, the results from MTN-P have some artifacts in the frame due to thermal contrast ratio issues. However, we note that adjusting the photometric correction method can train the model to be robust, providing it with the ability to estimate accurate depth images in such challenging scenarios.

Discussion

Our answer to the question in the introduction of *how can we estimate dense and accurate depth images all day* is to use a multispectral transfer approach for depth estimations from a single thermal image. The proposed multi-task approach with chromaticity can improve the performance, and the Interleaver module encourages better representations with less adverse effects of feature-sharing methods. As shown in Fig. 8, our predicted depth shows a good quality in day and night conditions. In the comparison with RGB-based models, we demonstrated that the depth of a single thermal image has realistic and reasonable quality and that our model has generalization ability sufficient to estimate the depth in various conditions.

Due to the different properties of multispectral domains, the chromaticity is not wholly plausible when used for reconstruction, and objects and boundaries strongly represented by chromaticity can be well reproduced. Compared to feature-sharing multitask learning, our model can learn selection and attention to control the relevant and useful features from this result using the proposed Interleaver module.

While photometric correction can regularize the variation of the thermal contrast ratios to some extent, there remain several issues to resolve before thermal image variants can be covered. However, we think that our correction method offers simple but effective guidance for dealing with one of the main issues when the applying thermal images to data-driven methods.

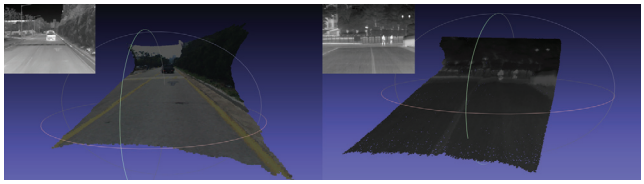


Figure 8: Examples of 3D reconstruction from thermal images at day (left) and night (right).

Conclusion

In this paper, we proposed the first multispectral transfer framework for depth estimation from a single thermal image. Our main concern is the generation of depth beyond day conditions using illumination-invariant thermal images. To realize this goal, we created a large-scale multispectral stereo dataset in various driving regions. Based on knowledge of the multispectral relationships, we designed an efficient multi-task learning framework using chromaticity without additional annotated data or data acquisition. For accurate predictions, we proposed Interleaver to encourage an efficient but accurate multitask learning using chromaticity features. Lastly, we explained the adaptive scaled sigmoid for stable convergence while covering a large disparity level, with photometric correction for thermal images to resolve the thermal variant problems for day and night depth predictions. To verify our contributions, we conducted experiments involving various cases compared to self-designed baselines, the results of previous works, and multi-modality approaches.

Acknowledgements

This work was supported by the Development of Autonomous Emergency Braking System for Pedestrian Protection project funded by the Ministry of Trade, Industry and Energy of Korea (MOTIE)(No.10044775). We also acknowledge supports to finish this work from NAVER LABS and NAVER, and the gold prize from 23th HumanTech Paper Award in Samsung.

References

Barrera, F.; Lumbreras, F.; and Sappa, A. D. 2013. Multispectral piecewise planar stereo using manhattan-world assumption. *PRL*.

Bell, S.; Zitnick, C. L.; Bala, K.; and Girshick, R. 2016. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*.

Chen, W.; Fu, Z.; Yang, D.; and Deng, J. 2016. Single-image depth perception in the wild. In *NIPS*.

Choi, Y.; Kim, N.; Park, K.; Hwang, S.; Yoon, J. S.; and Kweon, I. S. 2015. All-day visual place recognition: Benchmark dataset and baseline. In *CVPR Workshop*.

Choi, Y.; Kim, N.; Hwang, S.; and Kweon, I. S. 2016. Thermal enhancement network using convolution neural network. In *IROS*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Eigen, D., and Fergus, R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*.

Farabet, C.; Couprie, C.; Najman, L.; and Lecun, Y. 2016. Learning hierarchical features for scene labeling. *TPAMI*.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*.

Garg, R.; Kumar, B. V.; Carneiro, G.; and D. Reid, I. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *IJRR*.

Godard, C.; Aodha, O. M.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *CVPR*.

Hariharan, B.; Arbeláez, P. A.; Girshick, R. B.; and Malik, J. 2015. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*.

Heo, Y. S.; Lee, K. M.; and Lee, S. U. 2016. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *TPAMI*.

Hoffman, J.; Gupta, S.; and Darrell, T. 2016. Learning with side information through modality hallucination. In *CVPR*.

Hwang, S.; Park, J.; Kim, N.; Choi, Y.; and Kweon, I. S. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*.

Iizuka, S.; Simo-Serra, E.; and Ishikawa, H. 2015. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *TOG*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *NIPS*.

Jingjing, L.; Shaoting, Z.; Shu, W.; and Dimitris N., M. 2016. Multispectral deep neural networks for pedestrian detection. In *BMVC*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Kokkinos, I. 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*.

Kuznietsov, Y.; Stückler, J.; and Leibe, B. 2017. Semi-supervised deep learning for monocular depth map prediction. *CVPR*.

Ladicky, L.; Shi, J.; and Pollefeys, M. 2014. Pulling things out of perspective. In *CVPR*.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. Deeper depth prediction with fully convolutional residual networks. In *3DV*.

Li, B.; Shen, C.; Dai, Y.; van den Hengel, A.; and He, M. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*.

Limmer, M., and Lensch, H. P. A. 2016. Infrared colorization using deep convolutional neural network. In *ICMLA*.

Liu, F.; Shen, C.; Lin, G.; and Reid, I. D. 2015. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Mathieu, M.; Couprie, C.; and LeCun, Y. 2016. Deep multi-scale video prediction beyond mean square error. In *ICLR*.

Park, J.; Kim, H.; Tai, Y.-W.; Brown, M. S.; and Kweon, I. S. 2011. High quality depth map upsampling for 3d-tof cameras. *ICCV*.

Patricia L. Suarez, A. D. S., and Vintimilla, B. X. 2017. Infrared image colorization based on a triplet dcgan architecture. In *CVPR Workshop*.

Poujol, J.; Aguilera, C. A.; Danos, E.; Vintimilla, B. X.; Toledo, R.; and Sappa, A. D. 2016. A visible-thermal fusion based monocular visual odometry. In *Robot*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster-rcnn: Towards real-time object detection with region proposal network. In *NIPS*.

Steinbrucker, F., and Pock, T. 2009. Large displacement optical flow computation without warping. In *ICCV*.

Treible, W.; Saponaro, P.; Sorensen, S.; Kolagunda, A.; O'Neal, M.; Phelan, B.; Sherbondy, K.; and Kambhamettu, C. 2017. Cats: A color and thermal stereo benchmark. In *CVPR*.

Vedaldi, A., and Lenc, K. 2015. Matconvnet: Convolutional neural networks for matlab. In *ACMMM*.

Žbontar, J., and LeCun, Y. 2016. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*.

Yoo, D.; Kim, N.; Park, S.; Paek, A. S.; and Kweon, I. S. 2016. Pixel-level domain transfer. In *ECCV*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *ECCV*.

Zhou, T.; Brown, Matthew Sanvely, N.; and Lowe, D. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*.