

MuCAN: Multi-Correspondence Aggregation Network for Video Super-Resolution

Wenbo Li¹, Xin Tao², Taian Guo³, Lu Qi¹, Jiangbo Lu⁴, and Jiaya Jia^{1,4}

¹The Chinese University of Hong Kong

²Kuaishou Technology ³Tsinghua University ⁴Smartmore Technology

{wenboli, luqi, leojia}@cse.cuhk.edu.hk jiangsutx@gmail.com

gta17@mails.tsinghua.edu.cn jiangbo@smartmore.com

Abstract. Video super-resolution (VSR) aims to utilize multiple low-resolution frames to generate a high-resolution prediction for each frame. In this process, inter- and intra-frames are the key sources for exploiting temporal and spatial information. However, there are a couple of limitations for existing VSR methods. First, optical flow is often used to establish one-on-one temporal correspondences. But flow estimation itself is error-prone and hence largely affects the ultimate recovery result. Second, similar patterns existing in natural images are rarely exploited for the VSR task. Motivated by these findings, we propose a temporal multi-correspondence aggregation strategy to leverage most similar patches across frames, and also a cross-scale nonlocal-correspondence aggregation scheme to explore self-similarity of images across scales. Based on these two novel modules, we build an effective multi-correspondence aggregation network (MuCAN) for VSR. Our method achieves state-of-the-art results on multiple benchmark datasets. Extensive experiments justify the effectiveness of our method.

Keywords: Video Super-Resolution · Correspondence Aggregation

1 Introduction

Super-resolution (SR) is a fundamental task in image processing and computer vision, which aims to reconstruct high-resolution (HR) images from low-resolution (LR) ones. While single-image super-resolution methods design natural and clear structures mostly based on spatial information, video super-resolution (VSR) extends to exploit temporal information from multiple neighboring frames to recover missing details. Nowadays, VSR is widely applied in video surveillance, satellite imagery, etc.

Early methods [24, 26] for VSR propose delicate image models, which are solved via various optimization techniques. Recent deep neural network based VSR methods [19, 22, 38, 41, 37, 2, 23, 26, 11, 32] further push the limits and set new state-of-the-arts.

In contrast to previous methods that model VSR as separate alignment and regression stages, we view this problem as a kind of inter- and intra-frame correspondence aggregation task. Based on the fact that consecutive frames share

similar content, and different locations within a single frame may contain similar structures (known as self-similarity [20, 29, 9]), we propose to aggregate these similar contents from multiple correspondences to better restore HR results.

Inter-frame Correspondence Motion compensation (or alignment) is usually an important component for most video tasks to handle displacements between frames. A majority of methods [2, 37, 32] design specific sub-networks for optical flow estimation. In [19, 22, 15, 14], motion is implicitly handled using Conv3D or recurrent networks. Recent methods [38, 41] utilize deformable convolution layers [3] to explicitly align feature maps using learnable offsets. All the methods establish explicit or implicit one-on-one pixel correspondences between frames. However, motion estimation may suffer from inevitable errors and there is no chance for wrongly estimated mapping to locate correct pixels. Thus, we advocate that a better solution may be possible when considering multiple candidate correspondences for a pixel at a time, as illustrated in Figure 1(a).

In order to validate this point, we estimate optical flow with a simple patch-matching strategy on the MPI Sintel Flow dataset. After obtaining top- K most similar patches as candidate correspondences for the objective, we calculate the Euclidean distance between the best-performing one and ground-truth flow. As shown in Figure 1(b), it is clear that a better result is obtained by taking into consideration more correspondences for a pixel. Inspired by this, we propose a *temporal multi-correspondence aggregation module* (TM-CAM) for alignment. It uses top- K most similar feature patches as supplement. More specifically, we design a pixel-adaptive aggregation strategy, which will be detailed in Sec. 3. Our module is lightweight and, more interestingly, can be easily integrated into common frameworks. It is robust to visual artifact production, as shown in Sec. 4.

Intra-frame Correspondence From another perspective, similar patterns within each frame as shown in Figure 1(c) can also benefit detail restoration, which has been verified in several previous low-level tasks [20, 29, 9, 44, 47, 13]. This line is still new for VSR. For existing methods in VSR, the commonly used way to explore intra-frame information is to introduce a U-net-like [31] or deep structure, so that a large but still local receptive field is covered. We notice that valuable information may not always come from neighboring positions. Similar patches within nonlocal locations or across scales may also be beneficial.

Accordingly, we in this paper design a new *cross-scale nonlocal-correspondence aggregation module* (CN-CAM) to exploit the multi-scale self-similarity property of natural images. It aggregates similar features across different levels to recover more details. The effectiveness of this module is verified in Sec. 4.

The contribution of this paper is threefold.

- We design a multi-correspondence aggregation network (MuCAN) to deal with video super-resolution in an end-to-end manner. It achieves state-of-the-art performance on multiple benchmark datasets.
- Two effective modules are proposed to make good use of temporal and spatial information. The temporal multi-correspondence aggregation module

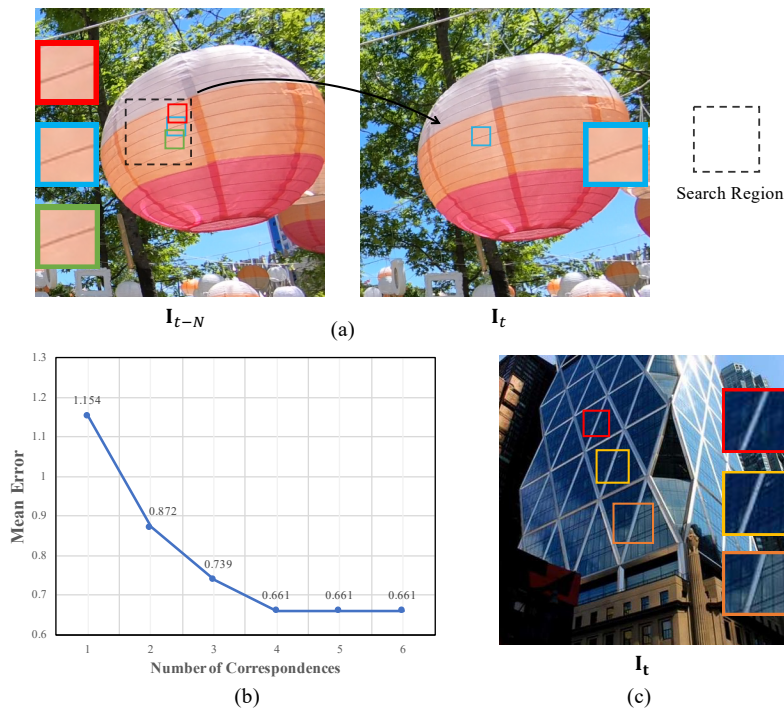


Fig. 1. Inter- and intra-frame correspondence. (a) Inter-frame correspondence estimated from temporal frames (e.g., I_{t-N} at time $t - N$ and I_t at time t) can be leveraged for VSR. For a patch in I_t , there are actually multiple similar patterns within a co-located search region in I_{t-N} . (b) Mean error of optical flow estimated by different numbers of inter-frame correspondences on MPI Sintel Flow dataset. (c) Similar patterns existing over different scales in an image I_t .

(TM-CAM) conducts motion compensation in a more robust way. The cross-scale nonlocal-correspondence aggregation module (CN-CAM) explores similar features from multiple spatial scales.

- We introduce an edge-aware loss that enables the proposed network to generate better refined edges.

2 Related Work

Super-resolution is a classical task in computer vision. Early works used example-based [8, 9, 46, 7, 39, 40, 33], dictionary learning [45, 28] and self-similarity [44, 13] methods. Recently, with the rapid development of deep learning, super-resolution has reached to a new level. In this section, we briefly discuss deep learning based approaches from two lines, i.e., single-image super-resolution (SISR) and video super-resolution (VSR).

2.1 Single-Image Super-Resolution

SRCNN [4] is the first method that employs a deep convolutional network in the super-resolution task. This work has inspired several following methods [5, 17, 34, 18, 21, 36, 10, 48, 49]. For example, Kim *et al.* [17] proposed a residual learning strategy using a 20-layer depth network, which shows a significant improvement in accuracy. Instead of applying commonly used bicubic interpolation, Shi *et al.* [34] designed a sub-pixel convolution network to effectively upsample low-resolution inputs. This operation reduces the computational complexity and enables a real-time network. Taking advantage of high-quality large image datasets, more networks such as DBPN [10], RCAN [48], and RDN [49] were proposed to further improve the performance of SISR.

2.2 Video Super-Resolution

Video super-resolution takes multiple frames into consideration. Based on the way to aggregate temporal information, previous methods can be roughly grouped into three categories.

The first group of methods process video sequences without any explicit alignment. For example, methods of [19, 22] utilize 3D convolutions to directly extract features from multiple frames. Although this approach is simple, the computational cost is typically high. Jo *et al.* [15] proposed dynamic upsampling filters to avoid explicit motion compensation. However, it stands the chance of ignoring informative details of neighboring frames. Noise in the misaligned regions can also be harmful.

The second line [23, 26, 16, 2, 25, 37, 32, 11] is to use optical flow to compensate motion between frames. Methods of [23, 16] first obtain optical flow using classical algorithms and then build a network for high-resolution image reconstruction. Caballero *et al.* [2] integrated these two steps into a single framework and trained it in an end-to-end way. Tao *et al.* [37] further proposed sub-pixel motion compensation to reveal more details. No matter optical flow is predicted independently or not, this category of methods needs to handle two relatively separated tasks. Besides, the estimated optical flow critically affects the quality of reconstruction. Because optical flow itself is a challenging task especially for large motion scenes, the resulting accuracy cannot be guaranteed.

The last line [38, 41] conducts deformable convolution networks [3] to accomplish video super-resolution. For example, EDVR proposed in [41] extracts and aligns features at multiple levels, and achieves reasonable performance. The deformable network is however sensitive to the input patterns, and may give rise to noticeable reconstruction artifacts due to unreasonable offsets.

3 Our Method

The architecture of our proposed multi-correspondence aggregation network (MuCAN) is illustrated in Figure 2. Given $2N + 1$ consecutive low-resolution

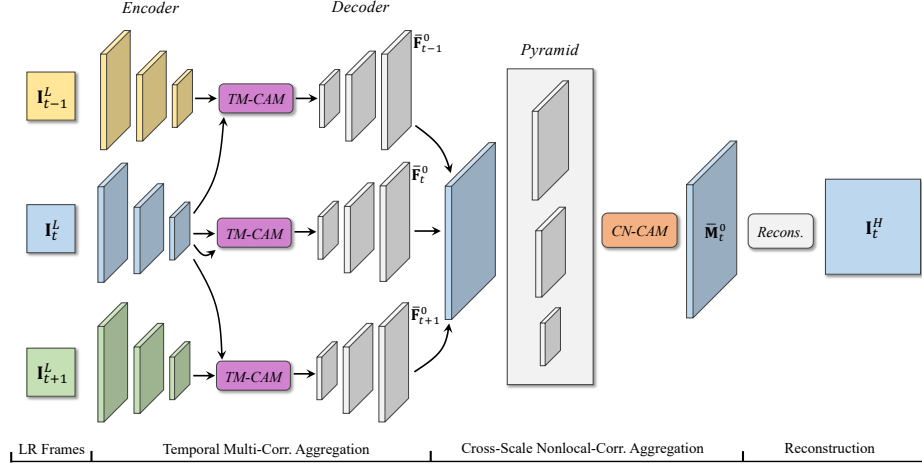


Fig. 2. Architecture of our multi-correspondence aggregation network (MuCAN). It contains two novel modules: temporal multi-correspondence aggregation module (TM-CAM) and cross-scale nonlocal-correspondence aggregation module (CN-CAM).

frames $\{\mathbf{I}_{t-N}^L, \dots, \mathbf{I}_t^L, \dots, \mathbf{I}_{t+N}^L\}$, our framework predicts a high-resolution central image \mathbf{I}_t^H . It is an end-to-end network consisting of three modules: a temporal multi-correspondence aggregation module (TM-CAM), a cross-scale nonlocal-correspondence aggregation module (CN-CAM), and a reconstruction module. The details of each module are given in the following subsections.

3.1 Temporal Multi-Correspondence Aggregation Module

Camera or object motion between neighboring frames has its pros and cons. On the one hand, large motion needs to be eliminated to build correspondences among similar contents. On the other hand, the accuracy of small motion (at sub-pixel level) is very important, which is the source to draw details. Inspired by the work of [30, 35], we design a hierarchical correspondence aggregation strategy to handle large and subtle motion simultaneously.

As shown in Figure 3, given two neighboring LR images \mathbf{I}_{t-1}^L and \mathbf{I}_t^L , we first encode them into lower resolutions (level $l = 0$ to $l = 2$). Then, the aggregation starts in the high-level/low-resolution stage (i.e., from $\bar{\mathbf{F}}_{t-1}^{l=2}$) compensating large motion, while progressively moving up to low-level/high-resolution stages (i.e., to $\bar{\mathbf{F}}_{t-1}^{l=0}$) for subtle sub-pixel shift. Different from many methods [2, 37, 32] that directly regress flow fields in the image space, our module functions in the feature space. It is more stable and robust to noise [35].

The aggregation unit in Figure 3 is detailed in Figure 4. A patch-based matching strategy is used since it naturally contains structural information. As aforementioned in Figure 1(b), one-on-one mapping may not be able to capture

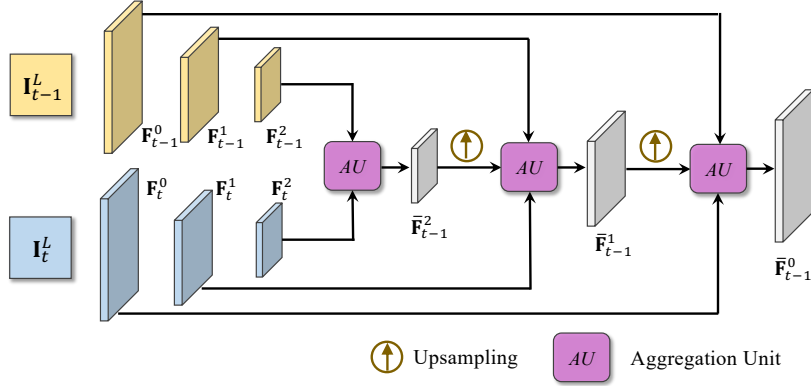


Fig. 3. Structure of temporal multi-correspondence aggregation module (TM-CAM).

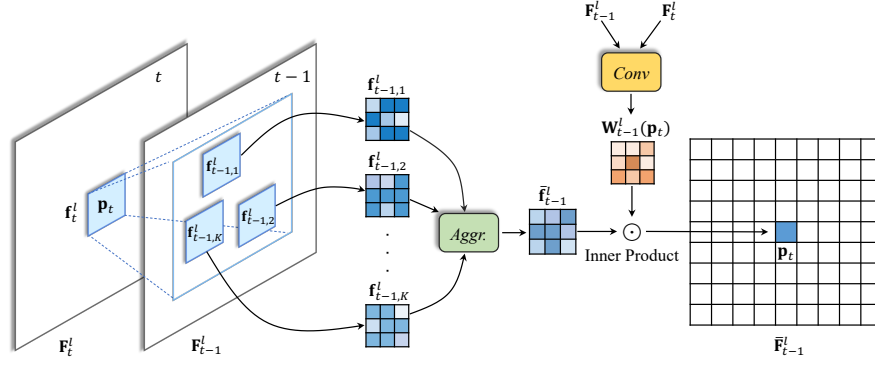


Fig. 4. Aggregation unit in TM-CAM. It aggregates multiple inter-frame correspondences to recover a given pixel at \mathbf{p}_t .

the true correspondences between frames. We thus aggregate multiple candidates to obtain sufficient context information as in Figure 4.

In details, we first locally select top- K most similar feature patches, and then utilize a pixel-adaptive aggregation scheme to fuse them into one pixel to avoid boundary problems. Taking aligning \mathbf{F}_{t-1}^l to \mathbf{F}_t^l as an example, given an image patch \mathbf{f}_t^l (represented as a feature vector) in \mathbf{F}_t^l , we first find its nearest neighbors on \mathbf{F}_{t-1}^l . For efficiency, we define a local search area satisfying $|\mathbf{p}_t - \mathbf{p}_{t-1}| \leq \mathbf{d}$, where \mathbf{p}_t is the position vector of \mathbf{f}_t^l . \mathbf{d} means the maximum displacement. We use correlation as a distance measure as in Flownet [6]. For \mathbf{f}_{t-1}^l and \mathbf{f}_t^l , their correlation is computed as the normalized inner product of

$$\text{corr}(\mathbf{f}_{t-1}^l, \mathbf{f}_t^l) = \frac{\mathbf{f}_{t-1}^l \cdot \mathbf{f}_t^l}{\|\mathbf{f}_{t-1}^l\| \cdot \|\mathbf{f}_t^l\|}. \quad (1)$$

After calculating correlations, we select top- K most correlated patches (i.e., $\bar{\mathbf{f}}_{t-1,1}^l, \bar{\mathbf{f}}_{t-1,2}^l, \dots, \bar{\mathbf{f}}_{t-1,K}^l$) in a descending order from \mathbf{F}_{t-1}^l , and concatenate and aggregate them as

$$\bar{\mathbf{f}}_{t-1}^l = \text{Aggr} \left(\left[\bar{\mathbf{f}}_{t-1,1}^l, \bar{\mathbf{f}}_{t-1,2}^l, \dots, \bar{\mathbf{f}}_{t-1,K}^l \right] \right), \quad (2)$$

where Aggr is implemented as convolution layers. Instead of assigning equal weights (e.g., $\frac{1}{9}$ when the patch size is 3), we design a pixel-adaptive aggregation strategy to enable varying aggregation patterns in different locations. The weight map is obtained by concatenating \mathbf{F}_{t-1}^l and \mathbf{F}_t^l and going through a convolution layer, which has a size of $H \times W \times s^2$ when the patch size is $s \times s$. More precisely, the adaptive weight map takes the form of

$$\mathbf{W}_{t-1}^l = \text{Conv} \left(\left[\mathbf{F}_{t-1}^l, \mathbf{F}_t^l \right] \right). \quad (3)$$

As shown in Figure 4, the final value at position \mathbf{p}_t on the aligned neighboring frame $\bar{\mathbf{F}}_{t-1}^l$ is obtained as

$$\bar{\mathbf{F}}_{t-1}^l(\mathbf{p}_t) = \bar{\mathbf{f}}_{t-1}^l \cdot \mathbf{W}_{t-1}^l(\mathbf{p}_t). \quad (4)$$

After repeating the above steps for $2N$ times, we obtain a set of aligned neighboring feature maps $\{\bar{\mathbf{F}}_{t-N}^0, \dots, \bar{\mathbf{F}}_{t-1}^0, \bar{\mathbf{F}}_{t+1}^0, \dots, \bar{\mathbf{F}}_{t+N}^0\}$. To handle all frames at the same feature level, as shown in Figure 2, we employ an additional TM-CAM, which performs self-aggregation with \mathbf{I}_t^L as the input and produces $\bar{\mathbf{F}}_t^0$. Finally, all these feature maps are fused into a double-spatial-sized feature map by a convolution and PixelShuffle operation, which is to keep sub-pixel details.

3.2 Cross-Scale Nonlocal-Correspondence Aggregation Module

Similar patterns exist widely in natural images that can provide abundant texture information. Self-similarity [20, 29, 9, 44, 47] can help detail recovery. In this part, we design a cross-scale aggregation strategy to capture nonlocal correspondences across different feature resolutions, as illustrated in Figure 5.

To distinguish from Sec. 3.1, we use \mathbf{M}_t^s to denote feature maps at time t with scale level s . We first downsample the input feature maps \mathbf{M}_t^0 and obtain a feature pyramid as

$$\mathbf{M}_t^{s+1} = \text{AvgPool}(\mathbf{M}_t^s), \quad s = \{0, 1, 2\}, \quad (5)$$

where AvgPool is the average pooling with stride 2. Given a query patch \mathbf{m}_t^0 in \mathbf{M}_t^0 centered at position \mathbf{p}_t , we implement a non-local search on other three scales to obtain

$$\tilde{\mathbf{m}}_t^s = \text{NN}(\mathbf{M}_t^s, \mathbf{m}_t^0), \quad s = \{1, 2, 3\}, \quad (6)$$

where $\tilde{\mathbf{m}}_t^s$ denotes the nearest neighbor (the most correlated patch) of \mathbf{m}_t^0 in \mathbf{M}_t^s . Before merging, a self-attention module [41] is applied to determine whether the

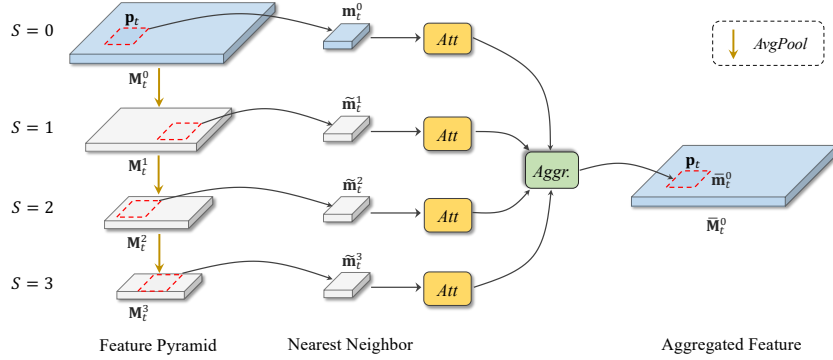


Fig. 5. Cross-scale nonlocal-correspondence aggregation module (CN-CAM).

information is useful or not. Finally, the aggregated feature $\tilde{\mathbf{m}}_t^0$ at position \mathbf{p}_t is calculated as

$$\tilde{\mathbf{m}}_t^0 = \text{Aggr}([\text{Att}(\mathbf{m}_t^0), \text{Att}(\tilde{\mathbf{m}}_t^1), \text{Att}(\tilde{\mathbf{m}}_t^2), \text{Att}(\tilde{\mathbf{m}}_t^3)]) , \quad (7)$$

where Att is the attention unit and Aggr is implemented as convolution layers. Our results presented in Sec. 4.3 demonstrate that CN-CAM reveals more details.

3.3 Edge-Aware Loss

Usually, reconstructed high-resolution images produced by VSR methods suffer from jagged edges. To alleviate this problem, we propose an edge-aware loss to obtain better refined edges. First, an edge detector is used to extract edge information of ground-truth HR images. Then, the detected edge areas are weighted more in loss calculation, enforcing the network to pay more attention to these areas during the learning.

In this paper, we choose the Laplacian filter as the edge detector. Given the ground-truth \mathbf{I}_t^H , the edge map \mathbf{I}_t^E is obtained from the detector and the binary mask value at \mathbf{p}_t is represented as

$$\mathbf{B}_t(\mathbf{p}_t) = \begin{cases} 1, & \mathbf{I}_t^E(\mathbf{p}_t) \geq \delta \\ 0, & \mathbf{I}_t^E(\mathbf{p}_t) < \delta, \end{cases} \quad (8)$$

where δ is a predefined threshold. Suppose the size of a high-resolution image is $H \times W$. The edge mask is also a $H \times W$ map filled with binary values. The areas marked as edges are 1 while others being 0.

During training, we adopt the Charbonnier Loss, which is defined as

$$L = \sqrt{\|\hat{\mathbf{I}}_t^H - \mathbf{I}_t^H\|^2 + \epsilon^2}, \quad (9)$$

where $\hat{\mathbf{I}}_t^H$ is the predicted high-resolution result, and ϵ is a small constant. The final loss is formulated as

$$L_{final} = L + \lambda \left\| \mathbf{B}_t \circ \left(\hat{\mathbf{I}}_t^H - \mathbf{I}_t^H \right) \right\|, \quad (10)$$

where λ is a coefficient to balance the two terms and \circ is element-wise multiplication.

4 Experiments

4.1 Dataset and Evaluation Protocol

REDS [27] is a realistic and dynamic scene dataset published in NTIRE 2019 challenge. There are a total of 30K images extracted from 300 video sequences. The training, validation and test subsets contain 240, 30 and 30 sequences, respectively. Each sequence has equally 100 images with resolution 720×1280 . Similar to that of [41], we merge the training and validation parts and divide the data into new training (with 266 sequences) and testing (with 4 sequences) datasets. The new testing part contains the 000, 011, 015 and 020 sequences.

Vimeo-90K [43] is a large-scale high-quality video dataset designed for various video tasks. It consists of 89,800 video clips which cover a broad range of actions and scenes. The super-resolution subset has 91,701 7-frame sequences with fixed resolution 448×256 , among which training and testing splits contain 64,612 and 7,824 sequences respectively.

Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [42] are used as metrics in our experiments.

4.2 Implementation Details

Network Settings The network takes 5 (or 7) consecutive frames as input. In feature extraction and reconstruction modules, 5 and 40 (20 for 7 frames) residual blocks [12] are implemented respectively with channel size 128. In Figure 3, the patch size is 3 and the maximum displacements are set to $\{3, 5, 7\}$ from low to high resolutions. The K value is set to 4. In the cross-scale aggregation module, we define patch size as 1 and fuse information from 4 scales as shown in Figure 5. After reconstruction, both the height and width of images are quadrupled.

Training We train our network using eight NVIDIA GeForce GTX 1080Ti GPUs with mini-batch size 3 per GPU. The training takes 600K iterations for all datasets. We use Adam as the optimizer and cosine learning rate decay strategy with an initial value $4e - 4$. The input images are augmented with random cropping, flipping and rotation. The cropping size is 64×64 corresponding to an output 256×256 . The rotation is selected as 90° or -90° . When calculating the edge-aware loss, we set both δ and λ as 0.1.

Testing During testing, the output is evaluated without boundary cropping.

Table 1. Ablation Study of our proposed modules and loss on the REDS testing dataset. ‘Baseline’ is without using the proposed modules and loss. ‘TM-CAM’ represents the temporal multi-correspondence aggregation module. ‘CN-CAM’ means the cross-scale nonlocal-correspondence aggregation module. ‘EAL’ is the proposed edge-aware loss.

Components				PSNR(dB)	SSIM
Baseline	TM-CAM	CN-CAM	EAL		
✓				28.98	0.8280
✓	✓			30.13	0.8614
✓	✓	✓		30.25	0.8641
✓	✓	✓	✓	30.31	0.8648

Table 2. Results of TM-CAM with different numbers (K) of aggregated temporal correspondences on the REDS testing dataset.

K	PSNR(dB)	SSIM
1	30.19	0.8624
2	30.24	0.8640
4	30.31	0.8648
6	30.30	0.8651

4.3 Ablation Study

To demonstrate the effectiveness of our proposed method, we conduct experiments for each individual design. For convenience, we adopt a lightweight setting in this section. The channel size of network is set to 64 and the reconstruction module contains 10 residual blocks. Meanwhile, the amount of training iterations is reduced to 200K.

Temporal Multi-Correspondence Aggregation Module. To make fair comparison, we first build a baseline without the proposed ideas. As shown in Table 1, the baseline only yields 28.98dB PSNR and 0.8280 SSIM, a relatively poor result. Our designed module brings about a 1.15dB improvement on PSNR.

To show the effectiveness of TM-CAM in a more intuitive way, we visualize residual maps between aligned neighboring feature maps and reference feature maps in Figure 6. After aggregation, it is clear that feature maps obtained with the proposed TM-CAM are smoother and cleaner. The mean $L1$ distance between aligned neighboring and reference feature maps are smaller. All these facts manifest the great alignment performance of our method.

Then, we evaluate how the number of aggregated temporal correspondences affects performance. Table 2 shows that the capability of TM-CAM rises at first and drops with the increasing number of correspondences. Compared with taking only one candidate, the four-correspondence setting obtains more than 0.1dB

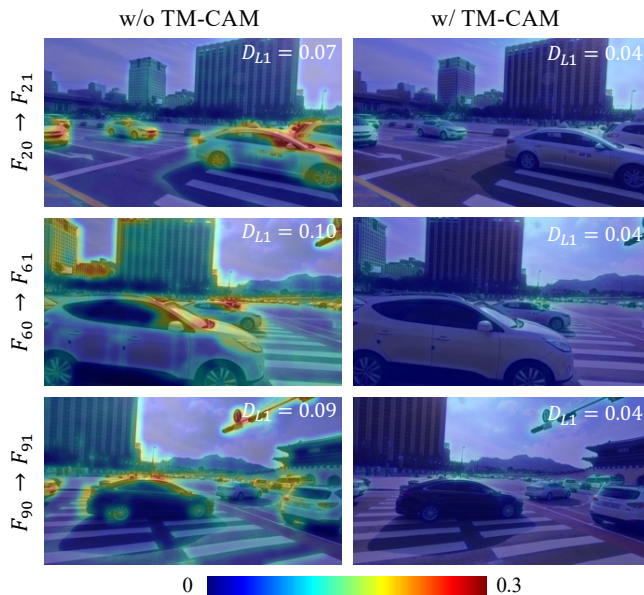


Fig. 6. Residual maps between aligned neighboring feature maps and reference feature maps without and with temporal multi-correspondence aggregation module (TM-CAM) on the REDS dataset. Values on the upper right represent the average $L1$ distance between aligned neighboring feature maps and reference feature maps.

gain on PSNR. It demonstrates that highly correlated correspondences can provide useful complementary details. However, once saturated, it is not necessary to include more correspondences, since weakly correlated correspondences actually bring unwanted noise. We further verify this point by estimating optical flow using a KNN strategy for the MPI Sintel Flow dataset [1]. From Figure 1(b), we find that the four-neighbor setting is also the best choice. Therefore, we set K as 4 in our implementation.

Finally, we verify the performance of pixel-adaptive weights. Based on the experiments, we find that a larger patch in TM-CAM usually gives a better result. It is reasonable since neighboring pixels usually have similar information and are likely to complement each other. Also, structural information is embedded. To balance between performance and computing cost, we set the size to 3. When using fixed weights (at $K = 4$), we obtain the resulting PSNR/SSIM as 30.12dB/0.8614. From Table 2, the proposed pixel-adaptive weighting scheme achieves 30.31dB/0.8648, which is superior to the fixed counterpart by nearly 0.2dB on PSNR, which demonstrates that different aggregating patterns are necessary for consideration of spatial variance.

More experiments of TM-CAM with regard to patch size and maximum displacements are provided in the supplementary file.

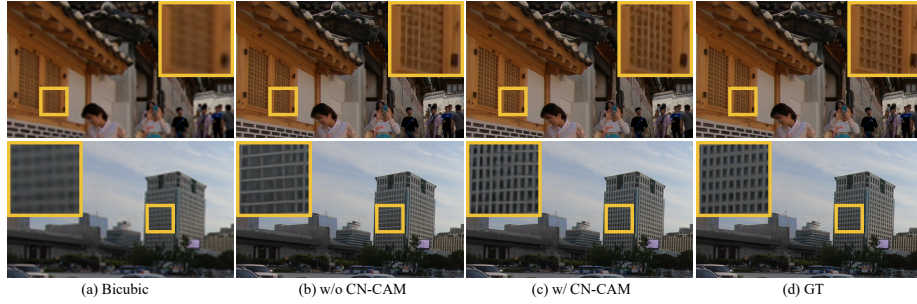


Fig. 7. Examples without and with the cross-scale nonlocal-correspondence aggregation module (CN-CAM) on the REDS dataset.

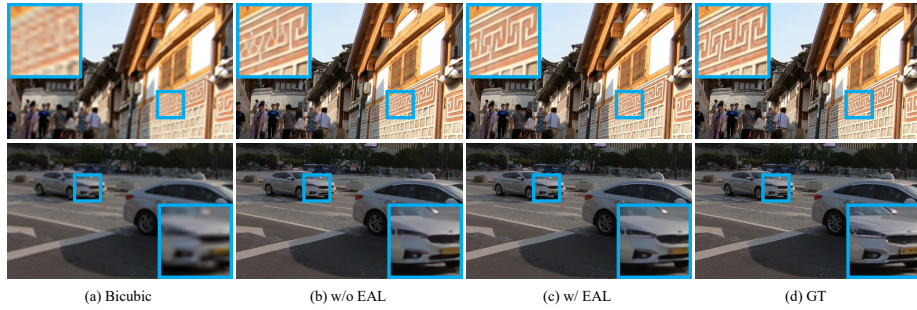


Fig. 8. Examples without and with the proposed edge-aware loss (EAL) on the REDS dataset.

Cross-Scale Nonlocal-Correspondence Aggregation Module In Sec. 4.3, we already notice that highly correlated temporal correspondences can serve as a supplement in motion compensation. To handle cases with different scales, we proposed a cross-scale nonlocal-correspondence aggregation module (CN-CAM).

As listed in Table 1, CN-CAM improves PSNR by 0.12dB. Besides, from Figure 7, we observe that this module enables the network to reveal more details when images contain repeated patterns such as windows and buildings within the spatial domain or across scales. All these results show that the proposed CN-CAM method further enhances the quality of reconstructed images.

Edge-Aware Loss In this part, we evaluate the proposed edge-aware loss (EAL). Table 1 lists the statistics. A few visual results in Figure 8 indicate that EAL improves the proposed network further, yielding more refined edges. The textures on the wall and edges of lights are clearer and sharper, which demonstrates the effectiveness of the proposed edge-aware loss.

Table 3. Comparisons of PSNR(dB)/SSIM results on the REDS dataset for $\times 4$ setting. ‘*’ denotes without pretraining.

Method	Frames	Clip_000	Clip_011	Clip_015	Clip_020	Average
Bicubic	1	24.55/0.6489	26.06/0.7261	28.52/0.8034	25.41/0.7386	26.14/0.7292
RCAN [48]	1	26.17/0.7371	29.34/0.8255	31.85/0.8881	27.74/0.8293	28.78/0.8200
TOFlow [43]	7	26.52/0.7540	27.80/0.7858	30.67/0.8609	26.92/0.7953	27.98/0.7990
DUF [15]	7	27.30/0.7937	28.38/0.8056	31.55/0.8846	27.30/0.8164	28.63/0.8251
EDVR* [41]	5	27.78/0.8156	31.60/0.8779	33.71/0.9161	29.74/0.8809	30.71/0.8726
MuCAN (Ours)	5	27.99/0.8219	31.84/0.8801	33.90/0.9170	29.78/0.8811	30.88/0.8750

Table 4. Comparisons of PSNR(dB)/SSIM results on the Vimeo-90K dataset for $\times 4$ setting. ‘-’ indicates results not available.

Method	Frames	RGB	Y
Bicubic	1	29.79 / 0.8483	31.32 / 0.8684
RCAN [48]	1	33.61 / 0.9101	35.35 / 0.9251
DeepSR [23]	7	25.55 / 0.8498	-
BayesSR [24]	7	24.64 / 0.8205	-
TOFlow [43]	7	33.08 / 0.9054	34.83 / 0.9220
DUF [15]	7	34.33 / 0.9227	36.37 / 0.9387
RBPB [10]	7	-	37.07 / 0.9435
MuCAN (Ours)	7	35.49 / 0.9344	37.32 / 0.9465

4.4 Comparison with State-of-the-art Methods

We compare our proposed multi-correspondence aggregation network (MuCAN) with previous state-of-the-arts including TOFlow [43], DUF [15], RBPB [10], EDVR [41], etc., on REDS [27], Vimeo-90K [43], Vid4 [24] and SPMCS [37] datasets. The quantitative results in Tables 3 and 4 are from the original publications. Especially, original EDVR [41] is initialized with a well-trained model. For fairness, we use the author-released code to train EDVR without pretraining.

On the REDS dataset, results are shown in Table 3. It is clear that our method outperforms other methods by at least **0.17dB**. As for Vimeo-90K, the results are reported in Table 4. Our MuCAN method works better than DUF [15] with nearly **1.2dB** enhancement on RGB channels. Meanwhile, it obtains **0.25dB** improvement on the Y channel compared with RBPB [10]. All of these results demonstrate the effectiveness of our method. Besides, the performance on the Vid4 and SPMCS datasets are reported in the supplementary file. Several examples are visualized in Figure 9.

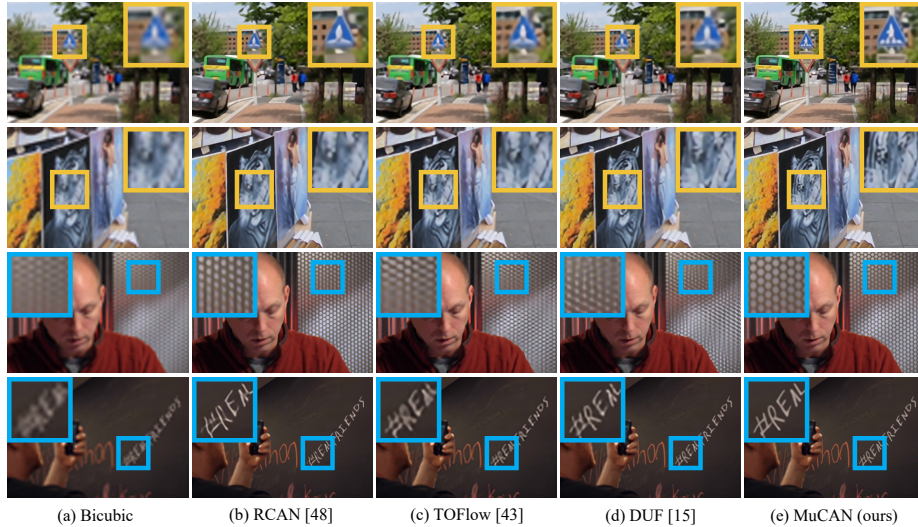


Fig. 9. Examples of REDS (top two rows) and Vimeo-90K (bottom two rows) datasets.

4.5 Generalization Analysis

To evaluate the generality of our method, we apply our model trained on the REDS dataset to test video frames in the wild. In addition, we test EDVR with the author-released model¹ on the REDS dataset. Some visual results are shown in the supplementary file. We remark that EDVR may generate visual artifacts in some cases due to the variance of data distributions between training and testing. In contrast, our MuCAN demonstrates its decent generality in the real world setting.

5 Conclusion

In this paper, we have proposed a novel multi-correspondence aggregation network (MuCAN) for the video super-resolution task. We showed that the proposed temporal multi-correspondence aggregation module (TM-CAM) takes advantage of highly correlated patches to achieve a better alignment-based frame recovery. Additionally, we verified that the cross-scale nonlocal-correspondence aggregation module (CN-CAM) utilizes multi-scale information and further boosts the performance of our network. Also, the edge-aware loss enforces the network to obtain more refined edges on the high-resolution output. Extensive experiments have demonstrated the effectiveness and generality of our proposed method.

¹ <https://github.com/xinntao/EDVR>

References

1. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. pp. 611–625. Springer (2012)
2. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4778–4787 (2017)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
4. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. pp. 184–199. Springer (2014)
5. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. pp. 391–407. Springer (2016)
6. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
7. Freedman, G., Fattal, R.: Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)* **30**(2), 12 (2011)
8. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. *IEEE Computer graphics and Applications* (2), 56–65 (2002)
9. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: 2009 IEEE 12th international conference on computer vision. pp. 349–356. IEEE (2009)
10. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1664–1673 (2018)
11. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2015)
14. Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: Advances in Neural Information Processing Systems. pp. 235–243 (2015)
15. Jo, Y., Wug Oh, S., Kang, J., Joo Kim, S.: 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. pp. 3224–3232 (2018)
16. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging* **2**(2), 109–122 (2016)

17. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
18. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1637–1645 (2016)
19. Kim, S.Y., Lim, J., Na, T., Kim, M.: 3dsrnet: Video super-resolution using 3d convolutional neural networks. arXiv preprint arXiv:1812.09079 (2018)
20. Kindermann, S., Osher, S., Jones, P.W.: Deblurring and denoising of images by nonlocal functionals. *Multiscale Modeling & Simulation* **4**(4), 1091–1115 (2005)
21. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
22. Li, S., He, F., Du, B., Zhang, L., Xu, Y., Tao, D.: Fast residual network for video super-resolution. arXiv preprint arXiv:1904.02870 (2019)
23. Liao, R., Tao, X., Li, R., Ma, Z., Jia, J.: Video super-resolution via deep draft-ensemble learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 531–539 (2015)
24. Liu, C., Sun, D.: On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence* **36**(2), 346–360 (2013)
25. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T.: Robust video super-resolution with learned temporal dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2507–2515 (2017)
26. Ma, Z., Liao, R., Tao, X., Xu, L., Jia, J., Wu, E.: Handling motion blur in multi-frame super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5224–5232 (2015)
27. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
28. Pérez-Pellitero, E., Salvador, J., Ruiz-Hidalgo, J., Rosenhahn, B.: Psycho: Manifold span reduction for super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1837–1845 (2016)
29. Protter, M., Elad, M., Takeda, H., Milanfar, P.: Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Transactions on image processing* **18**(1), 36–51 (2008)
30. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4161–4170 (2017)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
32. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6626–6634 (2018)
33. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3791–3799 (2015)
34. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient

- sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
35. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8934–8943 (2018)
 36. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3147–3155 (2017)
 37. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4472–4480 (2017)
 38. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally deformable alignment network for video super-resolution. arXiv preprint arXiv:1812.02898 (2018)
 39. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE international conference on computer vision. pp. 1920–1927 (2013)
 40. Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian conference on computer vision. pp. 111–126. Springer (2014)
 41. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
 42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
 43. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. *International Journal of Computer Vision* **127**(8), 1106–1125 (2019)
 44. Yang, C.Y., Huang, J.B., Yang, M.H.: Exploiting self-similarities for single frame super-resolution. In: Asian conference on computer vision. pp. 497–510. Springer (2010)
 45. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.: Coupled dictionary training for image super-resolution. *IEEE transactions on image processing* **21**(8), 3467–3478 (2012)
 46. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. *IEEE transactions on image processing* **19**(11), 2861–2873 (2010)
 47. Zhang, X., Burger, M., Bresson, X., Osher, S.: Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Sciences* **3**(3), 253–276 (2010)
 48. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)
 49. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018)