https://doi.org/10.2352/EI.2022.34.8.IMAGE-287

Correspondences for Image and Video Reconstruction

Xiaoyu Xiang[1] ***, Yapeng Tian***[2]
[1] ***Meta Reality Labs, Menlo Park, CA, 94025***
[2] ***Department of Computer Science, University of Rochester, NY, 14627***

Abstract

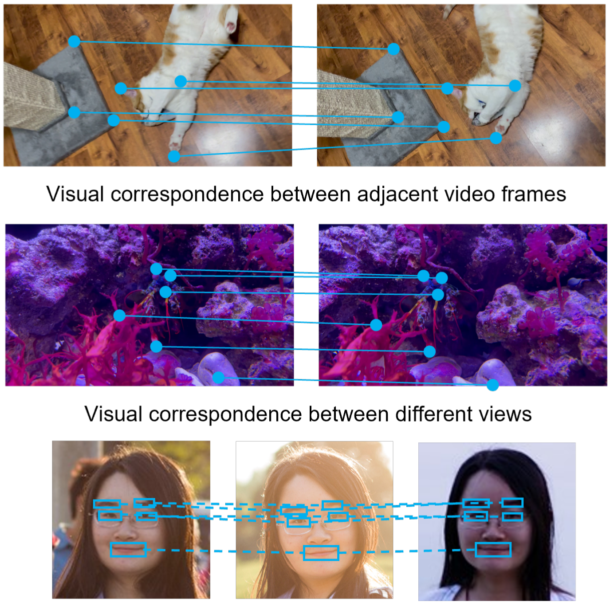
Correspondences are prevalent in natural videos among different frames, as well as a set of images sharing a common attribute. Dense correspondences are important for the core problem of many natural image and video reconstruction tasks: recovering texture details with high fidelity. In this paper, we will discuss recent methods in learning and utilizing such correspondences in image and video reconstruction. Specifically, we decompose the network design into several switchable components of different purposes and discuss their applications to different images and video restoration tasks such as super-resolution, denoising, and video frame interpolation. In this way, we can analyze the performance and uncover the generic and efficient network design. Benefiting from the above investigations, our proposed methods achieve state-of-the-art performance on multiple tasks with fewer parameters. Our findings could inspire the network design of multiple image and video reconstruction tasks for the future.

Introduction

Visual reconstruction aims to restore photo-realistic high-quality images and videos from their degraded counterparts. Since external images and internal neighboring video frames contain rich relevant content with input images or video frames, many image and video reconstruction problems rely on getting a dense estimation of the correspondence between a set of images to leverage useful visual contexts. Such sets of images usually possess some key attributes in common: environment, identity, *etc*. With careful designs, such similarities can be extracted and transferred to recover the fine details of degraded inputs. Compared with blindly recovering this information with priors only, utilizing correspondences can provide substantial information with higher fidelity and assist the reconstruction process.

The process of finding correspondences involves two images: *target image* that to be enhanced, and the *supporting image*. Generally, the supporting images are with similar texture and content structure to the target image. The supporting images could be from adjacent frames in a video [71, 81], images retrieved from the database [89, 86, 95], images from another view [96, 94], or images of the same identity [60, 41, 18, 73]. If we split the target image into separate patches and find the correspondences between them, it would be self-exampled reconstruction [22, 12, 21, 29]. The similarity of the supporting images varies greatly among the above methods. Thus, different matching methods are developed.

Finding correspondences in image sets has been a fundamental yet difficult problem in the computer vision area for the past 50 years. From early works like feature-based parameter estimation [77, 61], correspondence-based transformation [5, 72], optical flow [1, 3, 4], *etc*., to more recent works based on deep learning methods: segment-guided matching [75, 66] and deformable alignment [71, 74, 81]. These approaches use different matching strategies to discover image pixels or patches with similar local structure and texture information. In this paper, we will survey different visual correspondence detection methods and conduct a systematic study to investigate applications of discovering correspondence on image and video reconstruction.


Visual correspondence between adjacent video frames

Visual correspondence between different views

Correspondence between images of the same identity

Figure 1. *Visual correspondences exist widely among images and videos.*

Related Work

Feature Matching

SIFT feature descriptor [49, 50] are adopted in early works [44] to establish the pixel-level correspondence between the target and supporting images. Yue *et al*. [89] first retrieve similar images from the candidate pool and then search the correspondences with global registration and local matching. Usually, these algorithms involve a rectification of image pairs using a group of sparse correspondences. However, the coarse correspondence estimation and simple morphing scheme result in visual artifacts. Besides, such computation can be time-consuming.

Recent works [93, 92, 87, 83] adopt the idea of patch-wise texture searching with features extracted by a pretrained deep neural network, *e.g*. VGG [65]. Zhang *et al*. [93] proposes a deep model-based method that measures the patch similarity between neural features with an inner product and transfers the matched texture to the target with feature swapping. Yang *et al.* [87] utilizes hard attention to select the most similar patch and uses soft attention to transfer the features depending on the similarity.

Optical Flow

Optical flow [25, 3, 20] is widely used to describe motions with displacements at every voxel position between two images. It assumes a brightness constancy constraint. There are many methods for estimating the optical flow: Lucas-Kanade method [51, 52], Horn-Schunck method [28], and recent deep learning-based approaches like SpyNet [63], FlowNet2 [19, 32], and PWCNet [67]. The estimated optical flow is used for warping the supporting images, or directly as inputs in many low-level vision tasks, including reference-based super-resolution (RefSR) [94], video super-resolution (VSR) [43, 36, 47, 69, 26], video frame interpolation (VFI) [57, 33, 2, 84, 58], and video denoising [45, 85, 70]. Still, optical flow may not match long distance correspondences, thus may not perform well in handling large displacements between the target and the supporting images.

Segment-Guided Matching

Unlike dense correspondences between points, segment-guided matching tends to build a region-to-region correspondence, where these regions can be categorized depending on the semantic feature. Wang *et al.* [75] aims to synthesize realistic details for SISR by adopting a semantic segmentation map as a condition for texture feature transform. In this network, the semantic category is mapped to the learned affine transformation parameters that scale and shift the feature maps. For textureless inputs like anime, Li *et al.* [66] exploit global matching among color pieces of different frames to generate smooth interpolation outputs. It uses the color consistency of moving objects across cartoon frames to predict coarse optical flows.

Deformable Alignment

Previous pixel-wise alignment methods usually use optical flow between a target image and a supporting image to wrap the supporting input. Thus, the performance of these image-level wrapping-based models will highly depend on the estimation accuracy of optical flow, and inaccurate optical flow prediction will lead to artifacts in the wrapped supporting images, which also will be propagated into the reconstructed image. To reduce the limitation, deformable alignment-based methods are proposed [71, 74, 81, 64]. With the help of deformable convolution [17], the deformable alignment is able to adaptively align the target image and each supporting image at the feature level without explicitly motion estimation and compensation as in optical flow-based approaches. In particular, Tian *et al.* [71] firstly proposed to use deformable alignment to solve the video reconstruction problem, which utilizes a temporally-deformable alignment network to align video frames for video super-resolution. Later on, Wang *et al.* [74] further enhanced deformable alignment with a pyramid cascaded structure to address video deblurring and super-resolution tasks and Xiang *et al.* [81] incorporates deformable alignment to handle fast video motions for space-time video super-resolution.

Methodology

Representation

To get the most representative features for searching the correspondence and transferring the textures, recent works [93, 87, 71, 81] utilize learned neural features as representation. For CNN networks, features at different depths have different representations and abstraction of the image information [56, 68, 90, 87, 62]: in the early layers, the network perceives more local information; with the network depth grows, the effective receptive field becomes larger. While in vision transformer, the representation between lower and higher layers are more uniform: global information is Incorporated since early layers [62].

Some works [79, 81, 83, 87] adopt several convolution layers to turn the RGB images into feature space, which usually come from shallow layers. To get a better representation across different scales, [68, 90, 79, 87] adopt cross-scale feature integration to exchange information acquired at a different level by striding or downsampling. Other works [93] adopt a pretrained VGG [65] to extract representative features. By choosing outputs from different intermediate layers, it can acquire texture representation at different scales [23, 24]. For image and video reconstruction tasks, existing works [87, 83] have shown that the learned feature representation performs better than ones extracted by pretrained VGGs. In this paper, we explore the influence of such task-oriented feature extractors that are trained with the main network by controlling three key factors: network width and depth, and resolution space. We train 140 VFI networks of different structures for this experiment and evaluate the PSNR and SSIM [76] of the outputs to reflect the performance.

By changing the feature channel from 8 to 16, we plot the experimental results in Figure 2, where the x-axis is the number of trainable parameters (million), and the y-axis is the SSIM. We can observe that increasing the number of feature channels can increase the performance in most cases, while the number of parameters also increases greatly. Changing the feature channels from 8 to 16 can increase the PSNR by 1.24 and the SSIM by 0.0224 on average. Under the same number of trainable parameters, the models with more channels are usually not in the worst-performance tier. Still, they are not always to be the best-performed ones. These results inspire us that under a certain constraint of trainable parameters, we can always choose to increase the number of feature channels as a baseline. With careful designs, it is possible to make the model have better representation capability and thus demonstrate better performance.

For simplification, we choose residual blocks without the batch norm layer as the basic building block and connect them in sequential order for the feature extractor for this experiment. We change the number of feature extraction blocks from 1 to 3 to control the network depth. From Figure 3, it is interesting that models with one block perform better than the ones with three blocks. Under the same number of trainable parameters, the models with one block still perform better. On average, when decreasing the blocks from 3 to 1, the SSIM increases from 0.8691 to 0.8792, which is significant enough. These experiments show that, for this reconstruction task, we do not need deep extractors to acquire good feature representations.

For the common image and video reconstruction task, all inputs are within the same resolution space. But for RefSR, the input LR image and Ref images are with different resolutions and

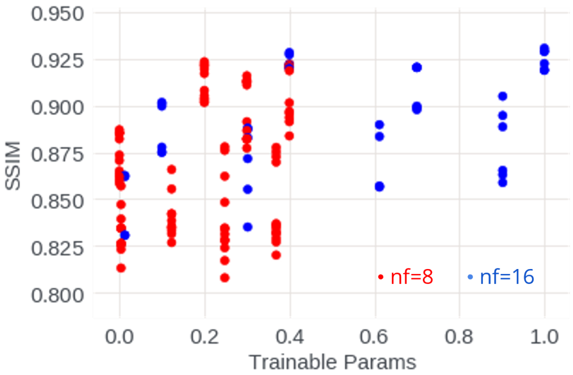


Figure 2. *Influence of the number of feature channels on the VFI task.*

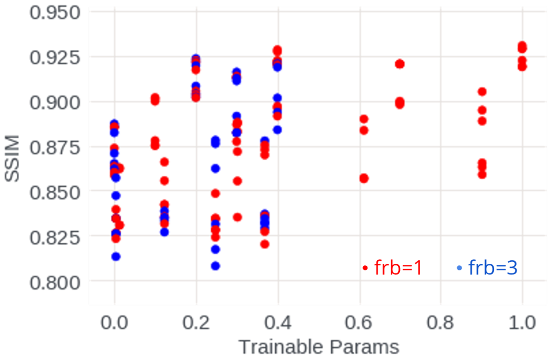


Figure 3. *Influence of the number of feature extraction blocks on the VFI task.*

Table 1. Influence of the feature resolution space on the RefSR task.

Methods	PSNR↑	SSIM↑	LPIPS↓
LR-LR ref	24.38	0.7339	0.2533
LR-ref	24.36	0.7333	0.2537

contain different levels of information. In order to find the correspondences between the LR and Ref images, it is necessary to turn the images or their features into the same resolution space to make them comparable. To investigate the effectiveness of such a design, we construct two models for the RefSR based on the framework in [80]. The first model utilizes the deformable convolution to build the correspondence between the LR and the downsampled Ref features that are both in the LR space. The second model changes the ways of computing the deformable offsets with LR and original Ref features, which is across the LR and HR spaces. According to the results in Table 1, we can see that the first model is slightly better than the second one, which indicates that matching the features in the same space is better than cross-space.

Alignment

Spatial alignment is the core part of the correspondence in reconstruction, which aims to align similar but misplaced pixels or features for transferring the texture. In terms of the alignment method, the existing works can fall into three categories: no-alignment, flow-based alignment, and deformable alignment:

No-Alignment In video reconstruction area, some works [30, 34, 38, 30, 88, 39] do not perform alignment but adopt recurrent modules or 3D convolution when processing consequent video frames. However, the unaligned frames cannot provide useful information for the following processing (like aggregation) for a given perceptive window, which loses the superiority of multi-frame correspondences and thus lead to inferior performance. We verify such influence in our VFI experiment by removing the alignment module from our backbone network. As a substitution, we directly concatenate the unaligned input features and pass them to the reconstruction trunk. Since the common convolution can only provide a very small perceptive field, it is unable to capture the correspondences with large displacement. As shown in the first row in Table 2, the performance is much lower in terms of PSNR and SSIM. This result indicates that it is necessary to acquire and integrate the corresponding information with a large spatial distance.

Flow-based Alignment As introduced in previous Section *Optical Flow*, the optical flow field can provide explicit warping guidance for the supporting inputs. Calculating the optical flow requires the target and the supporting images to be in the same resolution, while the warping operation can be conducted on both the RGB images and the feature space. [2] combines the warping results from the RGB images and the context features to achieve impressive performance. Other works [9, 10] have demonstrated that warping on a feature level performs better than warping on the image level. Thus, we conduct our experiment on the feature level only. We adopt the SpyNet [63] as the optical flow estimator and show the results at the second row of Table 2. Compared with the other alignment modules, it has the least trainable parameters.

Table 2. Influence of the alignment modules on the VFI task.

Alignment	Params (M)	Trainable params (M)	PSNR	SSIM
null	1.7	1.700	31.94	0.9180
SpyNet [63]	3.1	1.700	32.02	0.9308
DCN [81]	2.5	2.500	32.45	0.9259
FDA [11]	4.2	2.800	33.02	0.9381

It performs well in terms of the SSIM, while not for PSNR. These results indicate that the flow-based alignment method is good in retaining the details. Still, since it relies on the pretrained estimator, the alignment accuracy might be influenced when tackling different datasets.

Deformable Alignment Unlike the flow-based alignment, the deformable alignment directly calculates the required offset field from the input features. Thus, the module design is more flexible and does not require the input RGB images. Although the offset field and the flow field are conceptually similar, their ways of building correspondences are different: for flow-based warping, we explicitly copy the value from the source and paste it at the target position. While in the deformable alignment, the kernel of a convolution is modified to acquire longer-range contexts. The corresponding info is convoluted to form the aligned results. Thus, the alignment is more implicit, and the reconstruction results come from the synthesized features. As introduced in Section *Deformable Alignment*, it is widely adopted in RefSR, VFI, and VSR areas for efficient and effective architecture design. We adopt the DCNv2 in our experiment and show the results in the third row in Table 2. Compared with the optical flow network, it achieves a better PSNR with smaller network size. However, the SSIM is worse. These results suggest that the feature synthesis is not good at reconstructing fine details. The better performance in terms of the PSNR might also attribute to the consistency with our training objective, L2 loss.

Besides, the training of the deformable alignment module is difficult and may fail because of instability in the offset estimation. These issues might impede the network's final performance. Along with the above issues, we seek to combine the optical flow and the deformable alignment to reach an improvement. Considering that the learned offset field and the estimated optical flow both represent the correspondences, we can directly use the flow field as the offset field to guide the deformable convolutions. This changes the goal of the DCN from estimating the displacement into the residual of the optical flow estimation. Thus, we use the features pre-warped by the optical flow as inputs. This flow-guided alignment (FDA) model's performance is shown in the last row of Table 2. Adopting both modules does increase the overall model size as well as the number of trainable parameters. Correspondingly, it shows a 0.57 improvement for PSNR compared with the DCN-based model and a 0.0073 improvement in terms of the SSIM compared with the flow-based model. These results suggest that such a combination can reach a balance between the reconstruction accuracy and the fine details, which could be a promising direction for future researches.

We also conduct the above experiments on the RefSR task. Similarly, we switch the alignment module in the backbone network and list the results in Table 3. We experiment on multiple input resolutions (denoting as LR) and scale factors (s). We can observe that such a trend also exists in the RefSR task, which implicates that the improvement of combining the optical flow and deformable alignment can generalize to more image and video reconstruction tasks to build a better correspondence. Still, the performance might be influenced by the input content and the resolution.

Table 3. Influence of the feature alignment module on the RefSR task.

(LR, s)	Methods	PSNR↑	SSIM↑	LPIPS↓
$(32,4)$	DCN	29.11	0.8794	0.1136
	FDA	29.23	0.8817	0.1102
$(64,4)$	DCN	31.24	0.8785	0.1611
	FDA	31.28	0.8789	0.1600
$(16,8)$	DCN	24.54	0.7411	0.2433
	FDA	24.68	0.7467	0.2361

Table 4. Influence of the feature aggregation methods.

Methods	PSNR↑	SSIM↑	LPIPS↓
Average	22.120	0.6350	0.4332
Max-pool	22.118	0.6349	0.4331
CoFA	24.381	0.7339	0.2533

Aggregation

For video reconstruction tasks, the multi-frame information is usually concatenated together and then aggregated with a convolution. This method works well with fixed-length inputs, where the convolution can learn a weighted combination of the given feature maps. However, if the number of supporting frames is unknown, such a method cannot be applied.

In this case, we need to consider a more general method for set aggregation. We construct an application scenario for RefSR with a reference set of any length. Our goal is to find the best set representation for the following reconstruction. From the statistical point of view, we can simply average all the features, or select one feature with the highest similarity, or acquire a weighted average result. We experiment with all three methods and show results in Table 4. In order to learn the aggregation weights, we design a content-conditioned feature aggregation(CoFA) module based on soft attention [80]. By multiplying the feature vector of the target LR and downsampled Ref, we can get a similarity map μ_i for the i-th Ref. Then the aggregated feature can be expressed as:

$$\mathscr{F}(F_1^{ref},\dots,F_n^{ref}|F^L) = \frac{\sum_{i=1}^n \mu_i F_i^{ref}}{\sum_i^n \mu_i}, \tag{1}$$

where $\mathscr{F}$ denotes the aggregation operation, F^L is the LR input, F_i^{ref} is the i-th aligned reference feature map. Therefore, the final representation of the set is a fusion of each feature weighted by its similarity score.

From Table 4, we can observe that the CoFA module improves the PSNR by over 2 dB compared with average and max-pooling. Such performance improvement appears on all metrics, which indicates that our proposed module can learn a better set representation, helping to restore the LR information and enhance the output quality.

Experiments

In this section, we conduct experiments on several images and video reconstruction tasks: Spatio-temporal video super-resolution, burst denoising, video frame interpolation, and reference-based super-resolution. The experimental setups and comparisons with SOTA methods are described in each section.

Spatio-Temporal Video Super-Resolution

The goal of spatio-temporal video super-resolution is to reconstruct more high-resolution frames from the low-resolution inputs. It can also be achieved by two consequent methods: video frame interpolation and video super-resolution. In this task, our network [82] learns to find the correspondences between the input frame and the adjacent supporting frames and interpolates an intermediate frame based on them.

We train our network on the Vimeo90K [85] dataset, which includes over 60,000 7-frame training video sequences. We take out the even-indexed frames. Then we downsample the rest frames by 4× to construct the input sequences and use the original sequences as the ground truth. We compare our results with several two-stage methods that are composed of the SOTA VFI, and VSR works on the widely-used Vid4 [46] testset. Our quantitative comparison results are shown in Figure 5. From this table, we can observe the following facts: both VFI and VSR models matter. Although RBPN and EDVR perform much better than RCAN for SR, however, when equipped with more recent SOTA VFI network DAIN, DAIN+RCAN can achieve a comparable or even better performance than SepConv+RBPN and SepConv+EDVR on the Vimeo-Fast testset; Equipped with the same VFI network DAIN, EDVR keeps achieving better STVSR performance than other SR methods. Moreover, we can observe that our model outperforms the DAIN+EDVR by 0.19dB on Vid4 in terms of PSNR. Such significant improvements demonstrate that our one-stage approach can leverage space-time contexts with better synergy than two-stage methods.

Burst Denoising

Burst denoising leverages the temporal information from multiple frames with minor motions to reconstruct higher-quality targets. In this task, our network [42] adopts an FDA module to find the correspondences between the target input and the other burst frames. We conduct experiments on the grayscale burst denoising dataset [55], where the bursts are generated by applying random translations to the target images from the Zurich Raw to RGB dataset [31]. Then the generated images are corrupted by adding heteroscedastic Gaussian noise. At the training stage, the read and shot noise parameters are uniformly sampled from the following ranges: $\log(\lambda_s) \in [-4, -2]$, and $\log(\sqrt{\lambda_r}) \in [-3, -1.5]$. At inference,we evaluate the network on four gains: [1, 2, 4, 8], which matches noise parameters $(-2.2, -2.6)$, $(-1.8, -2.2)$, $(-1.4, -1.8)$, and $(-1.1, -1.5)$, respectively.

The comparison of our method and other SOTA methods is shown in Table 6. By comparing with the other results, we can see that our method achieves a significant improvement.

Video Frame Interpolation

Video interpolation aims to reconstruct the missing intermediate frame(s) from the given inputs. Thus, the network needs to model the temporal correspondences and estimate the missing info based on them. For this experiment, our model aims to reconstruct one in-between frame from two given inputs.

We train our network on the Vimeo90K [85] dataset, which has been introduced in the above sections. We randomly sample a 3-frame sequence and take out the middle one to construct the inputs for training. We also augment the sequence temporally by increasing the sequence interval so that the acquired sample is with larger motion, which is more challenging for the network. For a fair comparison, all the compared methods are trained with the same data and hyper-parameters. We evaluate the performance of different methods on the Vimeo90K testset, which contains 7815 clips of 7 frames. The results are shown in Table 7. We report two designs of our model in this table: one large model with the FDA module and one small model with the DCN for alignment. From this table, we can observe that our large model outperforms all the other methods by a large margin in terms of PSNR and SSIM while keeping a relatively small model size. Especially, our small model achieves the third-best performance with 4% to 24% parameters of the other methods. These results demonstrate the effectiveness and efficiency of our models.

Reference-based Super-Resolution

Reference-based super-resolution aims to improve the reconstruction of the LR target image with high-resolution references. Thus, the correspondence is built across two images of different resolutions. To investigate the influence of feature aggregation, we choose multiple HR images as references in previous sections. Here we keep the same experimental setting since the network may benefit from the rich information among various exemplars.

The CelebAMask-HQ is used as the training and test datasets [37], which includes over 30,000 1024×1024 faces selected from the CelebA dataset [48]. To construct the LR-ref image sets, We read the identity information from the original CelebA meta info and remove 3,300 out of 6,217 identities with less than four images. Then we randomly split the remaining identities into a training set and a test set of 2,600 and 287 identities, respectively. We synthesize images of different scales by bicubic downsampling with factor $s = [4, 8]$. The results are shown in Table 8. Since the experiment is conducted on face images, we choose both RefSR and face hallucination methods for comparison. Our model greatly outperforms the other methods by a large margin.

Ablation Studies

In this section, we discuss two details in the network design: long-range skip connection and the influence of choosing different optical flow estimators. Although they might not make the key contribution of the paper, they do substantially influence the network performance.

Long-range Skip Connection

The long-range skip connection refers to adding the input to acquire the final result. For SR tasks where the input and target images are of different resolutions, we first upsample the input with naive methods before the addition. For VFI-related tasks that miss the corresponding frame, we will either add one of the inputs or an overlap of the two inputs. This skip connection fundamentally changes the learning goal of the main network from reconstructing the image itself to recovering its residue. Such a long-

Table 5. Quantitative comparison of two-stage VFI and VSR methods and our results on Vid4 [46] dataset. The best results are highlighted in bold. We measure the total runtime on the entire Vid4 dataset [46]. Note that we omit the baseline methods with Bicubic when comparing in terms of runtime.

VFI	SR	Parameters	Runtime-VFI	Runtime-SR	Total	Average	Vid4	
Method	Method	(Million)	(s)	(s)	Runtime (s)	Runtime (s/frame)	PSNR	SSIM
SuperSloMo [33]	Bicubic	19.8	0.28	-	-	-	22.84	0.5772
SuperSloMo [33]	RCAN [91]	19.8+16.0	0.28	68.15	68.43	0.4002	23.80	0.6397
SuperSloMo [33]	RBPN [26]	19.8+12.7	0.28	82.62	82.90	0.4848	23.76	0.6362
SuperSloMo [33]	EDVR [74]	19.8+20.7	0.28	24.65	24.93	0.1458	24.40	0.6706
SepConv [59]	Bicubic	21.7	2.24	-	-	-	23.51	0.6273
SepConv [59]	RCAN [91]	21.7+16.0	2.24	68.15	70.39	0.4116	24.92	0.7236
SepConv [59]	RBPN [26]	21.7+12.7	2.24	82.62	84.86	0.4963	26.08	0.7751
SepConv [59]	EDVR [74]	21.7+20.7	2.24	24.65	26.89	0.1572	25.93	0.7792
DAIN [2]	Bicubic	24.0	8.23	-	-	-	23.55	0.6268
DAIN [2]	RCAN [91]	24.0+16.0	8.23	68.15	76.38	0.4467	25.03	0.7261
DAIN [2]	RBPN [26]	24.0+12.7	8.23	82.62	90.85	0.5313	25.96	0.7784
DAIN [2]	EDVR [74]	24.0+20.7	8.23	24.65	32.88	0.1923	26.12	0.7836
Ours [82]		**11.10**	-	-	**10.36**	**0.0606**	**26.49**	**0.8028**

Table 6. Comparison between different method for burst denoising on the gray scale burst denoising dataset [55]

Gain	HDR+	BM3D	NLM	VBM4D	Single	KPN	MKPN	BPN	Deep-Rep	Ours [42]
	[27]	[16]	[7]	[53]	Image	[55]	[54]	[78]	[6]	
Gain ∝ 1	31.96	33.89	33.23	34.60	35.16	36.47	36.88	38.18	39.37	**39.67**
Gain ∝ 2	28.25	31.17	30.46	31.89	32.27	33.93	34.22	35.42	36.51	**36.63**
Gain ∝ 4	24.25	28.53	27.43	29.20	29.34	31.19	31.45	32.54	33.38	**33.52**
Gain ∝ 8	20.05	25.92	23.86	26.52	25.81	27.97	28.52	29.45	29.69	**29.75**
Average	26.13	29.88	28.75	30.55	30.65	32.39	32.77	33.90	34.74	**34.89**

Table 7. Comparison with SOTA methods on VFI task

Methods	Number of Params (M)	Vimeo-90K	
		PSNR	SSIM
ZSM [81]	7.6	35.10	0.956
DAIN [2]	24.0	33.35	0.945
FLAVR [35]	42.1	32.22	0.929
Ours (large)	4.2	**35.47**	**0.959**
Ours (small)	1.8	34.85	0.955

range skip connection is easy to conduct without too much extra consumption of computation or storage. Besides, such shortcuts can pass the abundant information from the inputs, thus making the whole network easier to optimize.

We conduct experiments on 140 models: half of them are with the long-range skip connection, while the other half are without. We plot the results in Figure 4, where the x-axis is the number of trainable parameters, and the y-axis is the SSIM. It is obvious that the models with skip connections perform better than the ones without. On average, adding the long-range skip connection improves the PSNR by 1.67, SSIM by 0.0184 without increasing the model size. Such improvement is very significant, and the required operation is almost a "free lunch". This finding can inspire us to design models for other reconstruction tasks.

Influence of Optical Flow Estimator

In this section, we try to answer the question of how much the optical flow estimator can influence the final reconstruction performance. We swap the optical flow network of the FDA mod-

Table 8. Quantitative comparison of our results and other SOTA methods. The best results are shown in bold.

(LR, *s*)	Methods	PSNR	SSIM
$(32,4)$	Bicubic	25.64	0.7752
	SRNTT [93]	28.02	0.8434
	TTSR [87]	27.31	0.8346
	SPARNet [13]	20.50	0.6118
	PSFR-GAN [14]	25.47	0.7709
	Ours [80]	**29.23**	**0.8817**
$(64,4)$	Bicubic	28.40	0.8169
	SRNTT [93]	30.41	0.8552
	TTSR [87]	29.87	0.8484
	SPARNet [13]	23.26	0.6990
	PSFR-GAN [14]	26.62	0.7685
	DFDNet [40]	21.55	0.6587
	Ours [80]	**31.28**	**0.8789**
$(16,8)$	Bicubic	21.83	0.5929
	PFSR[8]	21.44	0.5778
	FSRNet [15]	20.03	0.5749
	GWAINet [18]	21.96	0.5844
	SPARNet [13]	19.00	0.5022
	PSFR-GAN [14]	22.05	0.6102
	Ours [80]	**24.68**	**0.7467**

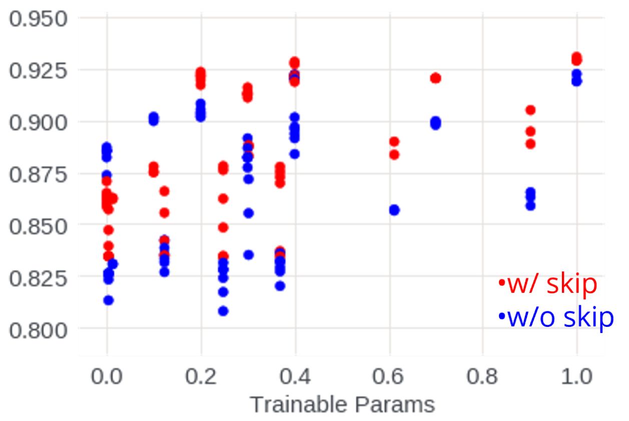


Figure 4. Influence of long-range skip connection.

Table 9. Performance comparison of the FDA module using different optical flow estimators.

Alignment	Params (M)	Trainable params (M)	PSNR	SSIM
PWCNet [67]	9.8	0.397	33.31	0.9436
SpyNet [63]	1.8	0.397	33.61	0.9471

ule: PWCNet [67], and SpyNet [63] while controlling the other parts of the network unchanged. By comparing the results in Table 9, we can see that the model with SpyNet performs slightly better than the other one, which suggests that a suitable optical flow estimator can influence the final reconstruction results to some extent.

In all the previous experiments, the weights of the optical flow estimator are frozen. In the following experiment, we investigate the influence of freezing/unfreezing the parameters during training. We conduct experiments on the simple flow-based feature warping module and the FDA module, and put the results in Table 10. Note that although the number of trainable parameters greatly increases for the unfrozen models, the inference cost remains the same. All results become better when unfreezing all parameters: for the flow-based warping module, the PSNR increases by 2.86, and SSIM improves by 0.063. For the FDA module, the PNSR improves by 1.24, and SSIM leverages by 0.0082. These results provide us with inspiration for more general image and video reconstruction: it is worth training a reconstruction-oriented flow model for multiple image/video reconstruction tasks.

Table 10. Performance comparison of freezing/unfreezing the optical flow estimator during training.

Alignment	Params (M)	Trainable params (M)	PSNR	SSIM
SpyNet-f	1.5	0.013	29.60	0.8896
SpyNet-u	1.5	1.5	32.46	0.9257
FDA-SpyNet-f	1.8	0.397	33.61	0.9471
FDA-SpyNet-u	1.8	1.8	34.85	0.9553

Conclusion

This work focuses on several general topics of learning the correspondences for image and video reconstruction tasks. We revisit various components: feature extraction, alignment, and aggregation modules. With experiments across many different tasks, we uncover the strength of existing approaches and propose our unique solutions that outperform existing state-of-the-art methods with high efficiency. With the modular designs, our models can serve as good baselines for upcoming researches. We believe that the findings for each component may inspire more future works on architecture design and have the potential to be extended to other image and video reconstruction tasks.

References

[1] JK Aggarwal and N Nandhakumar. On the computation of motion from sequences of images-a review. *Proceedings of the IEEE*, 76(8):917–935, 1988.

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019.

[3] John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.

[4] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466, 1995.

[5] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. *ACM SIGGRAPH computer graphics*, 26(2):35–42, 1992.

[6] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proc. ICCV*, pages 2460–2470, 2021.

[7] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Proc. CVPR*, 2005.

[8] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. In *Proceedings of the European Conference on Computer Vision*, pages 185–200, 2018.

[9] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4:3, 2020.

[10] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021.

[11] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. 2021.

[12] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[13] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231, 2020.

[14] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. *arXiv preprint arXiv:2009.08709*, 2020.

[15] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018.

[16] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *TIP*, 2007.

[17] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[18] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[19] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[20] David Fleet and Yair Weiss. Optical flow estimation. In *Handbook of mathematical models in computer vision*, pages 237–257. Springer, 2006.

[21] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011.

[22] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[23] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28:262–270, 2015.

[24] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[25] James J Gibson. The perception of the visual world. 1950.

[26] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.

[27] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM ToG*, 35(6):1–12, 2016.

[28] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[29] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[30] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *Advances in neural information processing systems*, 28:235–243, 2015.

[31] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proc. CVPR Workshops*, pages 536–537, 2020.

[32] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017.

[33] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.

[34] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018.

[35] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. *arXiv preprint arXiv:2012.08512*, 2020.

[36] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE transactions on computational imaging*, 2(2):109–122, 2016.

[37] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[38] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019.

[39] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *European Conference on Computer Vision*, pages 335–351. Springer, 2020.

[40] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Proceedings of the European Conference on Computer Vision*, pages 399–415, 2020.

[41] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European Conference on Computer Vision*, pages 272–289, 2018.

[42] Yawei Li, Lucas Young, Yuchen Fan, Xiaoyu Xiang, Jingyun Liang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Redesigning hypernetworks and self-attention for image and burst denoising. unpublished, 2021.

[43] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–539, 2015.

[44] Christian Lipski, Christian Linz, Thomas Neumann, Markus Wacker, and Marcus Magnor. High resolution image correspondences for video post-production. In *2010 Conference on Visual Media Production*, pages 33–39. IEEE, 2010.

[45] Ce Liu and William T Freeman. A high-quality video denoising algorithm based on reliable motion estimation. In *European conference on computer vision*, pages 706–719. Springer, 2010.

[46] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 209–216. IEEE, 2011.

[47] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang

Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2507–2515, 2017.

[48] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015.

[49] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[50] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[51] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia, 1981.

[52] Bruce David Lucas. *Generalized image matching by the method of differences*. Carnegie Mellon University, 1985.

[53] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IEEE TIP*, 21(9):3952–3966, 2012.

[54] Talmaj Marinč, Vignesh Srinivasan, Serhan Gül, Cornelius Hellge, and Wojciech Samek. Multi-kernel prediction networks for denoising of burst images. In *Proc. ICIP*, pages 2404–2408. IEEE, 2019.

[55] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proc. CVPR*, pages 2502–2510, 2018.

[56] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

[57] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1710, 2018.

[58] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020.

[59] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *IEEE Int. Conf. Comput. Vis.*, pages 261–270, 2017.

[60] Jeong-Seon Park and Seong-Whan Lee. An example-based face hallucination method for single-frame, low-resolution facial images. *IEEE Transactions on Image Processing*, 17:1806–1816, 2008.

[61] Richard Radke, Vitali Zagorodnov, Sanjeev Kulkarni, and Peter J Ramadge. Estimating correspondence in digital video. In *Proceedings International Conference on Information Technology: Coding and Computing*, pages 196–201. IEEE, 2001.

[62] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[63] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[64] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2020.

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[66] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6587–6595, 2021.

[67] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[68] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[69] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.

[70] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809. IEEE, 2019.

[71] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[72] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992.

[73] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Multiple exemplars-based hallucination for face super-resolution and editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[74] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[75] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018.

[76] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[77] Juyang Weng, Narendra Ahuja, and Thomas S. Huang. Matching two perspective views. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(08):806–825, 1992.

[78] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *Proc. CVPR*, pages 11844–11853, 2020.

[79] Xiaoyu Xiang, Qian Lin, and Jan P Allebach. Boosting high-level vision with joint compression artifacts reduction and super-resolution. *arXiv preprint arXiv:2010.08919*, 2020.

[80] Xiaoyu Xiang, Jon Morton, Fitsum Reda, Lucas Young, Federico Perazzi, Rakesh Ranjan, Amit Kumar, Andrea Colaco, and Jan Allebach. Hime: Headshot image super-resolution with multiple exemplars. unpublished, 2021.

[81] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3370–3379, 2020.

[82] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slowmo: An efficient one-stage framework for space-time video super-resolution. *arXiv preprint arXiv:2104.07473*, 2021.

[83] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *European Conference on Computer Vision*, pages 230–245, 2020.

[84] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *arXiv preprint arXiv:1911.00627*, 2019.

[85] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

[86] Xu Yan, Weibing Zhao, Kun Yuan, Ruimao Zhang, Zhen Li, and Shuguang Cui. Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 52–68. Springer, 2020.

[87] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[88] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3106–3115, 2019.

[89] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, 22(12):4865–4878, 2013.

[90] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019.

[91] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. Eur. Conf. Comput. Vis.*, pages 286–301, 2018.

[92] Yulun Zhang, Zhifei Zhang, Stephen DiVerdi, Zhaowen Wang, Jose Echevarria, and Yun Fu. Texture hallucination for large-factor painting super-resolution. *arXiv preprint arXiv:1912.00515*, 2019.

[93] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[94] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision*, pages 88–104, 2018.

[95] Yu Zhu, Yanning Zhang, and Alan L Yuille. Single image super-resolution using deformable patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2917–2924, 2014.

[96] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004.

Author Biography

Xiaoyu Xiang *received a Ph.D. degree in the department of electrical and computer engineering at Purdue University in 2021, a B.E. degree in engineering physics from Tsinghua University in 2015. She is currently a research scientist in the Meta Reality Lab. Her primary area of research has been image and video restoration.*

Yapeng Tian *received the B.E. degree in electronic engineering from Xidian University, Xian, China, in 2013, M.E. degree in electronic engineering at Tsinghua University, Beijing, China, in 2017, and is currently working toward a Ph.D. degree in the department of computer science at the University of Rochester, USA. His research interests include audio-visual scene understanding and low-level vision.*