Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion (Supplementary Material)

Anonymous ECCV submission

Paper ID 4260

1 Implementation Details

Network Architecture. The whole network follows the U-net architecture as shown in Fig.A 1. The basic module is the ResNeSt block [4]. Specifically, we first employ a simple 3×3 convolutional layer followed by a ReLU layer to transform the source images to obtain the shallow feature representations. And then, we feed these features to the following encoder network E. There are two decoders in the network, denoted as D^u , D^c . The decoders are similar to the encoder, except that the pooling layers are replaced with the bilinear upsample layers. As for the additional encoder E_c , it only has one ResNeSt block. The block numbers and output channel numbers are presented in the figure for better understanding.



Fig.A 1: The details of network architecture.

⁰⁴⁰ **Experimental Settings.** As briefly introduced in the manuscript, the mask ⁰⁴¹ M has the same size as the training image and consists of two types of patches ⁰⁴² with different sizes, *i.e.* 10 × 10 and 20 × 20 pixels, respectively. In Tab.A 1, we ⁰⁴³ present more details about the experimental settings. In the training stage, the ⁰⁴⁴ generated masks perform great diversity due to the randomness introduced in the preprocessing step. Specifically, the mask ratio of each source image is randomly drawn from a uniform distribution in $\mathcal{U}(0.55, 1.0)$, so that the common ratio (i.e. the proportion of the area of common regions to the area of the whole image) varies from 0.01 to 1.

In addition to parameter settings, some tricks are used to improve the training stability. We found that it is hard for the network to converge for directly training the model with four loss items. Therefore, at the early 3,000 iterations, we ask that only the losses of projected images from common and unique components are calculated while the reconstruction losses of P_r are excluded from the loss computation.

Config Value Common Ratio (0.01, 1.0)Mask Ratio (0.55, 1.0)Mask Resolution f10 10.20 20a RandomCrop (3/4, 4/3)

RandomResize

(0.08, 1.0)Tab.A 1: Training settings of the data preprocessing.

Ablation Studies

To explore the importance of the self-supervised learning with CUD pretext task. we perform comprehensive ablation studies. In general, the ablation studies can be classified into two categories: mask-related and loss-related.

Mask-related Ablation Studies. In the manuscript, the image augumatation is defined as follows:

$$\boldsymbol{x}^{i} = M_{i}(\boldsymbol{x}) + \bar{M}_{i}(n), \qquad (1)$$

where n denotes the Gaussian noise, \boldsymbol{x} is original scene, and M_i is the mask. Obviously, the noise patch $\overline{M}_i(n)$ is independent to the counterpart (region) of the other image. To verify the effectiveness of noise patches, we replace the noise patches with zero patches in the ablation study. Recently, the masked image modeling [2] demonstrates that the patch size of the mask is related to the performance of the model to some extent. To verify this opinion, we also conduct an another simple ablation study about the patch size of mask M.

We show qualitative results of mask-related ablation studies in the multi-focus image fusion task in Fig.A 2. The fused result of Fig.A 2c is generated by the model trained with zero patches. As can be seen, the model performs worse than the DeFusion shown in Fig.A 2e. Considering that the performance of multi-focus image fusion mainly depends on the quality of unique information. we can infer that using the noise patches in image transformation may be more conducive to learning unique semantic information. The results of the quantitative evaluation are consistent with the qualitative results. Compared with the performance of DeFuion, the ablation study using the zero patches shows lower



Fig.A 2: Qualitative comparisons of mask-related ablation studies. The (c) is generated by the model whose transformed source images replace the noise patches with zero patches. The resolution of mask patch size is the combination of 16×16 and 32×32 pixels in (d), while the default resolution of DeFusion is the $\{10 \times 10, 20 \times 20\}$ pixels.

metrics (PSNR:-2.24dB, SSIM:-0.017). The fused results of the model trained with bigger patch sizes are shown in Fig.A 2d. It also produces the worse fused result in this example. The model that uses data with the bigger noise patch to train may generate inappropriately semantic information about the scene for the leaf in Fig.A 2d is incorrectly colored into yellow rather than green. Furthermore, the quantitative evaluation show similar results. Compared with the performance of DeFuion, the ablation study using bigger noise patches shows lower metrics (PSNR:-3.41dB, SSIM:-0.031).

Loss-related Ablation Studies. To verify the effectiveness of decomposition components, we conduct some ablation studies about the decomposed common and unique components. Specifically, we design four groups of ablation studies to analyze the disparity in fusion results of the same example when we remove parts of reconstruction losses. In addition, during training the DeFusion, the ground truth of the projected image $\hat{\boldsymbol{x}}_c = P_c(f_c)$ is the $\boldsymbol{x}_c = M_1(\boldsymbol{x}) \cap M_2(\boldsymbol{x})$, where the projected images are asked not only to keep the useful information in the common regions of source images, but also to predict the zero values in the semantically irrelevant regions (*i.e.*, unique regions). In a similar vein, the $\hat{x}_c = P_c(f_c)$ is also required to predict the zero values in the common regions, except for keeping the unique semantic information of source images. The reason why the network needs to predict the zero values at the semantically irrelevant regions is because we hope that the extracted common information can be more accurate. In order to verify the effectiveness of the setting, we remove the constraint that predicting zero values at the semantically irrelevant regions, and only focus on keeping the information in the masked regions.

As shown in Fig.A 3, we employ the multi-exposure image fusion task to verify the capability of each reconstruction loss item. The first row in Fig.A 3 is generated by the model trained without the common construction loss. We can see that the whole visualized common feature map is activated. As for the second row, since we remove the unique reconstruction loss item during training,



Fig.A 3: Qualitative comparisons of loss-related ablation studies. (First row:
training without common reconstruction loss; second row: training without
unique reconstruction losses; third row: training with only original scene reconstruction loss; fourth row: training with only predicting the masked regions;
fifth row: DeFusion in the manuscript.)

the two visualized unique feature maps look similar. It denotes that both the over-exposed and under-exposed images all focus on the same region in the ex-tracted unique features, which may be redundant for fusion. In the third row, we only keep one loss to reconstruct the original scene during training and remove the other reconstruction losses. In this case, the quality of feature representa-tion obviously decreases, e.q. the visualized unique feature map f_u^1 exhibits some artifacts that even present in the final fusion results. In addition, the above-mentioned three ablation studies all show inconsistent edge. The worst edge artifacts are generated by the model train without any common or unique re-construction loss. As for the fourth row, the corresponding model is trained to focus on keeping the information in the masked region. Compared to the De-Fusion models, the visualized feature maps in this case cannot distinguish the common and unique region, so that the corresponding fused result preserves insufficient details of under-exposed image while remaining too much brightness information of over-exposed image.

3 Fused Results

In this section, we present more qualitative fused results on both the MEFB benchmark [6] and SICE dataset [1] for the multi-exposure task, as shown in Fig.A 4. The Fig.A 5 presents more qualitative fused results on the Real-MFF dataset [5] for the multi-focus image fusion task. The Fig.A 6 presents more qualitative fused results on the RoadScene dataset [3] for the visible-infrared image fusion task.

180									180
181		() (11)		() 1000101					181
182	(a) Under-expo (b) Over-expo	(c) CU-Net	(d) DeepFuse	(e) IFCNN	(f) MEFNet	(g) PMGI	(h) U2Fusion	(i) DeFusion	182
183						1	1		183
184			A CONTRACTOR						184
185									185
186									186
187									187
188									188
189									189
190		1 see			1		1	1100	190
191					THE R.L	A SURAL			191
192			- 30			- 30			192
193			P.				H	N N	193
194									194
195		No.	No.	a state	10.00		No.	No.	195
197	NUMIT		N911111	- STE		- AL	Killing Killing	N911111	197
198		and a start of the	and the shadow of the same	VALUE AND A		and the second second	and and and and and an and a	and the state of the second	198
199									199
200									200
201			Phil -					and there are	201
202									202
203		the second	1	12	100	55	and the second	S.	203
204		201	- 2- D	- 24		- 24	-	- 24 -	204
205	THE	T'SAN)	-	TRA	-	-14	7.149	TRAN	205
206				The sector	E			The second second	206
207	A A A	- Add	- Add	A Contraction		- A	- Adda	- Hora	207
208						and the sea			208
209								7	209
210									210
211									211
213		10C	Ne	AC	100	Ni	AC	10C	213
214		Carlos Conto	The set	Partie and South	First of Cale		and an and	The second	214
215		ADD	ALLER OF	A.DE	ALLER A	A.D.	ANALDESS	- Day	215
216			210			A	1975 ····	A COLOR OF COMPANY	216
217		Name of Street	And Street	1			And the second s		217
218		ARE	AAA			haf	hâf	AAA	218
219		Same Same S		frame and	Inter States			mine: Sunnt.	219
220									220
221									221
222				_					222



225	(a) Naar	(b) E	(a) CU N-4	(d) IECNN		(f) DMCI	(g) U2E-star	(b) D - F		22
226	(a) Near	(b) Far	(c) CU-Net		(e) MFFGAN	(I) PMGI	(g) U2Fusion	(f) DeFusion		22
227										22
228		A A				the second se				22
229	Serent .		A Same	Carried S	Constants -	*	Contraction of	Comments)		22
230			-			Contracting of A				23
231										23
232	- And Contraction	Jame	- And	Imme	- And -	- James W		- And	Jame Will	23
233										23
234	LE Sam		KB M	AT Sol		65 M		A Binn		23
235						ALC .				23
236		COLUMN A	NO DE RUA	N. T. S. M. A				CONTRACTOR		23
237										23
238										23
239	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1					-		1 28 19 1		23
240										24
241										24
242				The						24
243	To Marco and	Talling and	T Marrie Co	Comaria Co	The second	The second	and the second second	T. Marriel	Tomas -	24
244	The free		1 Sec							24
245	5									24
246							ili decid			24
247										24
248										24
249			The Plan							24
250	a we se	12.12.30		2				Star Star		25
251										25
252									R.D.R.	25
253										25
254	L S	* 14			TEO					25
255										25
256										25
257										25
258		-								25
259										25
260		Marke Mark	Waxlar Maga		-	Tableton			1	26
261		., /		£ , Y						26
262										26
263	all	- A CH -	"all all	all all	- A CH MI	"all a	- A CHANNEL	- A CH -	in Karl	26
264	Street States	Ser Strack	Sugar State	Sen Dated	and see . S.		Brock States	Store Stand		26
265	and the second	123	State 1	2552	the start		A AN			26
266										26

Fig.A 5: Fused results on the multi-focus image fusion task. We provide the enhanced residual maps for each result of comparison and input images to highlight the difference with GT.

270								270
271	(a) IR	(b) Vis	(c) FusionGAN	(d) IFCNN	(e) PMGI	(f) U2Fusion	(g) DeFusion	271
272								272
273	Trapa		IL - TA		There	III	The second	273
274		1		1				274
275	T	T		Ŧ	and the second s			275
276	and a provide the second	Maria La La	I RANGENS TO A	Manual day		A Share a	MARCH AND AND	276
277				12/ 512	12/ 22-	and the	12 325	277
278		the Tal		the John	Bistone - /	a state of the second s	ane -	278
279								279
280				52 . 1			A H	280
281								281
282								282
283		and a spin of the						283
284			Mar 1					284
285								285
286								286
287				E				287
288	all sign of the second		All Marine and		1	all the second second	ALL STREET	288
289	48	11-6-6	14	- AA	$ \Lambda_{i}$	$ \Delta x$	$-\Lambda_{1}$	289
290		-						290
291								291
292	120	1	× 12	10		12	10	292
293								293
294	Terre Distriction	A CONTRACT OF CONTRACT			THE PARTY IN	The second secon	The second secon	294
295				-				295
297	t _{nyn}		Ante .	AND ANT	Anya.	AND AND		297
298	and the second s		Part Carlos		A ward			298
299		Frank State	1 - Contraction	F	Ft.		H	299
300			And a second second					300
301		international and internationa	7	Kill				301
302	he start	Sector States	he is the second		Kennen (Long- i	302
303	the party of		- The -					303
304			a Research II				A land	304
305		AFT.						305
306		5100	an an	A A				306
307		- Aller		1 - Le faite a	A total	- Harris	1 millio	307
308		A REAL PROPERTY AND IN THE	A MACHANINE MAN	A REAL PROPERTY IN THE	AND ROWNELD RUN	A MARINE MARINE DIAL	A A A A A A A A A A A A A A A A A A A	308
309			18. 18.		· · · · · · · · · · · · · · · · · · ·			309
310		7.		· · · · ·				310
311	2	Mart 34						311
312								312
313	D .				c 1 ·	c •	1	313



8	ECCV-22 supplementary material ID 4260
0	LOOV-22 supplementary material ID 4200

References	315
1 Cai I Cu C Thang I. Learning a deep single image contrast anhance	316
 Cai, J., Gu, S., Zhang, E.: Learning a deep single image contrast emant multi-exposure images. IEEE Transactions on Image Processing 27(4), 20 (2012) 	49–2062 318
(2018) 2 Xio 7 Zhang 7 Cao V Lin V Bao I Vao 7 Dai O Hu H·S	immim:
A simple framework for masked image modeling arXiv preprint arXiv:211	1 09886 ³²⁰
(2021)	321
3. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2Fusion: A unified unsupervise	d image ³²²
fusion network. IEEE Transactions on Pattern Analysis and Machine Inte	elligence ³²³
(2020)	324
4. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., M	uller, J., 325
Manmatha, R., Li, M., Smola, A.: ResNeSt: Split-attention networks. arXiv	preprint 326
arAiv:2004.08955 (2020) 5 Zhang I Liao O Liu S Ma H Vang W Xuo IH: Roal MEE:	A largo
realistic multi-focus image dataset with ground truth Pattern Recognition	Letters 328
138 , 370–377 (2020)	329
6. Zhang, X.: Benchmarking and comparing multi-exposure image fusion alg	orithms. ³³⁰
Information Fusion (2021)	331
	332
	333
	225
	336
	337
	338
	339
	340
	341
	342
	343
	344
	345
	346
	347
	348
	349
	350
	351
	352
	354
	354
	356
	357
	358