Supplementary Material
Guided Deep Decoder: Unsupervised Image Pair Fusion

Tatsumi Uezato[1][0000−0002−8264−201X], Danfeng Hong[2,3][0000−0002−3212−9584],
Naoto Yokoya[4,1][0000−0002−7321−4590], and Wei He[1][0000−0003−3410−0643]

[1] RIKEN AIP, Tokyo, Japan
[2] German Aerospace Center, Wessling, Germany
[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
[4] The University of Tokyo, Tokyo, Japan
{tatsumi.uezato,naoto.yokoya,wei.he}@riken.jp, {danfeng.hong}@dlr.de

1 Ablation Study

The results of the ablation study that examines the contributions of each unit in GDD are shown in Table 1. The results showed that DD that did not incorporate URU and FRU performed poorly. This shows that it is difficult to generate high resolution images from random noise without the help of guidance images. The features of the guidance images are important to reconstruct high quality images in the unsupervised image fusion problems. The results also showed that the incorporation of URU or FRU significantly reduced RMSE. In addition, GDD that incorporated both URU and FRU led to further performance boost.

We have also evaluated the effect of guidance feature augmentation at different levels of the deep decoder. The level of the deep decoder used in this study is 5. The results are shown in Table 2. The maximum gains are coming from the layer 5. The result shows that the upsampling refinement at the final layer is more important than other layers. It also shows that the lower-level features from a guidance image can be more effective in the upsampling refinement unit as a regularizer than the higher-level features.

Table 1: Ablation study using the RMSE metric. PAN represents the results of panchromatic and multispectral image fusion. HS represents the results of hyperspectral and RGB image fusion. The average results of RMSE for all datasets are shown.

Method	URU	FRU	PAN	HS
DD			12.4157	6.7016
GDD	✓		5.2035	2.3017
GDD		✓	5.2332	2.2554
GDD	✓	✓	**4.7475**	**2.0213**

Table 2: Ablation study: effect of guidance feature augmentation at various levels. PAN represents the results of panchromatic and multispectral image fusion. HS represents the results of hyperspectral and RGB image fusion. The average results of RMSE for all datasets are shown.

	layer 1	layer 2	layer 3	layer 4	layer 5
PAN	11.8	10.62	9.97	8.49	**4.91**
HS	6.43	5.54	4.41	3.04	**2.37**

2 Runtime analysis

We have calculated the computational time of the compared methods as shown in the Table 3. The computational time of the hyperspectral super-resolution task is shown because the task is the most computationally expensive and the unsupervised deep learning method (uSDN) can be compared in this task. GDD and DIP are slower than classical matrix factorization methods, but are comparable to the unsupervised deep learning method (uSDN) or a Bayesian approach (BSR). Although supervised deep learning method is much faster for the inference, the training takes more than 12 hours, which may be required for each task.

Table 3: Computational time to compute one high resolution image.

	CNMF	BSR	NSSR	NLSTF	uSDN	MHF	DIP	GDD
Time (s)	**21**	1325	138	64	1773	22	1483	1605

3 GDD and existing network architectures

GDD is closely related to existing network architectures. URU and FRU in GDD generate multiplicative transformation parameters from the guidance image for the spatial and channel-wise feature modulation. A similar feature modulation has been also used in [8, 2, 1, 6, 4]. In [8], affine transformation parameters have been generated from segmentation probability maps for the feature modulation to achieve more realistic textures in image super-resolution. The affine transformation has been also considered in the style transfer [3]. Although the affine transformation considers the scaling and bias values, URU and FRU consider only the scaling values because we find that similar results can be obtained at lower computational cost for unsupervised optimization problems.

The conditional weights can be also interpreted as attention layers across all channels. In [2, 5, 1], attention gates are incorporated to refine the spatial details and highlight salient features. The conditional attention weights are generated from a label map for semantic image synthesis [4]. GDD is closely related to the conditional attention weights in that it uses the multi-scale features from a guidance image to generate the conditional attention weights. However, all of the aforementioned studies consider and require a large size of training data. The network architectures have not been fully explored as a regularizer for the unsupervised optimization problems. Our study is different from previous studies in that it uses the network architecture as a regularizer to solve a variety of unsupervised image fusion problems.

4 Early stopping

The early stopping is required, depending on a task. For the super-resolution task, the early stopping is not required because the value of the loss function converges. We stopped the iteration by empirically validating the changes of MSE derived by the loss function. The number of iterations is fixed for all data within each task and is large enough to converge. For the denoising task, the early stopping is required as the reviewer pointed out. We stopped the iteration so that we obtained the most qualitative results as done in [7].

5 Additional Results

In Fig. 1 and 2, the additional results of the hyperspectral and RGB image fusion on the CAVE dataset are shown. In Fig. 3, 4 and 5, we show the additional results of the panchromatic and multispectral image fusion.

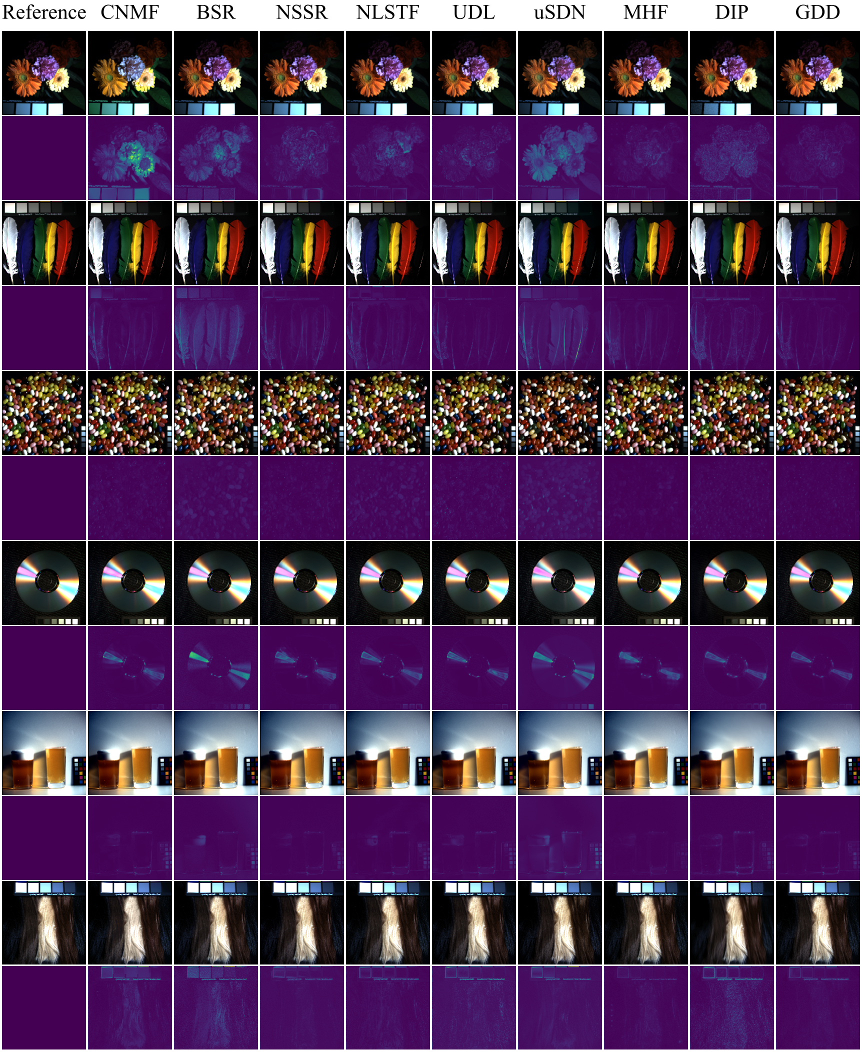

Fig. 1: Additional results of HS and RGB image fusion from the CAVE dataset. RGB images of the enhanced HS images and the corresponding error maps are shown.

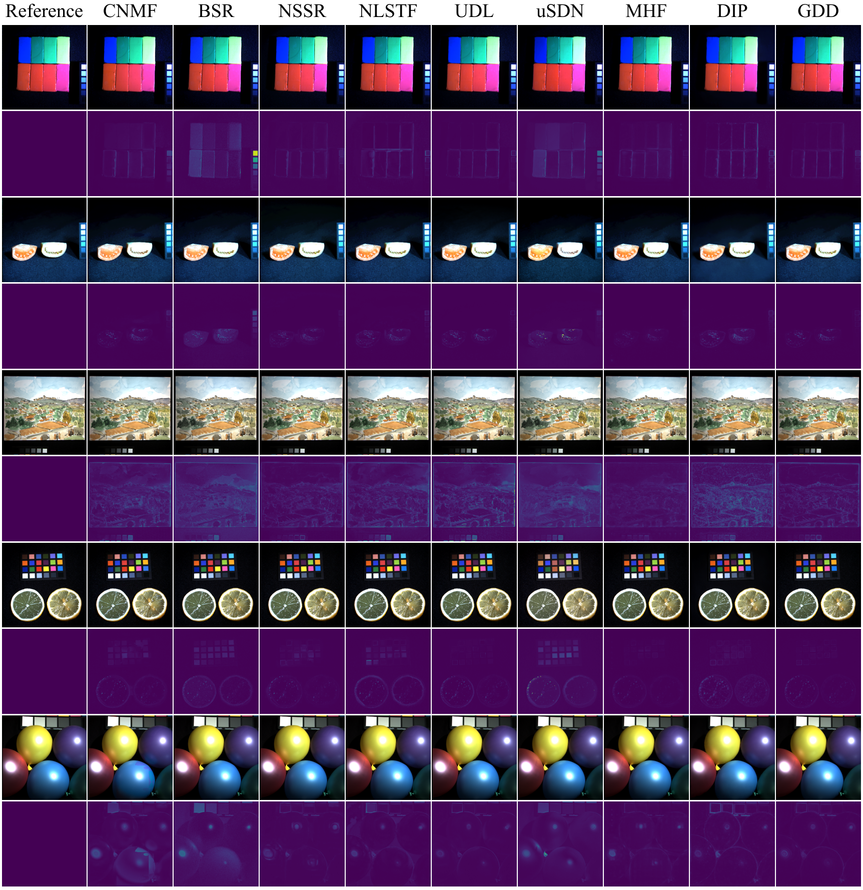

Fig. 2: Additional results of HS and RGB image fusion from the CAVE dataset. RGB images of the enhanced HS images and the corresponding error maps are shown.

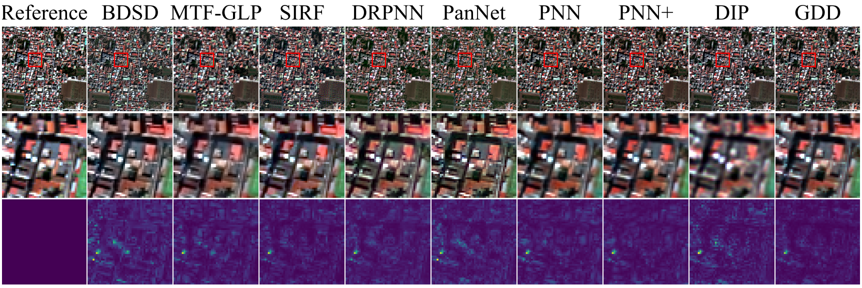

Fig. 3: Additional results of panchromatic and multispectral image fusion. First row: RGB images of the pansharpened MS images. Second row: The enlarged RGB images. Third row: The corresponding error maps.

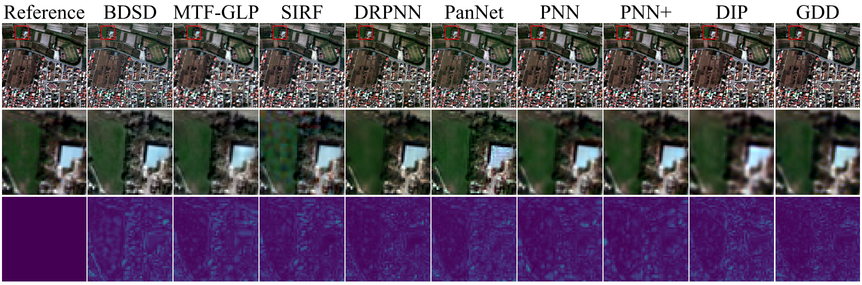

Fig. 4: Additional results of panchromatic and multispectral image fusion. First row: RGB images of the pansharpened MS images. Second row: The enlarged RGB images. Third row: The corresponding error maps.

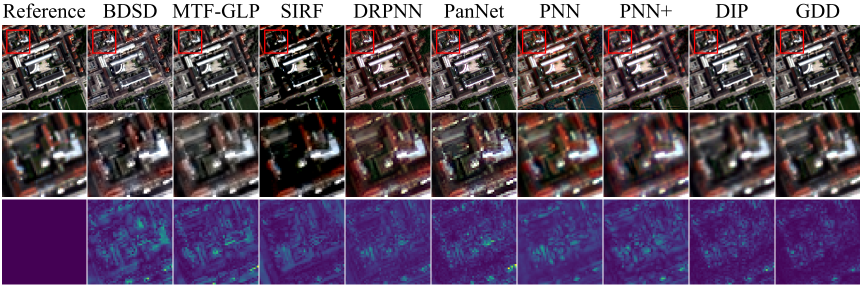

Fig. 5: Additional results of panchromatic and multispectral image fusion. First row: RGB images of the pansharpened MS images. Second row: The enlarged RGB images. Third row: The corresponding error maps.

References

1. Amirul Islam, M., Rochan, M., Bruce, N.D.B., Wang, Y.: Gated feedback refinement network for dense image labeling. In: CVPR (2017)
2. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019)
3. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
4. Liu, X., Yin, G., Shao, J., Wang, X., Li, h.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: NeurIPS (2019)
5. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: learning where to look for the pancreas. In: MIDL (2018)
6. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
7. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: CVPR (2018)
8. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018)