Supplementary of "Mumpy: Multilateral Temporal-view Pyramid Transformer for Video Inpainting Detection"

BMVC 2024 Submission # 318

1 More Analysis

Sanity Check. We assess the ability of our method to distinguish between authentic and inpainted frames. We perform experiments on the DVI dataset and train methods using VI+OP, and average the pixel-wise predictions as the frame-level result. As shown in Table 1. We can observe that our method performs best compared with others on all inpainting methods, demonstrating that our method can learn the discriminative inpainting clues. Surprisingly, HiFi-Net does not perform as expected. It may be because of the inappropriate margin setting between authentic and inpainted pixels, leading to a greater concentration on authentic pixels.

Table 1: Sanity check for image-level classification AUC comparison on DVI dataset. Each method is trained using VI and OP inpainting methods.

Methods	**VI***	**OP***	**CP**
HPF	0.718	0.640	0.845
GSRNet	0.762	0.758	0.834
VIDNet	0.778	0.768	0.884
FAST	0.795	0.787	0.898
OSNet	0.992	0.981	0.989
HiFi-Net	0.642	0.699	0.682
Ours	0.996	0.993	0.997

Variants of Interaction Strategies. We further discuss variants of interaction strategies. Denote the global vanilla temporal-view interaction as GVTI, the global deformable temporal-view interaction as GDTI, the window-based vanilla temporal-view interaction as WVTI and the proposed DWTI. Table 2 shows the results of each setting. It can be seen that the window-based strategy outperforms the global-based. It is perhaps because global interaction is more likely influenced by irrelevant semantic features in authentic regions, hindering the information exchange among inpainted regions. Furthermore, the window-based deformable interaction typically outperforms window-based vanilla interaction, indicating the feasibility of deformably selecting features conducive to inpainting traces.

Study on Loss Hyperparameters λ_1, λ_2. Table. 3 shows the results of using different settings of λ_1, λ_2. We can observe that the performance greatly drops if $\mathcal{L}_1$ is not employed. In contrast, $\mathcal{L}_2$ may play an auxiliary effect as the performance has no notable change with

Table 2: Effect of different interaction strategies on DVI dataset.

Type	VI* mIoU/F1	OP mIoU/F1	CP* mIoU/F1
None	0.720/0.820	0.632/0.758	0.820/0.894
GVTI	0.715/0.815	0.559/0.690	0.820/0.893
GDTI	0.731/0.827	0.631/0.748	0.825/0.896
WVTI	0.726/0.826	0.636/0.755	0.823/0.895
DWTI (Ours)	0.727/0.826	0.658/0.768	0.815/0.891

various λ_2. In the main experiment, we select the setting corresponding to the last row as it exhibits better cross-inpainting performance.

Table 3: Study on loss hyperparameters.

$\lambda_1 : \lambda_2$	VI* mIoU/F1	OP mIoU/F1	CP* mIoU/F1
1 : 0	0.728/0.826	0.610/0.730	0.825/0.897
0 : 1	0.698/0.804	0.501/0.637	0.792/0.875
0.1 : 1	0.707/0.810	0.532/0.664	0.814/0.890
1 : 0.1	0.729/0.827	0.627/0.744	0.825/0.897
1 : 1	0.727/0.826	0.658/0.768	0.815/0.891

Effect of n_{group} and k_{offset}. n_{group} denotes the groups of split offset for the diversity of deformed points, and $k_{offsets}$ means the kernel size used in θ (described in Sec.3.2). Table 4 shows the results. Note that n_{heads} denotes the heads of the current view. We can observe that the cross inpainting performance is positively related to k_{offset} and n_{group}, we choose to use setting that $k_{offset} = n_{heads}$ and $n_{groups} = 7$ to train the mumpy.

Table 4: Ablation on WDTI hyperparameters.

$\mathbf{n}_{group}$	$\mathbf{k}_{offset}$	VI* mIoU/F1	OP mIoU/F1	CP* mIoU/F1
1	3	0.732/0.828	0.632/0.750	0.825/0.897
1	5	0.725/0.824	0.653/0.765	0.817/0.892
1	7	0.718/0.818	0.649/0.762	0.821/0.894
3	7	0.729/0.828	0.651/0.764	0.824/0.896
n_{heads}	7	0.727/0.826	0.658/0.768	0.815/0.891

2 More Details in YTVI $\rightarrow$ DVI

In the main experiments on YTVI $\rightarrow$ DVI, our method is trained using three newly added methods (FF, EG2, and PP). In this part, we further validate our method on all combinations of three newly added methods from (FF, EG2, PP, and IS). The results are shown in Table 6, Table 7, and Table 8.

We can observe that our model generally outperforms the competitors by a large margin, averaging 6.2% in IoU and 6.7% in F1 score compared with the second-best OSNet. In particular, our method outperforms the video-based methods VIDNet and FAST by 18% in IoU and 18% in F1 score on average, under the cross-inpainting scenarios of the YTVI

Table 5: **Cross-dataset performance of different methods from YTVI to DVI dataset (YTVI → DVI)**. Each method is trained on YTVI with two inpainting methods and tested on DVI with all three inpainting methods.

Methods	**YTVI**	**DVI**		
		VI	**OP**	**CP**
		mIoU/F1	mIoU/F1	mIoU/F1
HPF	VI+CP	0.52/0.65	0.13/0.20	0.55/0.67
GSRNet		0.46/0.60	0.33/0.47	0.60/0.72
VIDNet		0.22/0.29	0.17/0.24	0.49/0.59
FAST		0.57/0.69	0.41/0.54	0.66/0.78
OSNet		0.58/0.70	0.48/0.60	0.66/0.78
HiFi-Net		0.27/0.35	0.42/0.52	0.68/0.79
IML-ViT		0.60/0.72	0.41/0.54	0.65/0.77
Ours		**0.67/0.78**	**0.66/0.77**	**0.72/0.83**
HPF	OP+CP	0.10/0.16	0.43/0.56	0.57/0.70
GSRNet		0.44/0.56	0.55/0.68	0.66/0.78
VIDNet		0.20/0.27	0.32/0.44	0.51/0.63
FAST		**0.52/0.65**	0.53/0.65	0.62/0.75
OSNet		0.44/0.55	0.57/0.69	0.67/0.78
HiFi-Net		0.06/0.08	0.64/0.73	0.70/0.80
IML-ViT		0.35/0.47	0.59/0.72	0.68/0.79
Ours		0.49/0.61	**0.72/0.83**	**0.71/0.82**

Table 6: **Cross-dataset Cross-inpainting Performance of different methods from YTVI to DVI (YTVI → DVI)**. Each method is trained on FF, EG2 and IS (marked *) in YTVI dataset.

Methods	**YTVI**							**DVI**		
	FF*	**EG2***	**PP**	**IS***	**VI**	**OP**	**CP**	**VI**	**OP**	**CP**
	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1
HPF	0.47/0.59	0.39/0.51	0.27/0.38	0.37/0.49	0.14/0.23	0.08/0.13	0.17/0.26	0.15/0.25	0.12/0.20	0.17/0.27
GSRNet	0.67/0.78	0.60/0.72	0.40/0.54	0.46/0.60	0.07/0.12	0.10/0.16	0.14/0.23	0.55/0.68	0.36/0.50	0.59/0.72
VIDNet	0.61/0.72	0.52/0.64	0.36/0.48	0.55/0.67	0.15/0.25	0.11/0.18	0.28/0.40	0.43/0.56	0.26/0.37	0.44/0.57
FAST	0.48/0.60	0.46/0.58	0.44/0.57	0.30/0.42	0.21/0.31	0.32/0.43	0.39/0.52	0.51/0.65	0.41/0.53	0.53/0.66
OSNet	0.74/0.82	0.61/0.71	0.61/0.71	0.64/0.74	0.27/0.38	0.34/0.44	0.54/0.65	0.65/0.77	0.48/0.61	0.63/0.74
HiFi-Net	0.37/0.49	0.36/0.48	0.33/0.44	0.28/0.39	0.15/0.24	0.21/0.31	0.22/0.32	0.61/0.73	0.53/0.65	0.65/0.76
IML-ViT	0.68/0.79	0.63/0.75	0.58/0.70	0.59/0.70	0.27/0.39	0.34/0.46	0.54/0.67	0.63/0.76	0.48/0.62	0.62/0.75
Ours	**0.77/0.86**	**0.72/0.82**	**0.67/0.78**	**0.69/0.80**	**0.29/0.40**	**0.42/0.53**	**0.61/0.72**	**0.68/0.80**	**0.69/0.80**	**0.71/0.82**

Table 7: **Cross-dataset Cross-inpainting Performance of different methods from YTVI to DVI (YTVI → DVI)**. Each method is trained on FF, PP and IS (marked *) in YTVI dataset.

Methods	**YTVI**							**DVI**		
	FF*	**EG2**	**PP***	**IS***	**VI**	**OP**	**CP**	**VI**	**OP**	**CP**
	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1
HPF	0.48/0.60	0.30/0.41	0.36/0.47	0.39/0.51	0.14/0.22	0.10/0.17	0.19/0.29	0.15/0.25	0.11/0.20	0.14/0.24
GSRNet	0.68/0.79	0.53/0.66	0.56/0.68	0.51/0.64	0.08/0.14	0.14/0.22	0.18/0.28	0.53/0.66	0.35/0.48	0.62/0.74
VIDNet	0.56/0.68	0.45/0.58	0.47/0.60	0.52/0.65	0.15/0.24	0.18/0.27	0.34/0.46	0.36/0.48	0.23/0.34	0.39/0.51
FAST	0.56/0.68	0.49/0.61	0.46/0.58	0.55/0.66	0.25/0.36	0.33/0.45	0.44/0.57	0.50/0.64	0.50/0.63	0.58/0.71
OSNet	0.73/0.81	0.63/0.73	0.63/0.73	0.66/0.76	0.27/0.37	0.35/0.46	0.56/0.67	0.65/0.77	0.48/0.61	0.63/0.74
HiFi-Net	0.41/0.54	0.37/0.49	0.38/0.50	0.30/0.42	0.15/0.24	0.20/0.29	0.22/0.31	0.54/0.67	0.48/0.61	0.65/0.77
IML-ViT	0.69/0.79	0.64/0.75	0.62/0.74	0.62/0.74	0.30/0.43	0.42/0.54	0.59/0.71	0.63/0.72	0.52/0.65	0.62/0.75
Ours	**0.76/0.85**	**0.70/0.81**	**0.68/0.79**	**0.68/0.79**	**0.27/0.38**	**0.42/0.54**	**0.59/0.71**	**0.68/0.80**	**0.68/0.80**	**0.72/0.83**

Table 8: **Cross-dataset Cross-inpainting Performance of different methods from YTVI to DVI (YTVI → DVI)**. Each method is trained on EG2, PP and IS (marked *) in YTVI dataset.

Methods	**YTVI**							**DVI**		
	FF	**EG2***	**PP***	**IS***	**VI**	**OP**	**CP**	**VI**	**OP**	**CP**
	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1	mIoU/F1
HPF	0.38/0.51	0.37/0.49	0.37/0.45	0.38/0.50	0.14/0.23	0.15/0.24	0.22/0.33	0.14/0.23	0.10/0.17	0.13/0.22
GSRNet	0.62/0.73	0.59/0.71	0.54/0.67	0.48/0.61	0.07/0.13	0.14/0.23	0.19/0.29	0.61/0.73	0.35/0.48	0.63/0.76
VIDNet	0.50/0.63	0.50/0.63	0.48/0.60	0.55/0.67	0.16/0.26	0.23/0.33	0.33/0.45	0.31/0.43	0.24/0.35	0.35/0.47
FAST	0.52/0.64	0.51/0.63	0.48/0.60	0.53/0.64	0.21/0.31	0.31/0.42	0.47/0.59	0.51/0.64	0.43/0.55	0.56/0.69
OSNet	0.70/0.79	0.63/0.73	0.67/0.76	0.63/0.73	0.26/0.36	0.36/0.47	0.55/0.66	0.64/0.77	0.50/0.63	0.62/0.73
HiFi-Net	0.34/0.46	0.34/0.46	0.33/0.45	0.31/0.42	0.15/0.23	0.21/0.31	0.18/0.28	0.46/0.59	0.35/0.48	0.49/0.62
IML-ViT	0.68/0.79	0.66/0.77	0.63/0.74	0.62/0.74	0.33/0.45	0.41/0.53	0.60/0.72	0.63/0.76	0.52/0.65	0.63/0.75
Ours	**0.73/0.83**	**0.71/0.82**	**0.69/0.79**	**0.68/0.79**	**0.29/0.40**	**0.41/0.53**	**0.59/0.71**	**0.66/0.78**	**0.67/0.78**	**0.68/0.80**

dataset. It demonstrates the flexible collaboration of spatial and temporal clues can better handle complex scenarios than the fixed version.

We also observe that the performance of all methods under cross-dataset cross-inpainting scenarios is much better than the cross-inpainting scenario only inside the YTVI dataset. This partially demonstrates that the proposed YTVI dataset contains more complex inpainting scenarios, which can easily be generalized to the simple DVI dataset.

3 More Training Details

In the training YTVI, we set the accumulated batch size to 64 and employ the SGD optimizer with a learning rate $1e-3$ for the encoder, $1e-2$ for the decoder, and weight decay $1e-4$ is adopted to optimize the model, and the same learning rate decay strategy as mentioned in the main paper. We set the hyper-parameter λ_1 and λ_2 both to 1. We employ only flip augmentation in the training of cross-dataset cross-inpainting settings and no augmentation for settings on YTVI. When training on the DVI dataset, we adopt common data augmentations such as horizontal and vertical flip, crop, scaled rotate, contrast, brightness, and mixup strategies. The training epoch for the DVI and YTVI dataset is set to 50 and 5 respectively.

4 More Qualitative Results

This section shows more visual results under the settings of in-inpainting, cross-inpainting, and cross-dataset, to qualitatively evaluate our method in both in-domain and cross-domain scenarios.

All the figures are organized as follows: The first row presents the inpainted video frames. From the second to the eighth row, we show the detection results of HPF, GSR-Net, OSNet, HiFi-Net, VIDNet, FAST, and our method (Mumpy). The ninth row shows the corresponding ground truth masks.

In-inpainting Visualization. Fig. 1 and Fig. 2 are the qualitative results of in-inpainting evaluation. All these methods are trained using OP+CP and tested on CP on DVI. Note the examples in Fig. 1 contain more spatial relationships with less movement. It can be seen that our method can obtain more accurate detection results than others, demonstrating the effectiveness of the flexible combination of spatial-temporal clues. Differently, the examples in Fig. 2 have notable movement. The results show that our method can also identify more details compared with others, corroborating our favorable temporal relationship modeling capability.

Cross-inpainting Visualization. Fig. 3 and Fig. 4 show the cross-inpainting qualitative results on the YTVI dataset trained using FF, EG2, and IS inpainting methods, and tested on PP inpainting method. HPF misclassifies the real regions because of the single noise modality, limiting its detection on complex and unseen scenarios. The same phenomenon is observed in VIDNet and FAST, showing the limitation of the fixed combination of spatial and temporal clues. GSRNet and HiFi-Net are easily influenced by relevant semantic features, which may be because only the use of the spatial inpainted features can cause false semantic correlations due to the limited training data. In contrast, our qualitative results significantly outperform others, which can be attributed to the adjustment of contribution strength of spatial and temporal clues helps capture more general inpainting clues.

Cross-dataset Visualization. Fig. 5 shows cross-dataset in-inpainting qualitative results. A similar trend can also be observed that our method can accurately predict the inpainted regions, notably outperforming others. Moreover, in Fig. 6, we visualize the cross-dataset

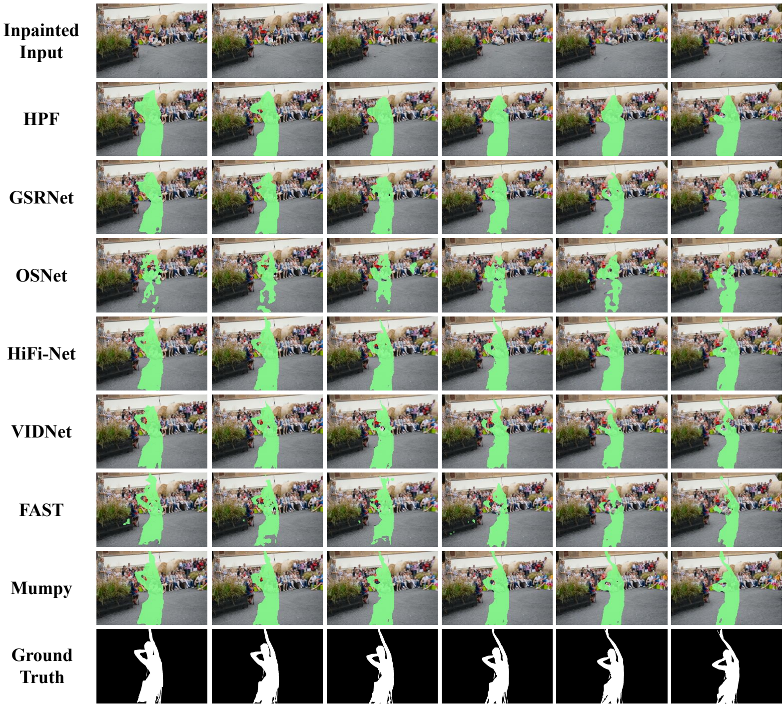

Figure 1: In-inpainting qualitative results on DVI dataset. The model is trained on OP+CP and tested on CP.

cross-inpainting qualitative results to further validate the generalization ability of our method to complex real-world scenarios. It can also observed that our method can obtain better accuracy by emphasizing the contribution of temporal clues.

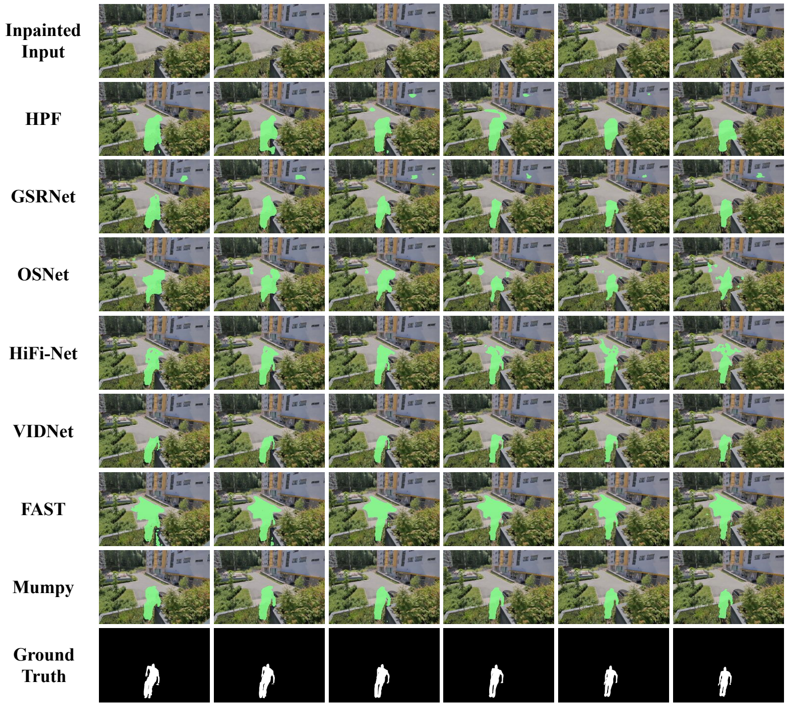

Figure 2: In-inpainting qualitative results on DVI dataset. The model is trained on OP+CP and tested on CP.

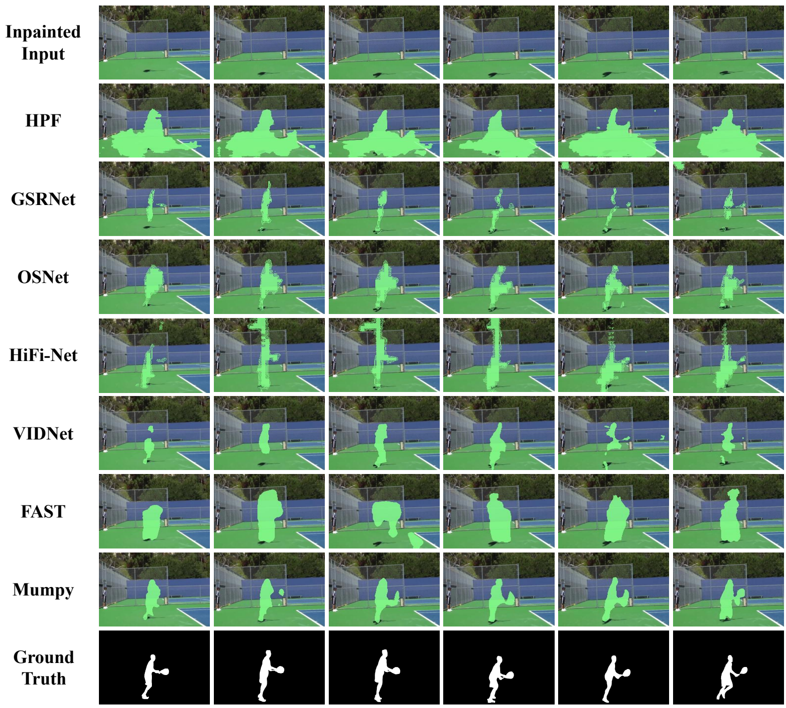

Figure 3: Cross-inpainting qualitative results on YTVI dataset. The model is trained on FF+EG2+IS and tested on PP.

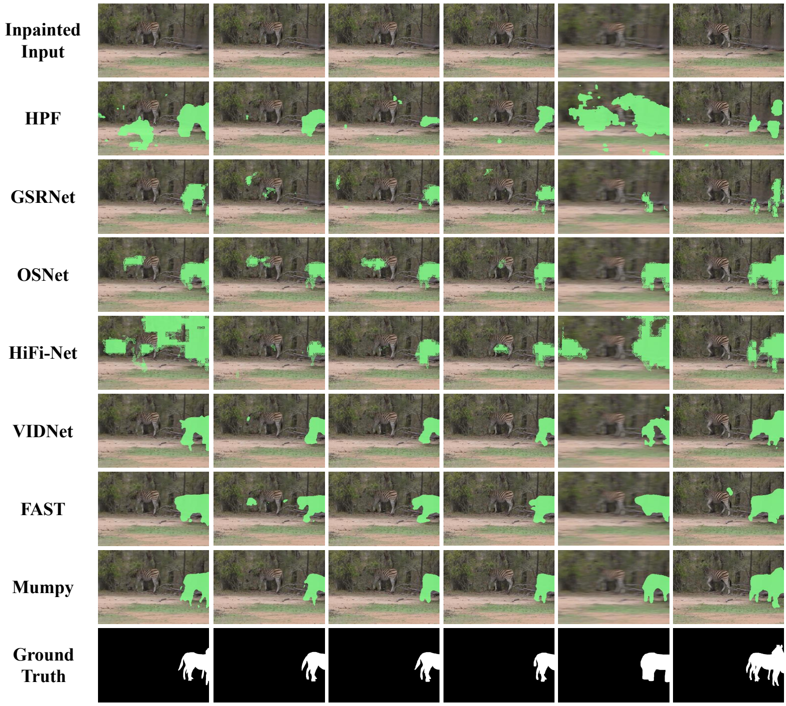

Figure 4: Cross-inpainting qualitative results on YTVI dataset. The model is trained on FF+EG2+IS and tested on PP.

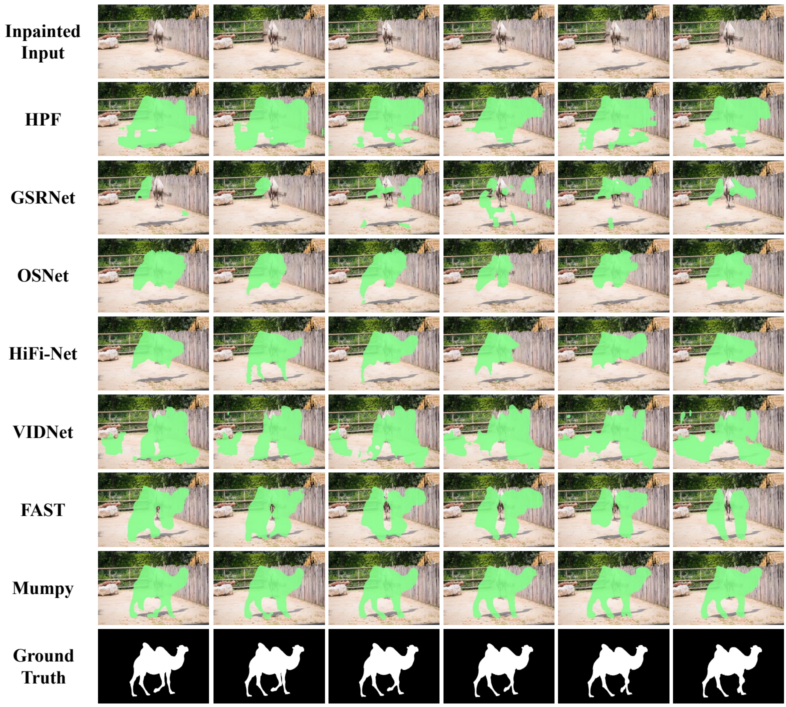

Figure 5: Cross-dataset in-inpainting qualitative results. The model is trained using VI+OP on YTVI dataset and tested using OP on DVI dataset.

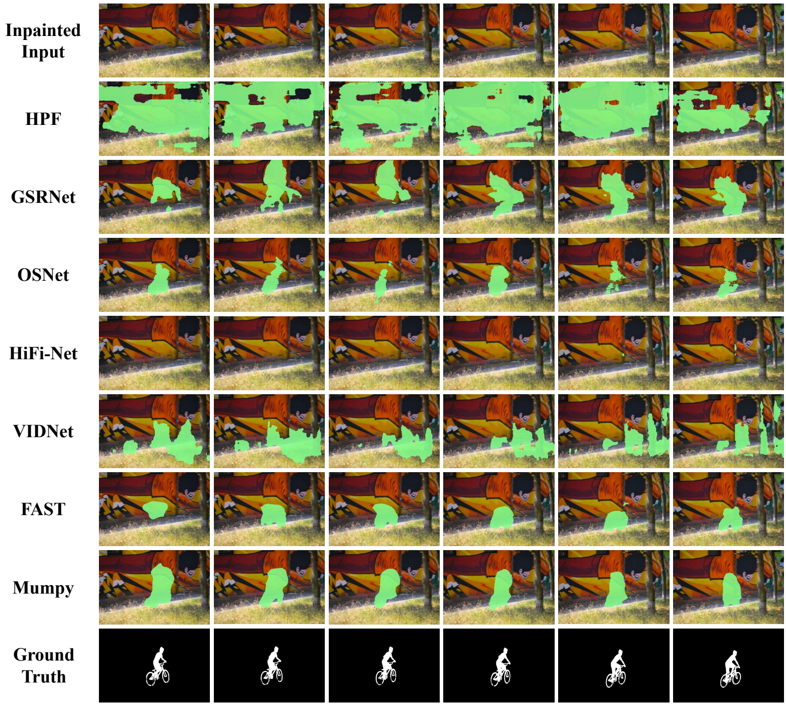

Figure 6: Cross-dataset cross-inpainting Qualitative results. The model is trained using VI+OP on YTVI dataset and tested using CP on DVI dataset.